LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Error-Correction Mechanisms in Language Learning: Modeling Individuals

Adnane Ez-zizi [a,b] Dagmar Divjak [a] and Petar Milin [a]

[a]University of Birmingham [b]University of Suffolk

Abstract: Since its first adoption as a computational model for language learning, evidence has accumulated that Rescorla–Wagner error-correction learning (Rescorla & Wagner, 1972) captures several aspects of language processing. Whereas previous studies have provided general support for the Rescorla–Wagner rule by using it to explain the behavior of participants across a range of tasks, we focus on testing predictions generated by the model in a controlled natural language learning task and model the data at the level of the individual learner. By adjusting the parameters of the model to fit the trial-by-trial behavioral choices of participants, rather than fitting a one-for-all model using a single set of default parameters, we show that the model accurately captures par-

CRediT author statement – **Adnane Ez-zizi**: conceptualization; methodology; investigation; data curation; formal analysis; software; visualization; writing – original draft preparation; writing – review & editing. **Dagmar Divjak**: conceptualization; methodology; funding acquisition; writing – review & editing. **Petar Milin**: conceptualization; methodology; funding acquisition; validation; writing – review & editing.

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

We are grateful to Emma Marsden and four anonymous reviewers for their insightful comments and suggestions on this article. We would also like to thank Maciej Borowski for his help with the choice of the stimuli and for discussion of relevant aspects of the Polish language. We are also grateful to Christian Adam for his help with running pilots of the experiment and for sharing his feedback on those pilots. This research was funded by Leverhulme Trust Leadership Grant RL-016-001 to Dagmar Divjak, which funded all authors. The first author was affiliated with the University of Birmingham while work on the paper was carried out.

Correspondence concerning this article should be addressed to Petar Milin, Ashley Building, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom. Email: p.milin@bham.ac.uk

The handling editors for this manuscript were Emma Marsden and Pavel Trofimovich.

DOI: 10.1111/lang.12569

ticipants' choices, time latencies, and levels of response agreement. We also show that gender and working memory capacity affect the extent to which the Rescorla–Wagner model captures language learning.

Keywords language learning; error-correction learning; Rescorla–Wagner model; morphology; agreement

Introduction

We humans share with other species many core learning mechanisms that allow us to adapt to our environment (Rescorla, 1988). These mechanisms include, among others, classical conditioning (i.e., Pavlovian conditioning; Pavlov, 1927), instrumental conditioning (also operant conditioning; Skinner, 1938), and forms of social learning, such as vicarious learning (Bandura, 1962). Arguably the most uniquely defining human learning ability is language learning, which also includes efficient transgenerational transmission and is foundational for social inclusion and cohesion. However, whereas core learning mechanisms are relatively well understood, language learning remains much of a mystery (Ambridge & Lieven, 2011). An early attempt by Skinner (1957) to account for language learning using the same principles as those governing lower-level cognitive tasks was quashed by Chomsky (1959). For much of the remainder of the 20th century, language was seen as a by-and-large innate system, governed by rules and handled by a uniquely human and specialized cognitive structure. This structure was initially conceptualized as the language acquisition device, and later extended to become universal grammar.

This dominant view was challenged from two sides simultaneously. The emergence of usage-based linguistics in the 1980s (Langacker, 1987) promoted a view of language as a dynamic and probabilistic system, resulting from general cognitive capacities acting on language input (Dąbrowska & Divjak, 2015). This view meshed well with connectionist frameworks, which showed that rulelike behavior can emerge from exposure to usage alone and that language knowledge is sensitive to properties of the input (Plaut & Gonnerman, 2000; Seidenberg & McClelland, 1989). Connectionism, arguably, paved the way for changes in theorizing too, toward a view of language as being learned like any other skill, and the early 2000s witnessed the start of a reintegration of the basic principles of learning into the body of work on language (e.g., see Bybee & McClelland, 2005; for more up-to-date works, see Ellis et al., 2016, which addresses both first and second language learning, as well as Chuang et al., 2021, which addresses lexical acquisition in second and third languages). Language was now seen as being amenable to the same general-purpose

cognitive capacities and learning mechanisms that humans and animals use to navigate and adapt to their environment (cf. Ellis, 2006a; Ellis & Sagarra, 2010, 2011; Sturdy & Nicoladis, 2017).

Among these learning models, Rescorla and Wagner's (1972) model of classical conditioning stands out for its simplicity and its ability to explain a range of empirical learning phenomena (Siegel & Allan, 1996). This model has been shown to be biologically plausibile (Chen et al., 2008) and to have an evolutionary advantage over other more powerful learning mechanisms, in the sense that it has a higher likelihood of being naturally selected and persisting in the evolutionary process, compared to other plausible learning mechanisms (for more details, see Trimmer et al., 2012).

Background Literature

The Rescorla–Wagner Model

As a model of classical conditioning, the Rescorla–Wagner (R–W) model is concerned with situations where an entity (a human, an animal, or a machine) has to learn the predictive relationship between objects and/or events (i.e., cues and outcomes) in an environment, and where cues compete for their predictive value for an outcome while iteratively (re)calibrating the learning (or association) weights. More specifically, an association weight reflects the tendency of an outcome to occur in the presence of a certain cue. A higher positive association weight value for a particular outcome corresponds to a higher likelihood of occurrence of that outcome in the presence of the cue; conversely, a highly negative value corresponds to a greater likelihood of nonoccurrence of that outcome (the cue is said to be inhibitory in this case). Values close to zero mean low chances of observing (if the weight is positive) or inhibiting (if the weight is negative) the outcome.

The R–W model assumes that the organism computes a simple error-correcting learning rule used to update the association weights in each new learning event (e.g., each trial in a behavioral experiment). The general idea behind the correction rule is that the association between a cue and outcome is (a) strengthened if both cue and outcome are present in the learning event, (b) weakened if the cue is present but the outcome is not, and (c) kept the same if the cue itself is absent. The updating of the association weights is driven by the discrepancy between the expected and the obtained outcome, such that the magnitude of the update—how much the association weights are adjusted—is determined by two parameters called learning rates, and the direction of the update—whether it increases the weight or decreases it—depends on the sign of the difference between the expected and the observed outcome. In this way,

most broadly, for the R–W model, learning is about the outcomes, and this sets it apart from related models where learning is about the input cues (e.g., Pearce & Hall, 1980).

Another feature of the R–W model is that, although the outcomes are updated independently from each other, input cues compete for the predictivity of outcomes. In other words, the adjustment of the weights depends not only on the single cue being updated but on all the cues present in the learning event through their sum of association weights. This cue competition principle is what allowed the R–W model to explain many of the puzzling phenomena of classical conditioning, some of which were also shown to be valuable for understanding the mechanics of language learning (see the next section for a discussion).[1] One of the best-known examples of such learning phenomena is the blocking effect (Kamin, 1969). This effect occurs when a cue is trained in compound with a second cue to predict an outcome but when the second cue is already a good predictor of the outcome. In such cases, the first cue cannot form a strong association with the outcome (i.e., the first cue is blocked by the second cue). More generally, the cue competition principle often results in the observation that the best cues for the outcome prevent other cues from developing a strong association with that same outcome.

The Rescorla–Wagner Model and Language Learning

Since its first mention within a linguistic context by Ellis (2006a), evidence has accumulated showing that the R–W model can capture several aspects of language learning (e.g., Baayen et al., 2011; Ellis, 2006b; Milin, Divjak, & Baayen, 2017; Milin, Feldman, et al., 2017). So far, the available empirical evidence stems from studies that train a R–W model that uses default parameter values (here we allude to the two learning rate parameters used to update the association weights after each new event), typically on either a small sample from experiments on artificial languages or a large corpus of texts.[2] Posttraining learning measures are then extracted from the simulated model and are compared against observed response measurements from an experimental task.

A first issue is that predictions for (and from) such models are typically generated independently from the experiment (with exceptions such as the studies of Ramscar & Yarlett, 2007, and Divjak et al., 2021, where the model generated the hypotheses to be tested experimentally). The parameters are typically set to their default values, missing the opportunity to take into account the variability that can arise from simulating the model with different parameter values (though see Olejarczuk et al., 2018, who used fixed parameter values but fitted a separate model to each participant's data using the same sequence

of examples encountered by the participant). Incorporating the variability arising from the model parameters when fitting learning models to language data has the potential to improve the explainability of the individual differences observed in the experiment, especially since language usage and representation is an area that shows huge individual variation (Dąbrowska, 2018).

Training the model on a large-scale corpus comes at an even greater cost. We leave aside here the issue of (lack of) similarity between the contents of a corpus and the input that language users receive (which plagues converging evidence studies generally; for a summary discussion, see Klavan & Divjak, 2016, and for collections of worked examples, see Divjak & Gries, 2012, and Gries & Divjak, 2012). Here we focus on another issue: Training on a corpus mutes the two main sources of variability of the model—namely, those related to the choice of model parameters and the order of training examples—which are mostly active during the early stages of learning (Shanks, 1995; also see Milin et al., 2020, for a more general discussion of the trial order effect in error-correction learning).[3] These early biases, as Ellis (2006a) called them, constitute a real test for the R–W model, before it can be deployed as a model of language learning at a large scale. Modeling the parameters' variability and training the R–W model on the same examples encountered by the participants represent novel opportunities for understanding language learning not yet fully explored in previous studies.

