UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Base Rate Neglect in Linguistic Category Learning

Permalink
https://escholarship.org/uc/item/6nf8q6dp

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors
Pertsova, Katya
Moreton, Elliott
Fennell, Josh
et al.

Publication Date
2025

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Base Rate Neglect in Linguistic Category Learning

Katya Pertsova (pertsova@unc.edu) **Elliott Moreton** (moreton@unc.edu)
Department of Linguistics, University of North Carolina at Chapel Hill
Chapel Hill, NC 27599 USA

Joshua Fennell (josh@jfnl.xyz)
Independent Researcher
Raleigh, NC

Brandon Prickett (brandonlprickett@umass.com)
Department of Linguistics, University of Massachusetts
Amherst, MA 01002 USA

Abstract

This paper presents a categorization experiment supporting the hypothesis that base-rate-neglect occurs in linguistic category learning (and, thus, is a cross-domain phenomenon), and that it is more likely for those learners who engage in explicit problem-solving, rather than implicit learning. We find that among those participants who were able to verbalize the cue that was probabilistically associated with category-membership (correct-staters) in a phonological learning task, about a third respond in a way consistent with base-rate-neglect. On the other hand, non-staters are more likely to respond randomly or by probability matching the base-rates. These results suggest that explicit learning is associated with base-rate-neglect to a greater extent. We found that no learners integrate the two probabilistic patterns present in the experiment into a single Bayesian estimate. Rather, some of them focus on the base-rates, others focus on category-internal cues, and others simply fail to learn anything.

Keywords: base-rate fallacy; linguistic category-learning; phonological learning; biases across domains

Introduction

A central question in cognitive science and in linguistic theory is the nature of learning biases. Psychologists and linguists have been investigating this issue largely independently using different types of stimuli and different discipline-specific theories. In our work, we attempt to bring these two strands of research together, by investigating whether the cognitive biases identified in the psychological literature are manifested in analogous cases of linguistic learning (Moreton & Pertsova, 2024; Moreton, Pater, & Pertsova, 2017). This approach allows us to get closer to understanding which types of biases are domain-specific vs. domain-general and to what extent. The tool we use is "artificial-language learning" (ALL) experiments that allow direct control over the learning situation. Such experiments arguably engage the same type of mechanisms used in naturalistic second language acquisition, and thus present an ecologically-valid way of probing the hypotheses about human learning. In this paper we apply the ALL methodology to study a prominent bias known as Base Rate Neglect/Fallacy. The main questions we ask are whether this bias can be found in linguistic learning and if so, whether it affects the *explicit* system more than the *implicit* one (this distinction is discussed in the next section).

Base Rate Neglect (BRN) refers to one of the cognitive biases described by Daniel Kahneman and Amos Tversky in their groundbreaking work on cognitive biases (Tversky & Kahneman, 1974). This bias involves undervaluing the probabilities ("base rates") of categories or events when making categorization decisions. Specifically, this bias manifests itself in a situation when base-rates are unequal and are in conflict with the category-internal cues: some category cue is more frequent within the rare category. A typical example would be diagnosing rare diseases. Given a rare disease and a test that is 99% accurate in diagnosing both presence and absence of the disease, most people think that if one tests positive, then the probability of them having the disease is high. However, in reality this probability is low since 1% of people who don't have the disease will test positive ("false positives"), and this number will be a lot larger than true positives given that most people do not have the disease.

In this paper, we present evidence that base-rate-neglect is more likely to occur among explicit learners in a linguistic categorization task. This finding confirms that the distinction between explicit vs. implicit learning modulates the effect of BRN. Our results also offer a possible explanation of the previous inconsistent findings in the literature – we find that different participants follow different strategies, something that becomes apparent only once we perform a clustering analysis on the by-subject results. Finally, our results are consistent with the view that linguistic categorization is subject to the same set of biases as other types of categorization.

