NEW BULGARIAN UNIVERSITY
DEPARTMENT OF COGNITIVE SCIENCE AND PSYCHOLOGY

Critical conditions for observing the inverse base-rate effect

Author's extended abstract of the thesis submitted in partial fulfillment of the
requirements for the degree of Ph.D. in Psychology

Yolina Petrova

Supervisor: Assoc. Prof. Penka Hristova

SOFIA , BULGARIA • 2022

The thesis was submitted in English, containing:

Main text: 97 pages

120 references

28 Figures

20 Tables

8 Appendices

Table of Contents

Introduction

The eagerness to understand what determines our abilities to organization knowledge (*concepts*) and to use it (*categorization*) led to enormous empirical and theoretical boom in the last 50 years. The two-millennial old "classical view" (Smith & Medin, 1981), which describes the conceptual organization and categorization as rules-driven, was finally questioned in favor of more dynamic views – i.e., similarity- and knowledge-based approaches. This scientific progress was even recognized as one of the "success stories" in cognitive science (Gardner, 1985; Gurova, 2013).

One of the contributions of this success story is the realization that it is rather easy for people to learn which examples go with label *A* and which with label *B*. This easiness is readily adopted in a classical cognitive task, called classification task. The task requires acquisition of a given set of categories by guessing the correct category of a series of examples, presented one at a time. Typically, each response is followed by corrective feedback, which enables gradual performance improvement (Goldstone et al., 2018). This very same classification learning tradition revealed one puzzling phenomenon called the "*inverse base-rate effect*" (*IBRE)* (Medin & Edelson, 1988).

1. The inverse base-rate effect (*IBRE*) in a gist

To grasp the gist of the *IBRE*'s paradigm, imagine the following. People are instructed that they should learn two categories (or *diseases*) – *A* and *B*. More specifically, what they need to learn is which features (or *symptoms*) go with which category. Initially, the participants do not know the correct answers, so they start by guessing, where each guess is followed by feedback whether it was correct or not. In several trials, people learn to respond with category *A* when presented with "*ear aches, skin rash*"; and to go with category *B* when they see "*ear aches, back pain*" (Kruschke, 1996). Two particularly noteworthy details are left implicit for the participants. First, the two categories do not appear with the same frequency. One of them appears three times more often than the other (i.e., there is a frequent category and a rare one). Secondly, each of the categories is defined by two features (or *symptoms*) – (1) a common one, which describes both categories, in the example above this is "*ear aches*"; and (2) a unique one, which predicts only one of the categories, i.e., "*skin rash*" for category *A* and "*back pain"* for category *B*.

The learning phase is immediately followed by a testing one. The test phase starts with instructions that new examples will follow, but the task remains the same – the participants should continue classifying what they see into one of the two categories – but without feedback. Usually, the

generalization preferences of the participants are tested in several ways. When the participants are presented with a unique feature only (i.e., "*skin rash*" or "*back pain*"), their preference is expected – they go with outcome *A f*or the first case and with *B* for the second case – as the test examples do not contain any information that could suggest membership to the other category. The common feature test (i.e., "*ear aches*") is usually classified as belonging to the more frequent category in line with the base-rate information. The three features presented together (i.e., "*ear aches, skin rash*, *back pain*") yield base-rate consistent classification as well, although not as often as the previous test case. Critically, when the two unique features are shown simultaneously (i.e., "*skin rash*, *back pain*"), the classification preference goes against the base-rate information and the example is classified as belonging to the rare category (Kruschke, 1996). Surprisingly, an odd, but consistent preference pattern emerges – in the same test phase, depending on the test case, participants choose to go with the base-rate information (i.e., the *Common* test), to almost ignore it (i.e., the *All together* test) or go against it (i.e., the *Combined* test). Exactly this preference reversal makes the effect so puzzling.