The Present Study

The aim of the present study is to model how individual language learners engage with the task at hand on a trial-by-trial basis, which constitutes a step-changing challenge for the application to language learning of discrimination or error-correction learning in general and the R–W model in particular. Whereas previous studies have provided general support for the R–W rule by using this model to explain the behavior of participants across a range of tasks (Divjak, 2019; Milin & Blevins, 2020; Milin, Feldman, et al., 2017; Pirrelli et al., 2020), we focus on testing predictions generated by the model in a controlled natural language learning task and model the data at the level of the individual language learner. In doing so, we treat each participant as a separate learning entity governed by different capacities, which are, crucially, formalized through the learning parameters of the chosen model.

Given that several studies have reported that classical conditioning performance can be affected by cognitive and personal characteristics such as working memory (Baetu et al., 2018; Sasaki, 2009), gender (Lonsdorf et al., 2015; Merz et al., 2018), and age (e.g., Mutter et al., 2012), we also

investigate whether such characteristics could affect the adoption of a R–W-like mechanism of language learning.

To achieve these goals and to address these questions, we designed a simplified natural language learning task: simplified in order to exploit the advantage of tight empirical control, but only partly so in order to maintain a commitment to ecological validity by offering a more naturalistic language input experience. The task represents, to a reasonable extent, how people would learn Polish subject–verb agreement mappings through natural exposure to examples.

We trained native English speakers on a set of carefully crafted examples, which had both auditory and visual dimensions, and which incorporated some of the complexities inherent to subject–verb agreement in Polish. Next, for each participant, individually, we selected the best-fitting model (i.e., the parameters that led to the closest match between the responses of the participant and the model), using the same training examples encountered by the participant. We then assessed the R–W model for its capacity to recover participants' language choices as well as their time latencies, and compared it to other plausible, yet rule-based response strategies. Finally, we tested whether cognitive and personal characteristics such as working memory capacity, age, and gender affect the extent to which the R–W model captures language learning.

Method

Participants

Sixty-six participants (Mdn_{age} = 20 years; range = 18–65; 41 females) took part in the experiment in exchange for a £7 Amazon voucher. Participants were university students and staff. All of them were native English speakers without knowledge of Polish or any other Slavic languages, had normal or corrected-to-normal hearing and vision, and did not declare any learning disabilities. Participants had different educational backgrounds, and many of them could speak other languages in addition to English (the distributions of education and language backgrounds are presented in Appendix S1 in the Supporting Information online).

Materials and Procedure

All our materials, including data and code, are openly available on Github (https://github.com/ooominds/Error-correction-mechanisms-in-language-learning) and the University of Birmingham's open-access repository, UBIRA (https://doi.org/10.25500/edata.bham.00000911). Participants completed three tasks and a short questionnaire in the following order: (a) a language learning

task (main task), (b) an explicit knowledge and demographic questionnaire, (c) an implicit learning task, and (d) a working memory (WM) task. (A detailed description of each task is provided in the next section.) The language learning and implicit learning tasks were implemented and presented to participants using OpenSesame (Mathôt et al., 2012; Mathôt & March, 2022). The demographic questionnaire was presented using Google forms, and the WM task was administered using Tatool (von Bastian et al., 2013). The experiment was run either individually or, whenever possible, in pairs, in a quiet room, on Intel Core i7-8700 computers running Windows 10 and equipped with Iiyama G-Master 24.5-in. monitors running at 59 Hz with a screen resolution of 1,920 × 1,080 pixels. Participants heard the auditory stimuli via Bose Quietcomfort 35 II noise-canceling headphones and registered their responses using a keyboard. The experiment took about 50 min to complete.

Language Learning Task

Our simplified natural language learning task was inspired by the challenge of learning subject–verb agreement in the plural past tense in Polish. In the past tense, verbs are marked for the grammatical gender of the subject according to the following rules:

1. If one of the referents is masculine personal (e.g., "man"), then the gender of the subject as a whole is *–li*, which is sometimes referred to as the masculine personal ending.
2. If the referents are feminine animate (e.g., "duck"), feminine personal (e.g., "girl"), or neuter (e.g., "child"), then the gender of the subject is *–ły*, which is sometimes referred to as the nonmasculine personal ending.
3. Prescriptive grammars and native speakers of Polish disagree as to what form should be assigned to a subject that includes multiple masculine animate referents that are not persons (e.g., "the dog and the cat went for a walk") or mixes masculine animate and feminine personal referents (e.g., "the girl and the dog went for a walk"). Grammar textbooks prescribe the use of *–ły* whereas native speakers appear to favor the use of *–li*, according to Kiełkiewicz-Janowiak and Pawelczyk (2014).

For the purpose of designing our simplified natural language learning task, we implemented the first two rules but assumed that a subject consisting of masculine animate referents should be used with the masculine personal form,

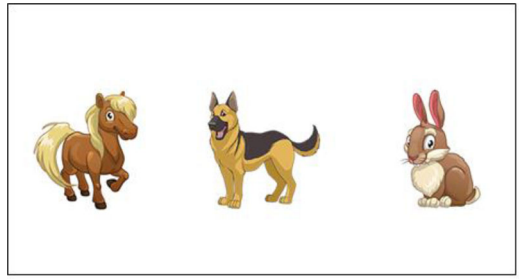

Figure 1 An example of stimuli presented in one trial to depict a scene.

as suggested by Kiełkiewicz-Janowiak and Pawelczyk (2014). We made such a concession to improve the testability of our task from a learning perspective, as explained in the task design below.

Stimuli

Each event in our learning task consisted of a scene that represented a joint action performed by a group of human and/or animal characters, and for each learning event, participants saw a picture that depicted the scene (Figure 1), along with an audio recording of a Polish clause describing it. A new trial started with a fixation dot that was shown at the center of the screen for about 500 ms, followed by the simultaneous display of the picture of the scene. Participants heard the audio recording of the clause describing the scene 250 ms after the onset of the picture of the scene while the picture remained on display. A new trial was then presented after about 1 s.

We used the verb *chodzić* ("walk"), with its two possible plural past tense forms *chodziły* (nonmasculine plural form) and *chodzili* (masculine plural form), as the common action in all learning events. An example of a clause heard by participants is *Chłopiec i kaczka chodzili* ("The boy and the duck were walking"). The first three columns in Table 1 provide a list of all characters used in the experiment, along with their linguistic categories in terms of gender and animacy; the last two columns concern the design of the task and will become relevant in the next section.

The images representing the different human and animal characters were extracted from Adobe Stock (https://stock.adobe.com) and then edited using Adobe Photoshop CC 2018. The audio recordings of both the character labels and the two verb forms were prepared using the speech synthesizer software Speech2Go (Harpo Software, 2018).

Table 1 Information about the characters used in the language learning task

Character used in the task	Polish word for the character	Linguistic category (gender & animacy)	Cue	Cue category
Duck	Kaczka	Feminine animate	FA1	uFA
Lamb	Owieczka	Feminine animate	FA2	
Goat	Koza	Feminine animate	FA3	bFA
Monkey	Małpa	Feminine animate	FA4	ibFA
Girl	Dziewczynka	Feminine personal	FP1	uFP
Woman	Kobieta	Feminine personal	FP2	
Granny	Babcia	Feminine personal	FP3	bFP
Dog	Pies	Masculine animate	MA1	uMA
Horse	Koń	Masculine animate	MA2	
Rabbit	Królik	Masculine animate	MA3	
Boy	Chłopiec	Masculine personal	MP1	uMP
Man	Mężczyzna	Masculine animate	MP2	

Note. F = feminine, M = masculine, A = animate, P = personal. The lowercase letters used with the cue categories reflect predictions from the Rescorla–Wagner theory for each cue, as explained in the design section: b = blocked (i.e., the cue is blocked according to the model); u = unblocked; ib = inhibitory blocked. For example, bFP means that the cue is a feminine personal one and is predicted to be blocked.

Design

First, participants were taught the Polish labels of the different animal and human characters used in the learning task. Specifically, participants were presented with the images of all the characters along with their corresponding labels, first individually and then in combination, as they appear later in the learning task (e.g., a dog; a boy, a dog, and a monkey). There were eight such character combinations, and participants were required to remember at least seven of them (i.e., to reach a retention accuracy of 87.5%) before they could proceed to the main task (see Appendix S2 in the Supporting Information online for more details). Participants were allowed up to 10 attempts to reach the required accuracy level.

The main task consisted of a training and a test phase. The design of the training part of the task is summarized in Table 2. The task contained 12 cues and two outcomes. The "+" sign indicates that the cues were presented in compound, and the arrow symbol "→" indicates that the outcome on the right-hand side followed the cues. Thus, for example, "FP1 + FP2 + FP3 → np" represents a clause such as *Dziewczyna, kobieta i babcia chodziły* ("The girl, the woman, and the grandma were walking"), where the subject of the clause is

Table 2 The learning events used for training

Block	Learning events
Block 1	MP1 (uMP) + FA1 (uFA) → mp
	MP2 (uMP) + FA2 (uFA) → mp
	FA1 (uFA) + FA2 (uFA) → np
	FP1 (uFP) + FP2 (uFP) → np
Block 2	FA4 (ibFA) + MP1 (uMP) + MP2 (uMP) → mp
	MA1 (uMA) + MA2 (uMA) + MA3 (uMA) → mp
	FA1 (uFA) + FA2 (uFA) + FA3 (bFA) → np
	FP1 (uFP) + FP2 (uFP) + FP3 (bFP) → np

Note. The category of each cue is provided in parentheses after the cue. F = feminine, M = masculine, A = animate, P = personal, mp = masculine plural verb form, np = nonmasculine plural verb form, b = blocked cue, u = unblocked cue, ib = inhibitory blocked cue.

made up of three female characters and the verb is in the nonmasculine plural (np) past form, as opposed to the masculine plural (mp) past form. There were two training blocks, each containing four events that were repeated 15 times each. The order of the events was fully randomized within each block. The events in the first block were composed of cue pairs, whereas those in the second block were composed of cue triples.