Previous work

Most of the early studies demonstrating BRN presented people with word-problems in which the base rates of the two relevant categories and the frequencies of identifying characteristics/cues for category membership were described verbally (cf. the famous "lawyer-engineer problem" Tversky and Kahneman (1974)). Bar-Hillel (1980) argued that these types of word-problems lead participants to ignore base-rates because category-specific information may be considered more prominent, salient, and more concrete compared to the abstract base-rates. A later strand of research shifted to a paradigm where base-rates and category-internal cues were instead learned experientially over a long series of trials (Manis, Dovalina, Avis, & Cardoze, 1980; Medin & Edelson, 1988; Gluck & Bower, 1988; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Goodie, 1997). The results were that depending on the nature of the task, the structure of the categories, and the number of cues used, some experiments led to

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

appropriate use of base-rates, others lead to base-rate-neglect, and yet others to alternative inappropriate uses of base-rates. Over the years a huge amount of literature has accumulated on this topic with no consensus on the nature of the phenomenon (for one comprehensive review, see Don, Worthy, and Livesey (2012)).

Several explanations of BRN have been proposed in the past. Kruschke (1996) attempts to explain BRN and its variability across experiments using the concept of *feature diagnosticity*. In particular, he proposes that "the dominant effect of base rates is to cause frequent categories to be learned before rare categories, so that the common categories are encoded in terms of their typical features, and the rare categories are encoded by whichever features distinguish them relative to the already-learned, common categories."

Other researchers propose that base-rate information is learned and applied more accurately during implicit but not explicit learning, and that some inconsistencies in the experimental findings may be due to participants using a different mix of learning strategies (Holyoak & Spellman, 1993; Koehler, 1996). For instance, Koehler (1996) speculated that, while gaining direct experience with a set of categories, base-rate sensitivity develops more robustly because information that is learned implicitly provides multiple "traces" (Hintzman, Nozawa, & Irmscher, 1982) and can be better remembered or more easily accessed compared to the information learned explicitly. The relationship between sensitivity to base-rates and implicit vs. explicit learning was directly investigated in Bohil and Wismer (2015). In this study, participants learned two simple perceptual categories (white rectangular bars) with unequal base-rates. Base-rate influence was lower in the conditions that according to the authors lessened implicit learning (i.e., observational training and delayed feedback).

Given that multiple studies suggest that implicit learning may result in better integration of base-rates into the category decision process, we will investigate whether the same is true in the domain of linguistic learning, which is typically assumed to be entirely implicit during the natural first language acquisition, but is often explicit in second language acquisition, especially at the early stages (Muñoz, 2012; VanPatten & Smith, 2022).

Implicit vs. Explicit learning in language

Multiple theories in psychology, neuroscience, and linguistics distinguish between two different systems of thought and types of memory which go by different names, such as System 1 vs. System 2, implicit vs. explicit, procedural vs. declarative (Reber, Allen, & Reber, 1999; Holyoak & Spellman, 1993; Ashby, Alfonso-Reese, Waldron, et al., 1998; Ashby FG, 2011; Kahneman, 2012). While the terms above do not always refer to the same thing, they are intended to describe two different types of thought: one that is unconscious, automatic, effortless, fast and intuitive (which we call "implicit" following the work of Reber), and one that is conscious, effortful, and involves deliberate reasoning and hypothesis testing (which we call "explicit").

In our previous work on learning of phonological categories (Moreton & Pertsova, 2024), we have shown that participants in artificial language learning experiments engage in both implicit and explicit reasoning, and that we can use post-experimental questionnaires in which participants are asked to verbalize what they learned as a proxy for identifying explicit learners. Furthermore, we find that this subjective measure of explicitness is correlated with more objective signs such as abruptness in the learning curves or the bimodal distribution at the end of the learning.

Experiment 1

The experiment described here tests the following research hypothesis: if BRN exists in learning of linguistic categories defined over phonological features, it will be more pronounced for those participants who engage in explicit learning (e.g., deliberately searching for a rule, testing hypotheses, mentally representing sounds as letters) compared to others. That is, we hypothesize, in line with the previous literature, that explicit learners, having limited short term memory resources, will not be able to integrate both the base-rates and category-internal probabilities into a single probability estimate that is close to the normative Bayesian response.