2. Where does the importance of the inverse base-rate effect come from?

2.1. Practical importance of the IBRE

One of the important practical real-life manifestations of phenomena like the *IBRE* is seen in medicine. Specifically with medical professionals, failure to take into account the base-rates of the events results in serious disease likelihood overestimation (Casscells et al., 1978) and underestimation (Bergus et al., 1995). In the name of the frequency information misuse, western medicine has even turned sayings like "*When you hear hoof beats behind you, don't expect to see a zebra*" into a mantra, intending to remind the diagnosticians to always investigate the most likely clinical conditions first and only then pursue the more exotic ones.

2.2. Bridging the gap between decision-making and categorization

Beyond the medical domain, the categorization and decision-making literature shows that whether people would rely on base-rate information or not is connected to the way they have acquired the frequency information (Koehler, 1996; Barbey & Sloman, 2007). When the base-rates are provided in the form of explicit summary, people mostly ignore it. Yet, if the prevalence information is acquired implicitly on a trial-by-trial basis, the likelihood that it will be used increases (Gigerenzer et al., 1988). The idea here is that, as the number of distinct memory traces increases, the

availability of the information associated with them increases as well. The *IBRE* has a noteworthy place in the scientific literature as it is both: 1) a manifestation of a trial-by-trial category learning effect; and 2) a choice preference, going against the events' base-rates. Thus, the explanation of the *IBRE* can be informative for both – the general properties of the decision-making; and the inference processes in the context of frequency information.

2.3. The IBRE as a challenge to classical categorization models

On another note, the effect is not predicted by any of the traditional normative or learning theories. The Bayes' theorem, for example, does not offer a normative principle which can unequivocally determine which is the rationally correct response for the *Combined* cases (Medin & Edelson, 1988). On the other hand, Medin and Schaffer's (1978) exemplar-based context theory expects more frequent choices for all of the ambiguous test patterns (including the *Combined* one, eliciting rare preference with people). By contrast, most prototype-based classification models expect no specific base-rate related preferences whatsoever as the classification is purely based on some form of similarity between the test case and the prototypes of the candidate outcomes.

All in all, the *IBRE* seems an important research adventure, as it has real-life occurrences, so – in case those behaviors are indeed irrational – its understanding can lead to the prevention of eventual base-rate information misuse. In addition, as the effect is on the crossroad between decision-making and categorization, it can be informative for both domains and used as a discriminating one between alternative cognitive models.

3. Theoretical accounts of the *IBRE*

Currently, there are two rivalry accounts of the *IBRE*. One of them relies on association-based learning roots (i.e., Kruschke 1996, 2009). The other imputes the effect to high-level rule-based reasoning processes taking place during the testing phase (i.e., Juslin et al., 2001; Winman et al., 2005).

3.1. Association-based approach to the IBRE

The dominant at the moment view imputes the *IBRE* to attentional and associative highlighting of some of the category's features (Kruschke, 1996). This highlighting occurs while the categories are being learned. As one of the categories appears more often, it just happens that it is the first one that is acquired, since one perceives more examples of it. At that time, the features of the frequent category are represented as having equal attentional weight. In contrast, when it comes to the rare

category, since classification based on the common feature leads to mistakes, the attention on the common feature is shifted towards the unique one of the same category (Kruschke, 1996, 2009). All this results in non-unitary attentional distribution, where the frequent category is assumed to be represented by both of its defining features (receiving relatively equal attention weight); while the rare category becomes represented mostly by its perfect feature (receiving much stronger associative strength than the common one).

In short, the effect is rendered to a kind of *represented asymmetry* and the explanation is instantiated in a connectionist model, called *Extended ADIT Model* (*EXIT*) (Kruschke, 2001b).