We structured our task in this way to create blocking-like effects as usually seen in Pavlovian learning experiments. For example, the addition of cues FA3 and FP3 to the compounds "FA1 + FA2" and "FP1 + FP2," respectively, in the second block should reduce the association strength that can be gained by FA3 and FP3 for outcome np. Likewise, training MP1 and MP2 with outcome mp in the first block should block FA4 from acquiring a positive association with mp. Besides predicting that FA4 could get blocked, we also predicted that it could become inhibitory for mp, that is, gain a negative association weight with mp, as will be seen when we present the model fit simulation results.[4] We thus refer to FA3 and FP3 as blocked cues, and to FA4 as an inhibitory blocked cue.

We categorized the cues into seven different categories based on their linguistic properties and the blockinglike effects they predict (see the rightmost column in Table 1). Specifically, the seven categories were based on whether the cue is masculine or feminine, whether it is personal or animate, whether it is predicted to be blocked or unblocked, and whether it is predicted to be an inhibitory blocked cue. The similarity between the cues within each of these categories is reinforced by the fact that they share the same association weights

with each outcome, according to the R–W theory, as will be shown in the Results section on learned noun–verb form association weights.

After training, the participant moved to the testing phase. The test consisted of two components. By using a randomly generated cue from each category, we tested learning once on all possible pairs mixing either cues from the same cue category (e.g., FP1 + FP2 from the uFP group) or cues from different categories (e.g., MA1 + FP3 from the uMA and bFP groups). We also included the four combinations consisting of cue triples presented in the training phase as a sanity check for participants' recall (these combinations were excluded from our main analyses). Overall, in the test phase, each learner encountered in total 29 cue combinations, which were randomly selected from a total of 70 possible cue combinations. (The exact format and instructions used while administering the task are provided in Appendix S2 in the Supporting Information online, and the list of all test cue combinations is provided in Appendix S3.)

Finally, let us return to the question of why we adopted Kiełkiewicz-Janowiak and Pawelczyk's (2014) rule, whereby any subject combination that contains a masculine referent takes the masculine personal plural form. First, having the combination "MA1 + MA2 + MA3" associated with "mp" rather than "np" made it possible to have a balanced number of mp and np events both within the full task and within each block. This reduced the likelihood of any bias towards np emerging purely due to the design. Second, this allowed us to have more challenging combinations that better probe participants' learning, notably combinations intermixing feminine and masculine cues.

Analysis

From the learning task, data from three participants were discarded because they persistently chose the same response across the test phase (27 or more out of 29 responses; i.e., rate > 93%).[5] To analyze participants' choices and response times, we used generalized mixed-effects modeling. The data contained repeated measurements from the same participants and items on multiple trials, hence we added random effects for both participants and items (i.e., cue combinations in the test phase). We selected the random effects structure of the models by using a top-down strategy starting with all random intercepts and slopes and then removing higher-order random effects step by step based on Akaike information criterion scores. We ran the mixed-effects models in R (R Core Team, 2019) using the lme4 package; the *p* values were obtained using the lmerTest package based on Satterthwaite's approximations, and the model summary tables were generated using the sjPlot package. To determine statistical significance, we used an alpha level of .05. In the analysis of

response times, we used the Box–Cox method as implemented in the car package to transform the distribution to normality and facilitate statistical modeling.

Explicit Knowledge and Demographic Questionnaire

After completing the language learning task, participants filled out a questionnaire that asked them whether they used any explicit rules to decide when to use each of the two verb forms, and if they did, what these rules were. The questionnaire also collected information about participants' gender, their age, the languages they spoke (other than English), and their highest education level. A full list of the questions used in the questionnaire is provided in Appendix S4 in the Supporting Information online. We focus specifically on the role of age and gender in explaining any individual differences observed when fitting the R–W model to the data. This is because age and gender have been shown to affect both associative learning and second language acquisition. For example, Mutter et al. (2012) showed that cue–outcome associations are less likely to be acquired by older adults than by young adults. It is also well established that older adults are less effective at learning a second language than young adults (for a review, see Muñoz & Singleton, 2011) and experience more difficulties with language production (Burke & Shafto, 2004). Several studies have also reported that females show higher conditioning levels in associative learning tasks (Lonsdorf et al., 2015; Merz et al., 2018) and acquire language more effectively (Adani & Cepanec, 2019; van der Slik et al., 2015) than males.

Alongside the main learning task, we included a standard task of implicit learning ability and one of working memory (WM). We selected these two tasks because they capture salient properties of the learning setup: (a) the fact that no explicit instructions were given; and (b) the fact that the linguistic phenomenon can be considered discontinuous in that properties of (constellations of) the agent, which is mentioned first, determine which past ending will be used on the verb, which is mentioned second, so that some maintenance of agent-related information in memory is required. Since our measure of implicit learning ability is nonstandard and did not play a significant role in our models, we report on this task in Appendix S5 in the online Supporting Information.

Working Memory Task

Stimuli

To measure participants' WM capacity, we used a slightly modified version of the operation span test (Turner & Engle, 1989) used by Medimorec et al.

(2021). In each trial, participants were asked to retain a list of digits (between 1 and 9) presented one at a time. Each digit presentation lasted for 1 s and was followed by a simple mathematical operation that could be either correct or incorrect (50% of the mathematical operations were correct). Participants had to verify the veracity of the mathematical operation before the next digit could be displayed. At the end of each trial, they had to type in the digits in the same order in which they had been presented to them. The length of the digit lists increased gradually from two to eight, with each length repeated three times. The task, thus, consisted of 21 trials.

Analysis

We calculated each participant's WM span by first summing the number of correct items they recalled in the correct order and then *z*-transforming the obtained score. We excluded one participant whose WM score was discontinuous from the rest of the sample (their WM score was −4.3 standard deviations from the mean, whereas the second furthest WM score was −1.8 standard deviations from the mean).

Computational Modeling

The Rescorla–Wagner Equations

The R–W model (Rescorla & Wagner, 1972) describes computationally how the associations between cues and outcomes are established. In the context of our experiment, a cue is the Polish label and image of one of the human or animal characters appearing in the scene on a given trial, and an outcome is the verb form describing their common action. For example, the clause *Chłopiec, mężczyzna i małpa chodzili* ("The boy, the man, and the monkey were walking") has as cues *chłopiec*, *mężczyzna*, and *małpa*, and as outcome *chodzili*. In our case, the association weight (or strength) measures the tendency of a verb form to occur in the presence of a certain noun.

After encountering a clause, the learner updates the association weight between a cue c_i and an outcome o, depending on whether the cue and outcome appear in the sentence, using a delta-type correction rule:

$$w_t\left(c_i, \mathrm{o}\right) = w_{t-1}\left(c_i, \mathrm{o}\right) + \alpha\beta\delta_{t-1}$$

where:

$$\delta_t = \begin{cases} 0, \text{ if } c_i \text{ absent} \\ \lambda - \sum\limits_{c_j \text{ present}} w_{t-1}\left(c_j, \mathrm{o}\right), \text{ if } c_i \text{ present and } o \text{ present} \\ 0 - \sum\limits_{c_j \textit{ present}} w_{t-1}\left(c_j, \mathrm{o}\right), \text{ if } c_i \text{ present and } o \text{ absent} \end{cases}$$

The subscript t refers to the present trial, thus w_t (c_i, o) is the association strength between c_i and o at trial t. α and β denote the learning rates for the cue c_i and outcome o respectively. λ refers to the maximum associability to an outcome and is almost always set to 1.

Based on the equation, three cases determine how an association weight is adjusted:

1. If the cue is absent, we make no adjustment to the weight.
2. If both the cue and outcome are present, then this provides positive evidence that should strengthen the association weight, and the sum of the weights of the cues present in the current event is adjusted towards the maximum associability value.
3. If the cue is present but the outcome is not observed, then this provides negative evidence that should weaken the association weight, and the sum of weights is adjusted towards 0.

For the implementation of the model, we used the package that was developed as part of the study by Milin et al. (2020).

Predicting Choices From the Model

To generate a verb form choice (or in the model's terminology, an outcome) from the model given a certain set of cues, we first calculate the activation of each form by summing the association weights between the form and each of the relevant cues. The predicted response from the model is then the form having the highest activation. For example, if at a certain trial in the test phase, a scene contained a girl and a monkey, then the activations of the masculine plural (mp; *chodzili*) and nonmasculine plural (np; *chodziły*) forms are calculated as follows:

$$activ\,(mp) = w\,(\text{girl},\ \text{mp}) + w\,(\text{monkey}, \text{mp})$$

$$activ\,(np) = w\,(\text{girl},\ \text{np}) + w\,(\text{monkey}, \text{np})$$

where for the formulae, we used the final weights obtained at the end of the training phase and hence omitted the trial subscripts (no learning happens in the test phase). If *activ(np)* > *activ(mp)*, the model would predict the np form, and otherwise it would predict the mp form.