The experiment required participants to categorize nonce words into two categories that depended probabilistically on a phonological cue. We chose to use a single binary cue to avoid confounding factors like cue-competition based on salience. Thus, we focus on a direct conflict between frequencies of two binary categories: a frequent category-internal binary cue (e.g., presence of a particular property in a word) goes with a rare category, and vice versa. The categories were city names in two fictional languages spoken in different countries, one of which was 3 times as frequent as the other. However, the probability of the rare category having a positive phonological cue was a lot higher compared to the common category, leading to the possibility of BRN. The correlation of phonological cues with city names in different countries is meant to loosely model soft phonotactic tendencies that exist in natural languages: e.g., some languages tend to have "open" syllables (ending in vowels), other languages may have certain restrictions on sizes of their words, stress-location, etc. Speakers (including young infants) are known to be sensitive to these kinds of cues and can recognize what language they are hearing based on what the words sound like (Mehler et al., 1988; Grosjean, 1982; Nakai, Strange, & Akahane-Yamada, 2020).

Table 1: Base-rates and category-internal probabilities of cues.

Category (base-rate)	−F	+F
Common (0.75)	0.79	0.21
Rare (0.25)	0.36	0.64

Predictions The base-rates and category-internal probabilities we used are reported in table 1; they were chosen to maximize the chances of distinguishing among different types of participants' possible behavior. In particular, we considered that: (i) participants may respond randomly, choosing each category at chance, (ii) participants may respond by following the Bayes Rule, taking into account both base-bates and category-internal probabilities of cues, (iii) participants may match the probabilities of base-rates (such *probability matching* is found in similar experiments, including linguistic ones (Gaissmaier & Schooler, 2008; Hudson Kam & Newport, 2009)), and (iv) participants may base their responses on category-internal cues only, ignoring base-rates (this would show evidence of BRN)[1]. The distinctive probabilities for the common category response predicted by these behaviors are provided in table 2 and their explanations appear below.

Table 2: Predicted probabilities of a common category given phonological cue under different behaviors

Behavior	P(comm|+cue)	P(comm|−cue)
Random choice	0.5	0.5
Normative Bayes	0.5	0.87
Probability matching	0.75	0.75
Base-rate neglect	0.25	0.69

1. **Random choice**: participants are equally likely to choose each category, regardless of the cue

2. **Normative Bayes**: participants compute the probability of a category given a cue, P(cat|cue), following the Bayes Rule: $P(cat|cue) = \frac{P(cue|cat)*P(cat)}{P(cue)}$. The derivations for specific values in our experiment are as follows:

   (a) $P(common|+cue) = \frac{0.75*0.21}{(0.75*0.21)+(0.25*0.64)} = 0.5$

   (b) $P(common|\text{-}cue) = \frac{0.75*(1-0.21))}{(0.75*(1-0.21))+(0.25*(1-0.64))} = 0.87$

3. **Probability matching**: participants choose the common category at the rate they saw it during training (in our case 75% of the time) regardless of cue

4. **Base-rate neglect**: participants ignore base-rates which can be modeled as a Normative Bayes formula in which the P(common) = P(rare) = 0.5. (Also see footnote 1).

   (a) $P(common|+cue) = \frac{0.5*0.21}{(0.5*0.21)+(0.5*0.64)} = 0.25$

   (b) $P(common|\text{-}cue) = \frac{0.5*(1-0.21))}{(0.5*(1-0.21))+(0.5*(1-0.64))} = 0.69$

[1]There are several versions of BRN: participants may assume that the two categories are equally likely or they may undervalue true category frequencies to various degrees. Participants may also simply equate the probability of a category given a cue to the probability of the cue given the category (this is known as the *Inverse Fallacy* (Villejoubert & Mandel, 2002)). The predictions of these two versions of base-rate-neglect for the specific values we have chosen in our experiment are too close to each other, so we will not consider Inverse Fallacy as a separate prediction.

Stimuli

The stimuli were made-up names of cities in two different fictional countries. There were three conditions, based on which phonological feature was selected to serve as category cue: the *onset* condition in which the cue was presence or absence of a word-initial consonant, the *coda* condition – presence or absence of a word-final consonant, and the *number of syllables* (2 vs. 3). All of these are possible ways in which natural languages have been observed to differ. In a pilot experiment, we verified that all three of these phonological features are learnable in the same set-up when they are deterministic. For each feature, one of the feature-values was chosen as "positive" for each experiment instantiation: that is, this feature-value was associated with the rare category while the opposite feature-value was associated with the common category.

The names of cities were created as follows: for each feature a pool of possible words was created using a bigram model trained on English nouns from the CMU pronouncing dictionary[2]. These words were restricted to the following segments: consonants [p, t, tʃ, k, b, d, ʒ, g, f, s, ʃ, v z, dʒ, m, n, ŋ, l] and vowels [i, ɪ, u, ʊ, ej, ɛ, ow, ɑ]. A MaxEnt grammar of (Hayes & Wilson, 2008)[3] was trained on nouns from the Celex corpus containing only segments from the inventory above. This grammar was then used to rank potential words based on their phonotactic well-formedness (i.e, how good each word was a potential word of English). The final set of stimuli was chosen by randomly picking 100 words for each feature-value from the top-ranked potential words. Some examples of the stimuli can be seen in table 3.

Table 3: Example words

	+F	−F
Onset	gilef, nulem	uzit, abing
Coda	zimof, bifang	chebi, fala
2syll	fadiv, jofom	chikozim, lotimil

These words were then synthesized using Google's Tacotron2 model fine-tuned to be able to create novel words based on phonemic transcriptions. All words were generated to have stress on the first syllable. Roughly 10% of synthesized words were excluded due to unexpected audio-artifacts or errors.

Procedure

Task We tested people in two different training modes, which were designed to encourage either explicit or implicit learning. However, whether a participant used implicit vs.

[2]http://www.speech.cs.cmu.edu/cgi-bin/cmudict?stress=-s&in=CITE

[3]This grammar induces a set of constraints on the phonological shape of words and weighs these constraints to maximize the likelihood of the data. As a result, it allows to make predictions about the phonological likelihood of novel words.

explicit learning was ultimately determined based on their responses in the post-experimental questionnaire (discussed in the next section). The two training modes were *observational* and *feedback training*. In the observational trial, a participant heard a city name, and saw a fictional flag associated with that city's country, as shown in Figure 1a. Participants could hear the word again if they wanted to before going on to the next trial. In the feedback mode, a participant heard a city name and saw two flags (see Figure 1b). They could hear the word again, and then they had to choose the flag for the country they thought the city was from. This was followed by feedback, either "Incorrect" accompanied by the sound of a sad trumpet or "Correct" accompanied by a ding. Based on prior work (Love, 2002; Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994), we hypothesized that the feedback mode would encourage more explicit learning.

Each participant saw a unique experiment instantiation, based on 24 profiles that differed based on which of the three features served as a cue (onset, coda, syll-number), which value of the feature counted as positive, which flag was assigned to which category, and which training mode was used (feedback vs. observational). For each of these profiles, we created experiment files by randomly sampling the stimuli that satisfied them. Participants were assigned one of these files at the beginning of the experiment in a round robin fashion. The experiment was designed using the EXPERt software written by Josh Fennell; experimental files as well as all the code is available at `https://github.com/jfnl88/brn`.

The training lasted for 112 unique trials. After it was complete, participants were told that they will hear more new city-names. This testing phase (N=56) was identical to the training trials of the feedback-mode, except it did not contain any feedback.