Model Fitting Procedure

In our simulations, we assumed that $\lambda = 1$ and $\beta = 1$ and considered the learning rate α as a free parameter to be estimated for each participant (henceforth,

whenever we refer to learning rate, we will always refer to the α parameter). Specifically, we ran 50 computer simulations per participant using grid-search for α ranging from .01 to .50. In each simulation, we programmed a virtual agent to behave according to the R–W model and presented it with the same training trials as the participant whose learning history we aimed to model. From the trained model, we then generated form choices for the same trials that the participant encountered in the test phase. We finally selected the learning rate (and hence the model) that maximized the match rate between the participant's observed responses and the model's predicted responses (i.e., the proportion of test items for which the model produced the same response as the learner). Due to the nonidentifiability of the best-fit model, where in some cases more than one learning rate value maximized the match rate, we selected the median learning rate as the best parameter.

Model Evaluation

To help explain participants' behavioral data, we derived an activation-based measure from the fitted R–W model, which we call activation support for an outcome. The measure aims to explain participants' form choices and response times, and is defined as the difference between the activation of the outcome of interest and the activation of the remaining outcome. For example, the activation support for the nonmasculine plural form (np) is given by the following:

$$activation\ support\,(np) = activ\,(np) - activ\,(mp)$$

We hypothesized that the higher the activation support for a verb form (i.e., the stronger the evidence from the model supporting the verb form relative to the other possible form), the higher the likelihood of that form being selected by participants. We also expected that the magnitude of the activation support would negatively correlate with participants' response times. In other words, the higher the magnitude of this measure, the quicker the participant's response would be. This should translate into a quadratic relationship between activation support and response times, with the slowest responses expected when activation support values are near zero, and the fastest responses expected for high positive or negative values.

Results

This section evaluates the extent to which the R–W model explains our participants' behavior by fitting a separate model to each participant's data, and tests whether the model fit quality is affected by individual differences such as WM span, age, and gender. We first present some descriptive results on the

association weights of the fitted models, which summarize the linguistic knowledge participants acquired in the language learning task. Next, we compare the quality of fit of the model with that of other plausible, yet rule-based response strategies. We then successively present analyses that assess the model's capacity to recover participants' language choices, time latencies, and levels of response agreement. The effect of cognitive and personal characteristics on the extent to which the R–W model captures language learning is analyzed at the end of the Results section.

Learned Noun–Verb Form Association Weights

Following the fit procedure described earlier in the section on computational modeling, we selected the model that best captured the choices each participant made over trials, by finding the "right" learning rate parameter (see Appendix S6 in the online Supporting Information). Each participant was characterized mostly by two regimes of model fit accuracy: one for learning rates ranging roughly between .05 and .11, and one for learning rates between .12 and .50 (with some exceptions, as for Participants 12, 19, 27, and 35, for whom there were three regimes of accuracy), with neither of the two regimes consistently leading to better model fit accuracy. Taken together, although the explained variability in choices contributed by the learning rate parameter was limited, making by-participants adjustments for that parameter was still beneficial and insightful: We observed that there was not a single learning rate value that led to the highest model fit accuracy for all participants. In other words, there appear to be considerable individual differences in the rate of learning. Figure 2 depicts the distributions of the acquired association weights of all possible noun–verb form pairs from the best-fitting models.

Overall, the distributions of association weights were similar within each cue category (e.g., MA1, MA2, and MA3 within the uMA category), reinforcing our grouping of the cues based on the grammatical gender and animacy of the nouns they represent. Secondly, and unsurprisingly, the (unblocked) masculine cues gained a positive association weight with the masculine plural form (i.e., these cues are more likely to result in a choice for the masculine plural form), whereas the unblocked feminine cues gained a positive association weight with the nonmasculine form (i.e., these cues are more likely to result in a choice for the nonmasculine plural form). The magnitudes of the weights also differed between participants for most of the cues, thus creating a potential tool for capturing individual differences in our data.

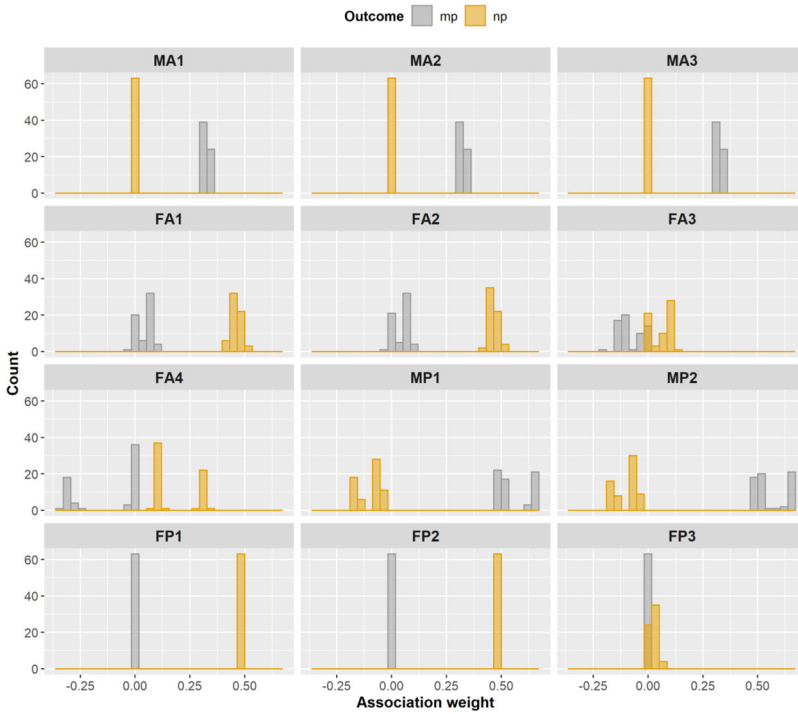

Figure 2 Histograms of cue–outcome association weights from the best-fitted Rescorla–Wagner models. F = feminine, M = masculine, A = animate, P = personal, mp = masculine plural verb form, np = nonmasculine plural verb form.

As predicted by (standard) blocking, the association weights between the feminine blocked cues (i.e., FA3 and FP3) and the nonmasculine form were more centered around zero than their unblocked counterparts (see the panes for FA3 and FP3 in Figure 2). The blocking, however, is not yet fully reflected in the acquired weights, since for many participants, the association weights between the blocked feminine cues and the nonmasculine form were different from zero. An inhibitory blockinglike effect (i.e., a negative weight between FA4 and the masculine plural form) appeared for about a third of the participants. For the remaining participants, FA4 was more like a standard blocked cue as its association weight with the masculine form was around zero. Taken together, blockinglike effects showed tendencies in the predicted directions. We assumed that their relatively mild magnitude was because our experiment captured early phases of learning, where expositions of the stimuli were

repeated only 15 times. This hypothesis was confirmed by rerunning the simulations presented in Figure 2, now with 1,000 repetitions per event, as shown in Appendix S7 in the Supporting Information online; blocking and inhibitory blocking effects occurred for all participants regardless of their learning rate or event ordering. These results confirm what we pointed out before: Biases and differences in learning are more likely to manifest early on in learning (Ellis, 2006a).

Participant–Model Match Rates

Next, we investigated to what extent these differences in learning can be captured by the R–W model if we take into account the order of events encountered by each participant as well as differences in their learning rates. The model's fit accuracy (i.e., the proportion of matches between the responses from a given participant and its best fitting R–W model) ranged from .24 to 1.00 ($M = .68$, $SD = .17$): 17 out of 63 participants had a fit accuracy $\geq .80$, and only nine participants had a proportion of matches lower than .50. Evaluating the model fit using leave-one-out cross validation[6] shows that model fit accuracy was equally high on unseen data, with an average accuracy of .68 ($SD = .17$) and 17 out of 63 participants reaching a fit accuracy $\geq .80$. The fit accuracy rates were highest for events containing a masculine personal cue ($M = .74$) or an unblocked feminine personal cue ($M = .68$), and they were lowest for events containing the inhibitory blocked cue ($M = .61$) or an animate cue (all means $\approx .65$).

These results suggest a reasonably good fit of the R–W model to participants' data, given that we considered a simple strategy for generating response predictions based on the model activations—that is, for each event, we selected the verb form that had the highest activation regardless of the difference in the activation magnitudes of the two possible verb forms. We will later analyze the sensitivity of the fitted models' activations to the observed form choice proportions and response times.