Post-experimental questionnaire After the testing phase, participants completed a post-experimental questionnaire which collected demographic data as well as information relating to learning strategies. There were two kinds of questions in the questionnaire: two multiple choice questions that asked participants to choose *all* strategies they used during training and testing. The options were: 'just listened to the words', 'went by intuition or gut feeling', 'looked for a rule or a pattern', and 'took notes.' Participants who indicated taking notes were subsequently removed from consideration. Additionally, participants were asked to describe any patterns they noticed in the data. These free responses were then scored by two trained research assistants using a developed scoring procedure. The three main questions that the scorers had to answer were (i) Did a participant mention the relevant cue? (ii) Did a participant state any rule/pattern? (iii) If a pattern was stated, was it correct? All disagreements between the scorers were resolved by the first author. The rate of agreement between the two scorers was reasonably high on stating the relevant feature (Kohen's Kappa = 0.79) and for stating a rule (Kohen's Kappa = 0.73). We counted responses that stated the critical pattern either as a tendency (e.g., "most words ending in a consonant went with the green flag") or as an absolute as correct.

Participants Participants were recruited using the online research platform *Prolific*; they were told they would participate in a language study and that they had to be a native speaker of English with normal hearing. They were also screened for prior experience with our previous studies (which have a similar design). 190 participants took part in the study and four of them were excluded for taking notes.

Results

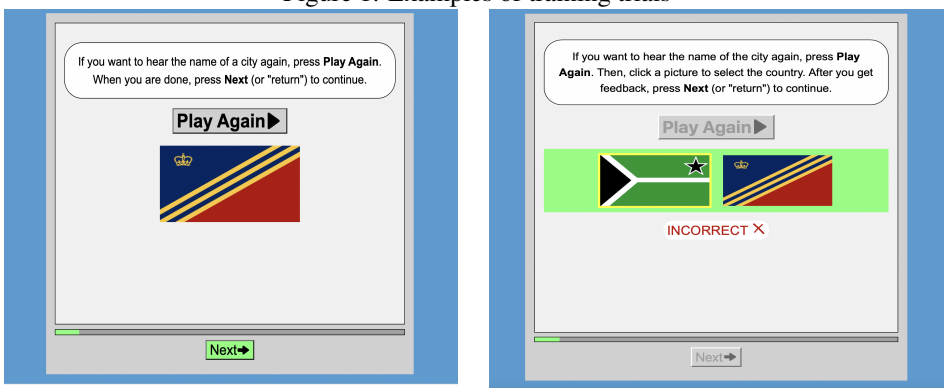
The results of the free-response in the questionnaires were analyzed to identify *correct staters* (28 participants), namely, those participants who were able to correctly verbalize the relevant category-internal cue after the experiment. Several analyses (not included here for reasons of space) on the multiple-choice questionnaire responses and on the numbers of correct staters showed that the two training modes did not significantly differ from each other on any measure of explicit/implicit learning.

To analyze if participants followed one of the predictions described in table 2, we modeled the proportion of common category responses based on the phonological cue (+F more characteristic of the rare category), training (observational vs. feedback), and the interaction between these two variables. We used a mixed effects logistic regression with random intercepts for participants and random slopes for the variable *cue*. The results of the regression model appear in table 4 and the predicted probabilities of the common-category response in the four cells of the experiment based on this regression model are given in table 5. (Note that in this probabilistic experiment there is not a right or wrong response on each trial.)

Comparing probabilities in table 5 to the expected probabilities in table 2, we see that they do not match any of the predictions. On average, regardless of cue, participants were more likely to choose the more common category, but slightly less likely when the word had the rare cue (effect of cue). They were also less likely to choose the common category in the observational training mode (effect of mode).