Comparison Between the Rescorla–Wagner Model and Other Decision Strategies

The results presented above show that the R–W model captures our participants' behavior reasonably well, but how does the model compare to other strategies that participants might have employed during the experiment? To answer this question, we considered four decision strategies. The first two are the prescriptive and normative strategies we presented earlier. The prescriptive strategy is the one described, or prescribed, by Polish grammar books,

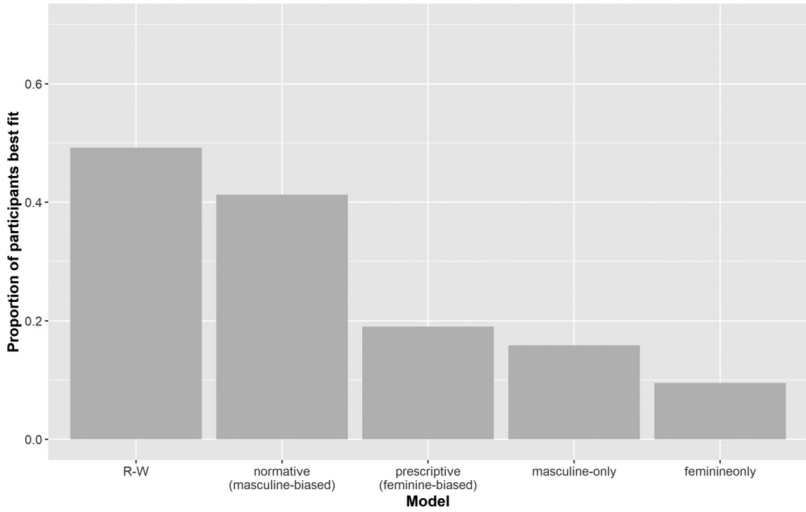

Figure 3 Proportion of participants that each model fitted the best. R–W = Rescorla-Wagner.

and whereby a participant always chooses the nonmasculine verb form except when a masculine personal cue is present (we also refer to this strategy as the "feminine-biased" strategy). The normative strategy is the one generally adopted by native speakers of Polish, whereby the masculine verb form is always selected except when all cues are feminine (referred to as the "masculine-biased" strategy). We also included two basic strategies, whereby a participant either always chooses the masculine verb form (referred to as the "masculine-only" strategy) or always chooses the nonmasculine verb form (referred to as the "feminine-only" strategy). The latter two strategies were included to capture participants' behavior at the extremes.

Figure 3 displays the proportion of participants best fitted by each of the five resulting models (R–W and our four decision strategies); we considered the model(s) with the highest participant–model match rate among the five models as the best-fit model(s). The R–W model was the model that best explained participants' responses (31 out of 63 participants), followed closely by the normative strategy (26 participants). The other three strategies explained participants' choices substantially less well than those two strategies (< 12 participants). The fact that the R–W model and the normative strategy were close in capturing participants' behavior is not very surprising since the verb forms used in the training events were selected based on the

Table 3 Fixed effects structures of the (generalized) linear mixed-effects models explaining participants' nonmasculine plural form choices (left) and response times (right) based on activation support from the fitted Rescorla–Wagner models

	Nonmasculine plural choice			Response times (transformed)		
Predictors	*OR*	95% CI	*p*	*b*	95% CI	*p*
(Intercept)	0.75	[0.60, 0.95]	.016	0.08	[−0.05, 0.21]	.244
Activation support(np)	6.78	[3.82, 12.03]	< .001	−0.02	[−0.14, 0.09]	.685
Activation support(np) squared				−0.20	[−0.35, −0.04]	.012

Note. $\alpha = .05$.

normative rules and the predictions of the R–W model were largely in accordance with the normative strategy (Figure 3). It is interesting, though, that the R–W model managed to learn this strategy implicitly without any prior experience based on a simple general learning rule. The average percentage of response matches between the R–W model and the normative strategy per participant was above 90%, and the average percentage of response matches between the R–W model and the prescriptive strategy was above 85%.

Relationship Between the Model's Activation-Based Measures and Participants' Choices and Response Times

In a further assessment of the quality of fit of the R–W model, we carried out a generalized linear mixed-effects modeling analysis looking at the relationship between participants' choices and the activation-based measure derived from the fitted R–W models—that is, activation support for the np form (see the section on computational modeling for more details). We also analyzed the relationship between participants' response times and activation support by fitting a polynomial linear mixed-effects model with both the linear and quadratic terms of activation support, since we expected activation support to have a quadratic effect on response times (Table 3). More detailed summaries of the models with the random effects structures are provided in Appendix S8 in the Supporting Information online.

As expected, activation support for the nonmasculine plural form was significantly positively associated with the likelihood of nonmasculine form

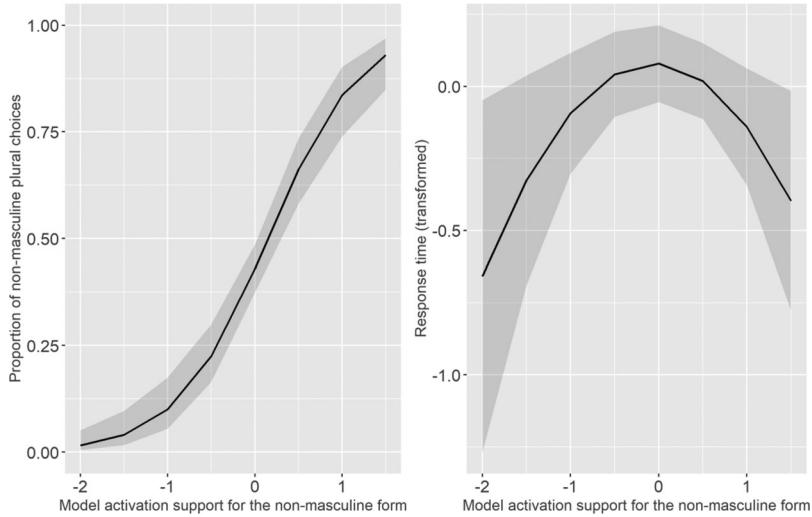

Figure 4 Relationship between the Rescorla–Wagner activation support for the nonmasculine plural form and the proportion of nonmasculine plural choices made by participants (left), and relationship between the Rescorla–Wagner activation support for the nonmasculine plural form and participants' response times (right).

choices, $OR = 6.78$, $p < .001$, 95% CI [3.82, 12.03]. Figure 4 (left pane) also shows that this relationship is asymmetrical around 0, reflecting a strong bias towards the masculine verb form that, even with (activation-based) evidence supporting the nonmasculine form, can still lead to a preference for the masculine form. Also, and in line with our hypotheses, the second-order polynomial term of activation support was a significant predictor of response time, as there was a quadratic relationship between the activation support and response time, with the slowest responses recorded for the least supported events, $b = -0.20$, $p = .012$, 95% CI [−0.35, −0.04]; see also Figure 4, right pane.

These results suggest that the fitted models performed well in predicting participants' form choices, and that the information encoded in the association weights—the basic currency of a R–W model—is a good predictor of both the likelihood of choosing a particular verb form and the speed with which the response is made. Participants' level of agreement regarding the choice of a certain verb form thus differed depending on the activation support of that particular form, with a high level of agreement expected and attested for high positive or low negative activation support values and with a

high level of disagreement expected and attested for activation support around zero.

Level of Agreement Between Participants Through the Lens of the Model

We further analyzed participants' behavior by exploring two additional questions. First, what level of agreement was there among language learners, given a particular type of event (e.g., events made up of cues from the same grammatical gender versus events intermixing cues of different grammatical gender)? Second, and crucially, can the differences in levels of agreement be adequately explained using the R–W model?

To answer these questions, we analyzed the effect of the presence of each event category on the proportion of participants who chose one of the two verb forms (for a full list of event categories used in the task, see Appendix S3 in the Supporting Information online). To obtain the model estimates of the proportions, we used the generalized linear mixed-effects model that we built in the previous analysis (Table 3, left) to model the relationship between activation support and the choice of the nonmasculine plural form. Specifically, for each event category, we averaged the model's predicted proportions of the nonmasculine forms based on its activation support values across participants (given that each event category was encountered once by each participant). The results are summarized in Figure 5. The events in the figure are sorted in ascending order by observed choice proportions in the experiment, so the leftmost and rightmost sides represent regions of high level of agreement between participants, whereas events situated in the middle part triggered a high level of disagreement between participants.

Participants had a clear preference for the masculine plural form when the event contained a masculine personal cue (uMP) or when the event included only animate masculine cues (i.e., "uMA1 + uMA2"). Likewise, participants clearly preferred the nonmasculine plural form when the events solely contained feminine cues, whether personal or animate (e.g., "uFA + uFP"). High levels of disagreement were mainly observed for events intermixing the inhibitory blocked cue and a feminine cue or a masculine animate and a feminine personal cue (with proportions ranging between .44 and .51).

Comparing the observed and predicted proportions, the model managed to capture the difference in the levels of agreement between participants across the different categories of events surprisingly well. The largest discrepancies between the observed and predicted proportions appear to have occurred for events involving the blocked and inhibitory blocked cues, suggesting that it

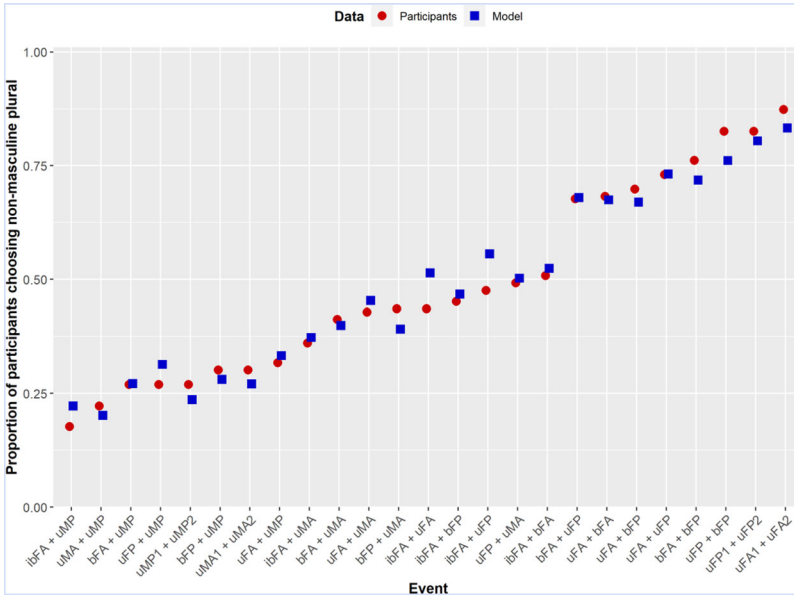

Figure 5 Observed and predicted proportion of participants choosing the nonmasculine plural form given a certain cue combination. F = feminine, M = masculine, A = animate, P = personal, b = blocked cue, u = unblocked cue, ib = inhibitory blocked cue.

was more challenging for the model to capture their effect on participants' choices.

Relationship Between Model-Fit Quality and Individual Difference Measures

The model fit results demonstrated that the R–W model successfully accounts for the behavior of a large proportion of our participants, including their response times. However, the quality of fit varied across participants, with data from a few participants fitted very poorly by the model. In an attempt to assess whether the other measures collected during the experiment (demographic characteristics, WM span, and implicit learning) can explain the observed individual differences in the model fit quality, we ran a multiple linear regression to predict participant–model match rates (logit-transformed) based on the different individual difference measures collected. Specifically, the predictors included WM span (*z*-transformed), gender, and age; the time-slopes extracted for each participant from the implicit learning task did not make a significant

Table 4 Summary of the linear regression model assessing the effect of working memory (WM) span and gender on the proportion of participant–model matches

	Participant–model match rate (transformed)		
Predictors	*b*	95% CI	*p*
(Intercept)	0.59	[0.44, 0.75]	< .001
WM *z* score	0.17	[0.05, 0.30]	.008
Gender = male	−0.28	[−0.54, −0.02]	.033
Observations	61		
R^2/R^2 adjusted	.148/.118		

Note. $\alpha = .05$.

contribution and are reported on in Appendix S5 in the Supporting Information online.

From the full linear regression model that included all four individual measure variables as well as the interaction between gender and WM span, we derived a final model containing only the significant variables by using backward variable selection based on the likelihood ratio test. Data from two participants were omitted from this analysis because one did not report their gender and one was identified as an extreme outlier, as explained in the section on the WM task. In total, data from 61 participants were fed to the linear regression model. The best model after variable selection included both WM span and gender, but not their interaction (Table 4).

The model fit accuracy increased significantly with increasing WM span, $b = 0.17$, $p = .008$, 95% CI [0.05, 0.30], as illustrated in Figure 6 (left pane), and was significantly higher for female participants ($M = .69$, $SD = .17$) in comparison with male participants ($M = .63$, $SD = .16$), $b = -0.28$, $p = .033$, 95% CI [−0.54, −0.02], as shown in Figure 6 (right pane). To check that removing influential residuals did not affect our findings, we applied model criticism as described in Baayen and Milin (2010); excluding the single extreme residual with a z score greater than 2.5 in our model resulted in even stronger effects ($p = .004$ for WM span and $p = .007$ for gender). Also, WM span did not significantly correlate with gender ($p = .058$), which suggests that their relationship is unlikely to have influenced the effects of gender and WM span on model fit accuracy (see Appendix S9 in the Supporting Information online for more detail). Our findings thus show that having a larger WM capacity or being female increased the likelihood of a participant

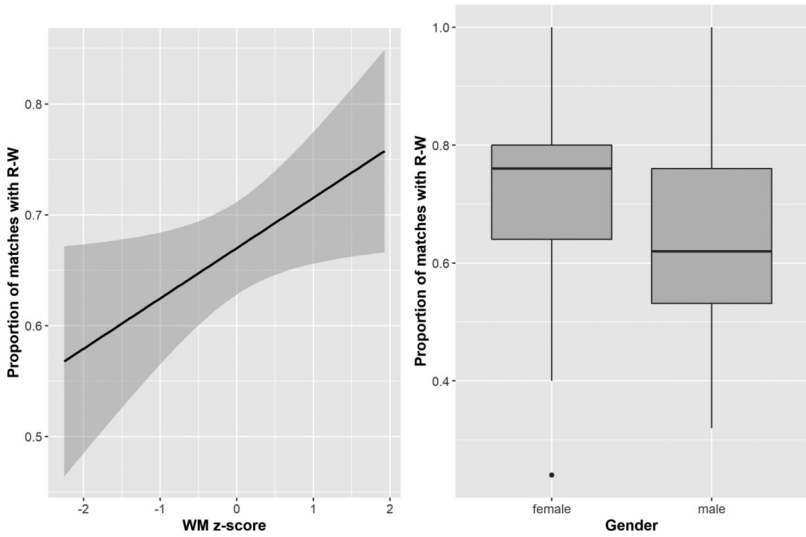

Figure 6 Effect of a participant's (*z*-transformed) working memory (WM) span (left) and gender (right) on the rate of matches between their responses and the predicted responses from their best-fitting model. R–W = Rescorla–Wagner.

choosing verb forms in accordance with the R–W model in our language learning task.

Discussion

Summary of Findings

Our findings show that a R–W mechanism captures well how participants learn subject–verb agreement in a morphologically complex language and, by extension, how they might learn language through mere exposure to it. With an average fit accuracy of 68%, based on a simple activation-based decision strategy, the model explained the verb form choices made by a large proportion of participants rather well.[7] More interestingly, an activation-based measure extracted from the best-fitting models correlated strongly with both the likelihood of a particular verb form choice and the time required to make that choice.

The model also provided insights as to why participants might display high or low agreement levels when choosing a verb form, depending on the nature of the subject of the clause. According to the model, this is due to the association strengths that the participants acquire, which are used to calculate the

activation support for each of the possible verb forms. These association strengths are mostly affected by (a) the learner's learning rate for the cues (the learning rate determines the magnitude of the correction of the weights, based on the estimated error in each trial) and (b) the distribution of the learning events they encountered during the learning stage (this would include the frequency of each learning event and the order of the learning events, among other things). Thus, one prediction from our study is that changing the order or the relative frequencies of the learning events during the training might lead to different choice patterns from those we observed here.

We also found a significant relationship between both gender and WM capacity and the participant–model match rates, which sheds light on what might have driven the observed differences in the quality of model fit. The fact that in our experiment a larger proportion of women than men acted in accordance with a R–W mechanism is in line with findings from several previous studies that highlighted the association between gender and classical conditioning for both humans (Lonsdorf et al., 2015; Merz et al., 2018) and animals (e.g., Velasco et al., 2019). This suggests a significant difference in learning between men and women, with women being better modeled by the R–W error-correction learning rule. Women are generally known to have a small language advantage over men (see Kimura, 1999, for an extensive assessment), specifically in areas pertaining to lexical retrieval (Balling & Baayen, 2008, 2012). It has been suggested that this might be due to women having a superior declarative memory, which they could use to generalize over stored neighboring forms (Hartshorne & Ullman, 2006).

The finding that the likelihood of a language learner behaving according to the R–W mechanism increases with WM capacity provides evidence that WM can play a role in classical conditioning by affecting the adoption of a classical conditioning mechanism such as the R–W rule. Sasaki (2009) and Baetu et al. (2018) previously provided evidence of disruption of classical conditioning performance when WM is loaded using dual-task paradigms. The present finding adds to the mounting evidence that, against the predominant belief, WM may be implicated in low-level cognitive processes such as instrumental learning, more commonly referred to as reinforcement learning within the neuroscience and machine learning communities (Collins & Frank, 2012; Ez-zizi, 2016) and in some forms of implicit learning (Medimorec et al., 2021).

Blocking and inhibitory blocking-like effects did not emerge from the R–W model for all participants. As shown, this was mainly due to the short duration of the training phase. Increasing the number of training trials not only resulted

in the reemergence of blocking effects for all participants' best-fitting parameters, but also removed the variability in the association weights, thus predicting that all participants should end up behaving in the same way in the long run (see Appendix S7 in the Supporting Information online). This is not surprising in light of Danks' (2003) work, which shows that, in many cases, the R–W system will converge to the same equilibrium regardless of the parameters used. In other words, the destination of learners is often the same (this is also desirable since we often want all learners to learn to make the same associations), but the paths to those destinations can differ substantially depending on the nature of the learning problem (e.g., amount of data available, relative frequencies of the events, number of cues and outcomes). This has many implications for language learning studies employing the R–W model, as discussed in the next section.

Implications for Language Learning Studies That Use the Rescorla–Wagner Model

Several language research groups have embraced the R–W model as a valuable approach for modeling language learning phenomena due to its simplicity, its cognitive plausibility, and its successes in explaining a wide range of language learning phenomena. The present study provides further support for simulating or modeling language learning using this model, but also draws attention to the important issue of individual differences, which has so far been overlooked in studies that combine computational modeling using the R–W model and experimental data.

Given the amount of individual difference we observed in our data, it would be prudent to move away from the currently predominant approach where the R–W model is run once with its default parameter values and used to explain data from all participants. Although the effect of the learning rate parameter on model fit accuracy was not substantial for the chosen task, in practice, R–W performance will always be affected by the choice of learning rate, irrespective of the particular modeling challenge (Milin, Divjak, & Baayen, 2017, pp. 1739–1741).