Next, we consider the possibility that the reason we found no easily interpretable effects is due to the fact that different groups of participants had different behavior (e.g., some neglected base-rates, others relied on base-rates only, etc.). To better understand whether our data is a result of averaging over such multiple sub-populations, we conducted a clustering analysis as follows: we plotted each subject's average proportion of common-category responses for all words that had a +cue (on the x-axis) and words that had a -cue (on the y-axis). On this plot, we overlayed the points that correspond to the predictions of the four types of behavior in Table 2. We then assigned each data-point to one of these four types/clusters based on shortest Euclidean distance. The result can be seen in Figure 2. In this plot the points that are shaded correspond to correct staters. A color of each point shows a cluster it was assigned to.

Figure 1: Examples of training trials



(a) Observational mode (b) Feedback mode

Table 4: Fixed effects in a mixed-effects logistic regression for Experiment 1 (on the probability of the common category)

Fixed effects:	Estimate	Std. Error	z value	Pr(>|z|)
Intercept (feedback, $+F$)	0.62	0.09	7.17	7.74e-13 ***
cue: $-F$ (common)	0.28	0.08	3.39	0.0007 ***
mode: observational	-0.39	0.12	-3.27	0.001 **
cue * mode	-0.08	0.11	-0.71	0.47

Table 5: Modeled probability of the common category response in Experiment 1

	feedback	observation
-F (common)	0.71	0.60
+F (rare)	0.65	0.56

We see two prominent clusters in the graph: in the top-right corner there are participants who center around the probability matching point (75% preference for the common category regardless of cue) or show *overmatching* behavior which consists of preferring the common category even more frequently. This is a rational type of response when presented with skewed base-rates (see the discussion section). The next largest cluster is in the middle, centering around random choice (with a few participants straying into the bottom left-corner, probably probability matching but reversing the two categories). The Bayesian prediction at the top does not look like a centroid of any cluster – almost all points that were closest to it lie below it. And finally, there is a small cluster to the left around the BRN center. This cluster corresponds to those participants who chose the common category more often (70% of the time) when the word had $-$cue and the rare category more often (75% of the time) when the word had $+$cue (matching the category-internal cue probabilities and ignoring base-rates).

Most notably, around one third of all correct staters, eight people or 29%, are in this cluster compared to only 7% of non-staters (this difference is statistically significant by Fisher's exact test, odds-ratio = 5.3, p = 0.002). Additionally, the correct-staters in the BRN cluster are marginally closer to the BRN-center, compared to the non-staters in the same cluster as revealed by a Welch independent one-tail test on the mean distance to the BRN-point (t = -1.68, df=16, p=0.05). These results suggest that BRN-type behavior is more likely among correct staters (who we use as a proxy for explicit learners). Some of the non-staters could potentially be explicit learners as well, who for some reason did not state what they learned. Given that non-staters are a heterogenous group, it is harder to make claims about them, but we can say that no identifiable group of learners performed in a way predicted by the Bayes' rule.

Although the results above are suggestive of an effect of explicit learning on BRN, they are due to a post-hoc clustering analysis. To replicate this finding, we conducted Experiment 2.

Experiment 2

Experiment 2 was just like Experiment 1, using the same stimuli and design but only in the observational mode. We chose the observational mode over the feedback mode because in the deterministic pilot experiment mentioned earlier, participants performed significantly better on the rare cue in the observational mode (they were equally good on the common cue).

Participants were recruited in the same way, on Prolific, among those who did not take part in previous experiments. After excluding note-takers, we had 169 valid participants with 21 of them identified as correct-staters through the questionnaire analysis as before. The results of this experiment were analyzed in the same way as the post-hoc analysis of

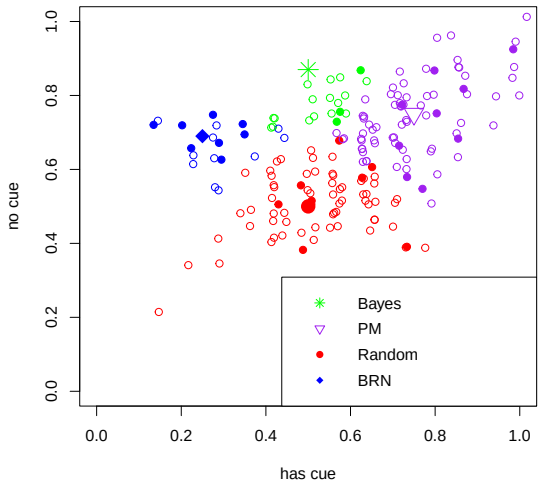


Figure 2: Experiment 1. By-subject proportion of common-category responses in the test phase for words with a cue (+cue) vs. without a cue (-cue). Shaded points correspond to correct staters.