Another common practice that might need to be reconsidered is training the model on one large dataset or a small subset of it. Consequently, the features that set the model apart from purely statistical classification models—namely, the possibility of choosing parameters that capture how fast an individual can learn, and the ability to account for input order effects—remain unused. It is not surprising, then, that several studies have reported only minor or no differences between the R–W model and other statistical or learning models such as

logistic regression, memory-based learning, and decision trees (Baayen, 2011; Baayen et al., 2013). With such an approach, the main advantage of the R–W model over purely statistical techniques is its ability to perform incremental learning, as data become available. However, the same advantage could be achieved from any neural network model with no hidden layer as such a model can also produce weights between cues and outcomes, albeit with a different and, arguably, less plausible learning rule (here we mainly allude to the back-propagation learning rule, which is currently the predominant approach when training neural network models). With such a model, the same model fitting approach we used here can still be applied but with a different set of parameters to tune, such as the type of activation function, number of neurons, and learning rate.

Recent work in usage-based frameworks has highlighted the vast individual differences characterizing language knowledge in first language populations. Individual differences in grammar have been found to be comparable in size to those in lexical knowledge and are related to both the quality of the input and the learner's cognitive abilities (Dąbrowska, 2018). Our findings demonstrate that individual model fitting should be the default option when comparing the R–W model or other computational models to participants' data. Specifically, model parameters and data inputs should be adjusted separately for each participant to allow for a better account of individual responses and obtain a more veridical picture of where knowledge of, for example, a rule resides, whether in the aggregated mind of the linguist or in the individual minds of the users (Divjak, 2018), and how it is distributed across the population. Our data also support the usage-based stance that our linguistic knowledge is shaped by our personal and cognitive characteristics, as attested by the significant role of WM and gender in the quality of model fit; such factors should be considered by default when modeling language.

Limitations and Future Directions

Our study is the first to fit the R–W model to the behavior of individual learners in an actual language learning task. We used the R–W model in the form available now, but our findings, despite being very promising, show that it might be interesting to extend the model to handle WM capacity limitations. Further investigations in such a direction could be inspired by work done in reinforcement learning—a closely related field to classical conditioning—where learning has also long been assumed to occur in areas, often associated with low-level cognitive functions, such as the dorsal and ventral striatum (Balleine & O'Doherty, 2010), but is now recognized to also involve high-level cognitive

control via WM. This has led to the development of new learning frameworks where WM is explicitly modeled as a key component that supports learning by retaining information from previous trials (e.g., see Collins & Frank, 2012; Ez-zizi et al., 2015). This could be the approach to take for R–W and other classical conditioning models, especially because in large simulation-based language studies, learning events typically contain a large number of cues (e.g., all trigraphs or words in one sentence), which cannot be processed at once by a human learner—as is required in the updates of the R–W model—due to known WM capacity limitations (see Glautier, 2013, and Baayen et al., 2016, for early attempts in this direction).

Another direction for future extension of our work is to collect participants' responses over time while they are trained on the cue–outcome associations rather than having a separate postlearning test phase. This would have the potential to improve the model fit further and to provide a broader picture of the behavior of participants while they are learning the task. In addition, this could allow the extraction of a learning measure based on time slope for the language learning task, such as we did for the implicit learning task, and thus would increase the likelihood of finding a link between implicit learning and the quality of fit of the R–W model (see also our discussion in Appendix S5 in the Supporting Information online). A link between the two measures can also be probed by fitting the R–W model to the response times collected in the implicit learning task, as was done in Notaro et al. (2018), rather than using time slopes only or a mixture of the two.

Finally, the particular structure of our language learning task favored the normative (masculine-biased) strategy, but an interesting question that remains unanswered is whether we can use the R–W model to predict the emergence of different strategies as we vary the structure of the language input and control for individual differences among language learners. The approach of using the R–W model to explain or predict the level of agreement among language users can be extended beyond Polish subject–verb agreement in the plural past tense to cover other facets of language where a lack of consensus in language use has been observed (e.g., see Geeraert et al., 2020; Milin, Divjak, & Baayen, 2017).

Conclusion

The R–W model is a very simple learning model, yet it has multiple sources of variation that can be used to explain participants' behavior in language learning experiments. These include the model's learning rate, the order of presentation of learning examples, and the relative frequencies of cue–outcome

cooccurrences. In the present study, we systematically incorporated these sources of variation when fitting the model to participants' data, thus enabling the model to successfully capture the choices and response latencies of most participants in a language learning task on Polish subject–verb agreement. In addition, cognitive and demographic characteristics such as WM and gender determined the extent to which language learning was driven by R–W-like learning principles.

Final revised version accepted 26 January 2023

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at https://github.com/ooominds/Error-correction-mechanisms-in-language-learning and https://doi.org/10.25500/edata.bham.00000911. All proprietary materials have been precisely identified in the manuscript.

Notes

1 The idea of cue competition is also at the core of the competition model of Bates and MacWhinney (1987) for language acquisition. Their model, however, uses mainly symbolic/linguistic cues such as word order or morphological features of words and is based on a connectionist approach requiring a much more complex architecture than the R–W model.

2 The contents of any corpus are, at best, a very rough approximation of the input that language users receive. Conversely, artificial languages are illustrative and informative for understanding natural languages but hardly a realistic reflection of the complexity found in any given natural language.

3 The early implementations of the R–W rule as the naïve discrimination learning model relied on a noniterative version of the algorithm, as provided by Danks (2003), which eliminates the possibility of any order effects emerging.

4 It is important to note that here we were not interested in testing the blocking effects per se as is typically done in behavioral experiments of classical conditioning. In those experiments, only the events relevant to blocking are included (blocking is tested separately from the other effects), and blocking is tested on a second cue rather than a third cue as in our case (e.g., Kamin, 1969). Also, the learner is usually trained for long enough to ensure that the "blocking" cue becomes a good predictor of the outcome of interest. Such a clean experimental setup would not

fairly represent the "disarray" so pervasive in natural languages. As our study is about language learning, we opted to mimic a realistic learning situation as closely as possible.

5 An additional participant experienced equipment malfunction in the middle of the test phase and had to retake the test. We retained this participant's data since they did not go through any extra training in the second run of the test phase and thus started the new run with the same knowledge as in the first run (recall that the outcome feedback is only provided in the training phase). The familiarization effect should not play a role here since all participants underwent a few practice trials in all phases before starting the actual experiment. We also reran all analyses without this participant's data and confirmed that removing them did not alter any of the results presented in the paper.

6 The quality of model fit on unseen data was evaluated using leave-one-out cross-validation. Specifically, for each participant, we held out one event of those they encountered in the test phase and fitted a R–W model to the remaining events. The model was then evaluated only on the reserved event by assessing whether the response from the participant matched that of its best-fitting R–W model. We repeated this for all test events and then computed the average (leave-one-out) fit accuracy, that is, the proportion of matches between the responses from a given participant on each of the reserved events and their associated predictions from the participant's best-fitting R–W model.

7 In fact, this accuracy level can be considered excellent since we assumed a very simple nonprobabilistic action selection process where a verb form is chosen if it has the highest activation. This does not take into account the variability that might arise from exploration, lapse of attention, or inherent brain noise.

References

Adani, S., & Cepanec, M. (2019). Sex differences in early communication development: Behavioral and neurobiological indicators of more vulnerable communication system development in boys. *Croatian Medical Journal*, *60*(2), 141–149. https://doi.org/10.3325/cmj.2019.60.141

Ambridge, B., & Lieven, E. V. M. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada*, *11*(2), 295–328. https://doi.org/10.1590/S1984-63982011000200003

Baayen, R. H., Endresen, A., Janda, L. A., Makarova, A., & Nesset, T. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, *37*(3), 253–291. https://doi.org/10.1007/s11185-013-9118-6

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. https://doi.org/10.21500/20112084.807

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481. https://doi.org/10.1037/a0023851

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, *31*(1), 106–128. https://doi.org/10.1080/23273798.2015.1065336

Baetu, I., Burns, N., & Child, B. (2018). *Individual differences in working memory capacity predict performance on an associative learning task* [Paper presentation]. Australian Psychologist, Sydney, Australia. https://doi.org/10.1111/ap.12372

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69. https://doi.org/10.1038/npp.2009.131

Balling, L. W., & Baayen, R. H. (2008). Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes*, *23*(7–8), 1159–1190. https://doi.org/10.1080/01690960802201010

Balling, L. W., & Baayen, R. H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition*, *125*(1), 80–106. https://doi.org/10.1016/j.cognition.2012.06.003

Bandura, A. (1962). Social learning through imitation. In M. R. Jones (Ed.), *Nebraska Symposium on Motivation* (pp. 211–274). University of Nebraska Press.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language aquisition* (pp. 157–193). Lawrence Erlbaum.

Burke, D. M., & Shafto, M. A. (2004). Aging and language production. *Current Directions in Psychological Science*, *13*(1), 21–24. https://doi.org/10.1111/j.0963-7214.2004.01301006.x

Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, *22*(2–4), 381–410. https://doi.org/10.1515/tlir.2005.22.2-4.381

Chen, Z., Haykin, S., Eggermont, J. J., & Becker, S. (2008). *Correlative learning: A basis for brain and adaptive systems*. Wiley.