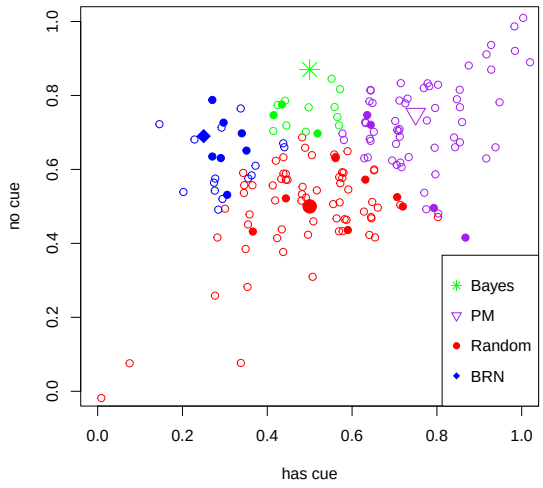


Figure 3: Experiment 2. By-subject proportion of common-category responses in the test phase for words with a cue (+cue) vs. without a cue (-cue). Shaded points correspond to correct staters.

Experiment 1.

Graph 3 shows by-subject data with correct-staters corresponding to the shaded circles. Overall, this experiment replicates our previous findings, with BRN-type behavior being more likely among correct-staters (33% vs. 10%, significant difference by Fisher's exact test, odds-ratio = 4.37, p=0.008), and with correct-staters in the BRN group being more tightly clustered compared to non-staters (Welch t-test on average distance to centroid t=-2.3, df=15.8, p=0.02).

Discussion

In both experiments, about a third of correct-staters relied primarily on the identified cue, ignoring base-rates. This is indicative of the fact that leaners who might be making categorization decisions explicitly, fail to integrate category-internal cues with base-rates. It has been suggested that implicit learners do not have similar limitations (Unsworth & Engle, 2005), however, we found that no learners in our experiment were centered around the normative Bayes' centroid. Non-staters tended to either respond randomly or probability match, possibly because the task was too difficult for them. Interestingly, many participants chose the common category even more frequently than expected based on trained frequencies. This type of response is known as *overmatching* in the literature on probability learning (see the review in Saldana, Claidière, Fagot, and Smith (2022)). It lies between the probability matching and the maximizing response (always choosing the most common category). The maximizing response is often discussed as the optimal and more rational guessing strategy compared to probability-matching because it maximizes the reward. Supporting this idea, we found that in Experiment 1 most of the overmatching participants were in the feedback condition. However, we also saw a lot of overmatching in Experiment 2 which did not include feedback.

Conclusion

Overall we found that in a categorization task that involved a probabilistic association between phonological cues and categories, a large group of participants failed to learn the pattern or only learned the base-rates of the two categories. However, a small number of participants showed evidence of base-rate-neglect and, hence, of learning the probabilistic association between cues and categories and ignoring base-rates. This type of behavior was significantly more common among correct-staters compared to non-staters. We take this result to support the hypothesis that BRN is possible in linguistic learning (just like in visual category-learning), and that it is affected by learning strategy – explicit learners are more susceptible to BRN. Additionally, we think these results can explain the previously inconsistent findings in the BRN literature. We found that different participants may be using different strategies which can be masked by averaging their results. Within the field of language learning, our results could help explain some overgeneralization mistakes committed by second-language (L2) learners, given that L2 acquisition proceeds explicitly at early stages (VanPatten & Smith, 2022). In particular, learners might focus too much on salient or diagnostic cues, ignoring category frequencies (e.g., by over-applying an irregular pattern like *bring* – **brang* due to similarity with *sing* – *sang*, underestimating the fact that irregular patterns apply to a small minority of verbs compared to the regular -/ed/ rule).

Acknowledgments

This project was partially supported by an NSF grant No. 1651105 and partially by research funds from the UNC's Institute of Arts and Humanities. The authors would like to thank Adam Aji for help with the Tacotron finetuning; Erin Humphreys and Aly Highnight for scoring questionnaires of Experiment 1. We are also grateful to comments received from the anonymous reviewers, Lisa Sanders, Joe Pater, and the members of P-side and Cognitive Tea groups at UNC-Chapel Hill.

References

Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442.

Ashby FG, M. W. (2011). *Human category learning 2.0.* Ann N Y Acad Sci. doi: 10.1111/j.1749-6632.2010.05874.x

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233.

Bohil, C. J., & Wismer, A. J. (2015). Implicit learning mediates base rate acquisition in perceptual categorization. *Psychonomic bulletin & review*, *22*, 586–593.

Don, H. J., Worthy, D. A., & Livesey, E. J. (2012). Hearing hooves, thinking zebras: A review of the inverse base-rate effect. *Psychonomic Bulletin & Review*, 1–22.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: a comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 556.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422. doi: https://doi.org/10.1016/j.cognition.2008.09.007

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227.

Goodie, A. S. (1997). *Base-rate neglect under direct experience*. University of California, San Diego.

Grosjean, F. (1982). Listening to spoken language and code-switching. In U. Ammon, N. Dittmar, & K. J. Mattheier (Eds.), *Code-switching: Anthropological and sociolinguistic perspectives* (pp. 170–186). The Hague: Mouton.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379–440.

Hintzman, D. L., Nozawa, G., & Irmscher, M. (1982). Frequency as a nonpropositional attribute of memory. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 127-141. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0022537182905011` doi: https://doi.org/10.1016/S0022-5371(82)90501-1

Holyoak, K. J., & Spellman, B. A. (1993). Thinking. *Annual review of psychology*, *44*(1), 265–315.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66. doi: 10.1016/j.cogpsych.2009.01.001

Kahneman, D. (2012). Two systems in the mind. *Bulletin of the American Academy of Arts and Sciences*, *65*(2), 55–59.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and brain sciences*, *19*(1), 1–17.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 3.

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

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. *Journal of Personality and Social Psychology*, *38*(2), 231.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*(1), 68.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178. doi: 10.1016/0010-0277(88)90035-2

Moreton, E., Pater, J., & Pertsova, K. (2017). Phonological concept learning. *Cognitive science*, *41*(1), 4–69.

Moreton, E., & Pertsova, K. (2024). Implicit and explicit processes in phonological concept learning. *Phonology*, 1–53. doi: 10.1017/S0952675724000034

Muñoz, C. (2012). Explicit learning in second language acquisition. *The Encyclopedia of Applied Linguistics*.

Nakai, S., Strange, W., & Akahane-Yamada, R. (2020). Phonotactic constraints affect the perception of non-native speech. *The Journal of the Acoustical Society of America*, *147*(3), EL235–EL241. doi: 10.1121/10.0000914

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, 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. doi: https://doi.org/10.3758/bf03200862

Reber, A. S., Allen, R., & Reber, P. J. (1999). Implicit versus explicit learning. *The nature of cognition*, 475–513.

Saldana, C., Claidière, N., Fagot, J., & Smith, K. (2022). Probability matching is not the default decision making strategy in human and non-human primates. *Scientific Reports*, *12*, 13092. doi: 10.1038/s41598-022-16983-w

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124–1131.

Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory & cognition*, *33*(2), 213–220.

VanPatten, B., & Smith, M. (2022). *Explicit and implicit*

learning in second language acquisition. Cambridge University Press.

Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from bayes's theorem and the additivity principle. *Memory & cognition*, *30*, 171–178.