Chomsky, N. (1959). A review of BF Skinner's Verbal behavior. *Language*, *35*(1), 26–58. https://doi.org/10.4159/harvard.9780674594623.c6

Chuang, Y.-Y., Bell, M. J., Banke, I., & Baayen, R. H. (2021). Bilingual and multilingual mental lexicon: A modeling study with linear discriminative learning. *Language Learning*, *71*(S1), 219–292. https://doi.org/10.1111/lang.12435

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *The European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460–9568.2011.07980.x

Dąbrowska, E. (2018). Experience, aptitude and individual differences in native language ultimate attainment. *Cognition*, *178*, 222–235. https://doi.org/10.1016/j.cognition.2018.05.018

Dąbrowska, E., & Divjak, D. (2015). Introduction. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 1–9). Walter de Gruyter.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121. https://doi.org/10.1016/S0022-2496(02)00016-0

Divjak, D. (2018). Binding scale dynamics. In D. Van Olmen, T. Mortelmans, & F. Brisard (Eds.), *Aspects of linguistic variation* (pp. 9–42). De Gruyter Mouton.

Divjak, D. (2019). *Frequency in language: Memory, attention, and learning*. Cambridge University Press.

Divjak, D., & Gries, S. T. (Eds.). (2012). *Frequency effects in language representation*. De Gruyter.

Divjak, D., Milin, P., Ez-zizi, A., Józefowski, J., & Adam, C. (2021). What is learned from exposure: An error-driven approach to productivity in language. *Language, Cognition and Neuroscience*, *36*(1), 60–83. https://doi.org/10.1080/23273798.2020.1815813

Ellis, N. C. (2006a). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24. https://doi.org/10.1093/applin/ami038

Ellis, N. C. (2006b). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194. https://doi.org/10.1093/applin/aml015

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*. Wiley-Blackwell.

Ellis, N. C., & Sagarra, N. (2010). The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition*, *33*(4), 553–580.

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition*, *33*(4), 589–624. https://doi.org/10.1017/S0272263111000325

Ez-zizi, A. (2016). *Reinforcement learning in partially observable tasks: State uncertainty and memory dependence* [Doctoral thesis, University of Bristol]. EThOS. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.707711

Ez-zizi, A., Farrell, S., & Leslie, D. (2015). Bayesian reinforcement learning in Markovian and non-Markovian tasks. *IEEE Symposium Series on Computational Intelligence* (pp. 579–586). https://doi.org/10.1109/SSCI.2015.91

Geeraert, K., Newman, J., & Baayen, R. H. (2020). Variation within idiomatic variation: Exploring the differences between speakers and idioms. *East European Journal of Psycholinguistics*, *7*(2), 9–27. https://doi.org/10.29038/eejpl.2020.7.2.gee

Glautier, S. (2013). Revisiting the learning curve (once again). *Frontiers in Psychology*, *4*, Article 982. https://doi.org/10.3389/fpsyg.2013.00982

Gries, S. T., & Divjak, D. (Eds.). (2012). *Frequency effects in language learning and processing*. De Gruyter.

Harpo Software. (2018). Speech2Go, Ivona, Nuance—Voices to Go—Text-to-Speech. http://speech2go.net

Hartshorne, J. K., & Ullman, M. T. (2006). Why girls say 'holded' more than boys. *Developmental Science*, *9*(1), 21–32. https://doi.org/10.1111/j.1467-7687.2005.00459.x

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behaviour* (pp. 279–296). Appleton-Century-Crofts.

Kiełkiewicz-Janowiak, A., & Pawelczyk, J. (2014). Language and gender research in Poland. In S. Ehrlich, M. Meyerhoff, & J. Holmes (Eds.), *The handbook of language, gender, and sexuality* (2nd ed., pp. 353–377). Wiley-Blackwell.

Kimura, D. (1999). *Sex and cognition*. MIT press.

Klavan, J., & Divjak, D. (2016). The cognitive plausibility of statistical classification models: Comparing textual and behavioral evidence. *Folia Linguistica*, *50*(2), 355–384. https://doi.org/10.1515/flin-2016-0014

Langacker, R. W. (1987). *Foundations of cognitive grammar: Vol. 1. Theoretical prerequisites*. Stanford University Press.

Lonsdorf, T. B., Haaker, J., Schümann, D., Sommer, T., Bayer, J., Brassen, S., Bunzeck, N., Gamer, M., & Kalisch, R. (2015). Sex differences in conditioned stimulus discrimination during context-dependent fear learning and its retrieval in humans: The role of biological sex, contraceptives and menstrual cycle phases. *Journal of Psychiatry and Neuroscience*, *40*(6), 368–375. https://doi.org/10.1503/140336

Mathôt, S., & March, J. (2022). Conducting linguistic experiments online with OpenSesame and OSWeb. *Language Learning*, *72*(4), 1017–1048. https://doi.org/10.1111/lang.12509

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: an open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Medimorec, S., Milin, P., & Divjak, D. (2021). Working memory affects anticipatory behavior during implicit pattern learning. *Psychological Research*, *85*, 291–301 (2021). https://doi.org/10.1007/s00426-019-01251-w

Merz, C. J., Kinner, V. L., & Wolf, O. T. (2018). Let's talk about sex… differences in human fear conditioning. *Current Opinion in Behavioral Sciences*, *23*, 7–12. https://doi.org/10.1016/j.cobeha.2018.01.021

Milin, P., & Blevins, J. P. (2020). Paradigms in morphology. In *Oxford Research Encyclopedia of Linguistics*. https://doi.org/10.1093/acrefore/9780199384655.013.551

Milin, P., Divjak, D., & Baayen, R. H. (2017). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1730–1751. https://doi.org/10.1037/xlm0000410

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS One*, *12*(2), Article e0171935. https://doi.org/10.1371/journal.pone.0171935

Milin, P., Madabushi, H. T., Croucher, M., & Divjak, D. (2020). *Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences*. arXiv:2003.03813. https://doi.org/10.48550/arXiv.2003.03813

Muñoz, C., & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*(1), 1–35. https://doi.org/10.1017/S0261444810000327

Mutter, S. A., Atchley, A. R., & Plumlee, L. M. (2012). Aging and retrospective revaluation of causal learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 102–117. https://doi.org/10.1037/a0024851

Notaro, G., van Zoest, W., Melcher, D., & Hasson, U. (2018). *Prediction and information integration determine subtle anticipatory fixation biases*. bioRxiv. https://doi.org/10.1101/252809

Olejarczuk, P., Kapatsinski, V., & Baayen, R. H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard*, *4*(s2), Article 20170020. https://doi.org/10.1515/lingvan-2017-0020

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552.

Pirrelli, V., Marzi, C., Ferro, M., Cardillo, F. A., Baayen, H. R., & Milin, P. (2020). Psycho-computational modelling of the mental lexicon. In V. Pirrelli, I. Plag, & W. U. Dressler (Eds.), *Word knowledge and word usage* (pp. 23–82). De Gruyter Mouton.

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*(4–5), 445–485. https://doi.org/10.1080/016909600501196 61

R Core Team. (2019). *R: A language and environment for statistical computing* (Version 3.6.2) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*(6), 927–960. https://doi.org/10.1080/03640210701703576

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, *43*(3), 151–160. https://doi.org/10.1037/0003-066X.43.3.151

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

Sasaki, T. (2009). The role of the central executive in associative learning. *Psychologia*, *52*(1), 80–90. https://doi.org/10.2117/psysoc.2009.80

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. https://doi.org/10.1037/0033-295x.96.4.523

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge University Press.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla–Wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–321. https://doi.org/10.3758/BF03210755

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century.

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts.

Sturdy, C. B., & Nicoladis, E. (2017). How much of language acquisition does operant conditioning explain? *Frontiers in Psychology*, *8*, Article 1918. https://doi.org/10.3389/fpsyg.2017.01918

Trimmer, P. C., McNamara, J. M., Houston, A. I., & Marshall, J. A. (2012). Does natural selection favour the Rescorla–Wagner rule? *Journal of Theoretical Biology*, *302*, 39–52. https://doi.org/10.1016/j.jtbi.2012.02.014

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

van der Slik, F. W. P., van Hout, R. W. N. M., & Schepens, J. J. (2015). The gender gap in second language acquisition: Gender differences in the acquisition of Dutch among immigrants from 88 countries with 49 mother tongues. *PloS One*, *10*(11), Article e0142056. https://doi.org/10.1371/journal.pone.0142056

Velasco, E. R., Florido, A., Milad, M. R., & Andero, R. (2019). Sex differences in fear extinction. *Neuroscience & Biobehavioral Reviews*, *103*, 81–108. https://doi.org/10.1016/j.neubiorev.2019.05.020

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods*, *45*(1), 108–115. https://doi.org/10.3758/s13428-012-0224-y

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. Distributions of Participants' Educational and Language Backgrounds.

Appendix S2. Experimental Procedure for the Language Learning Task.

Appendix S3. Cue Combinations Used in the Test Phase of the Language Learning Task.

Appendix S4. Explicit Knowledge and Demographic Questionnaire.

Appendix S5. Implicit Learning Task.

Appendix S6. Effect of Learning Rate Parameter on Model Fit Accuracy.

Appendix S7. Long-Run Simulations to Assess Blocking Effects.

Appendix S8. Generalized Linear Mixed-Effects Models Explaining Participants' Choices and Response Times.

Appendix S9. Correlations Between the Different Individual Difference Measures.