3.2. Rule-based approach to the IBRE

Previous work has acknowledged a potential role for higher-order reasoning when it comes to the *IBRE* as well (Kruschke, 2003; Johansen et al., 2007; Winman et al., 2003). One such rule-based inspired explanation of the effect – called *eliminative inference* – comes from Juslin and colleagues (2001). They raise the awareness that the *IBRE* could be due to some form of high-level reasoning, and not to learning mechanisms. Essentially, Juslin et al. (2001) argue that the categorization is a matching process, where each novel example is verified in terms of sufficient matching with the acquired representations of the categories. Most often, the rule of the frequent category is probed first, since it is the better-known one. If the test example is a plausible member of the category – in the context of the *IBRE* this means to have less than one differing features with the rule – the rule is *inferred* to be the correct one (Juslin et al., 2001). If the example is novel, ambiguous and seems implausible – i.e., it has more than one differing features with the rule – the rule is *eliminated* in favor of the rare one. Exactly the elimination of the frequent rule is what causes the preference of the rare category when the novel stimulus is a *Combined* one. The assumption of this view is that all categories are represented by the whole set of defining features – i.e., there is *represented symmetry,* where both of the categories are represented by their two features (the common and the unique one). This rule-based explanation of *IBRE* is also formalized in a model, called *Elimination Model* (*ELMO*) (Juslin et al., 2001).

3.3. Empirical support for the association-based and the rule-based explanations of the IBRE

The empirical support for the outlined explanations is quite equivocal. From one side, there is no *IBRE* without a common feature (e.g., when category *A* is defined through "*earaches*" and "*skin rash*" and *B* is defined through "*back pain*" and "*nausea*") (Johansen et al., 2007; Kruschke,

2001a). This observation is consistent with the associative-based prediction, as there would be no reason for attention shifting away from any of the features that could result in a represented asymmetry (Johansen et al., 2007; Kruschke, 2001a). Meanwhile, rule-based explanation of the effect (formalized in *ELMO*) is indifferent to whether the categories share a feature or not, as it imputes the effect to better learning of the rule representing the more frequent category, and not to any structural differences between the formed rules.

Moreover, in accordance with the associative-based approach, the visual attention measured through Selection Negativity and concurrent anterior Selection Positivity event related potentials (ERPs) was found to be greater for the rare test case (*Unique to R*), compared to the frequent unique one (*Unique to F*) (Wills et al., 2014). There are recent fMRI studies reporting results in the same line of thought. It seems that specific brain regions receive significantly more activation during presentation of the unique for the rare category feature (Inkster et al., 2022). Importantly, the regions displaying stronger activation during the rare test case (*Unique to R*) compared to the frequent one (*Unique to F*) have been associated with prediction error during learning (e.g., Fouragnan et al., 2018, as cited in Inkster et al., 2022).

However, in support of the rule-based account is a relatively recent fMRI study, using multivoxel pattern analysis (O'Bryan et al., 2017). O'Bryan and colleagues (2017) demonstrated that on the *Combined* test cases participants attend more to the unique frequent feature, including when they choose the rare category. This observation coincides with the rule-based account of the *IBRE* as it implies that rare preference on the ambiguous test cases is actually associated with stronger attention to the unique frequent feature (i.e., the more frequent rule is tested first and eliminated).

It is obvious that the empirical data do not give its watertight support for any of the outlined formalized explanations – neither the associative-based account explaining the effect as a learning one, nor the rule-based one, which imputes the effect to reasoning processes during the test phase.

4. Rationale behind the Current Work

This thesis' intents are three-fold: 1) to systematically investigate the role of learning in the occurrence of the *IBRE* under different conditions; 2) to explore alternative explanations of the effect; and 3) to test the dominant associative learning explanation of the *IBRE* against conditions where the acquired categories are represented symmetrically. With these aims in mind, six experiments and one simulation were planned.

The first experiment (Experiment 1: *IBRE* with Classification Learning) aimed to establish the magnitude of the effect with the classical classification task in order to use it as a norm for comparison with the further experiments. The expectation for the first experiment was to observe the generalization preference pattern associated with the *IBRE* after training with 3:1 frequency differences with categories sharing a common feature. The stimuli employed in this experiment (simple visual stimuli constructed specifically for this project) were used throughout the rest of the experiments presented below. Next, tested was whether the *IBRE* could be obtained with an inference learning task. The central issue of the second experiment concerns whether the key to observing the *IBRE* is, indeed, represented asymmetry (acquired throughout the learning of the categories) (Kruschke, 1996, 2009). Previous research in the categorization domain has highlighted the differences in the category representations formed through classification learning and other types of learning (Chin-Parker & Ross, 2004; Sweller & Hayes, 2010; Yamauchi & Markman, 1998), and in particular through inference learning. Thus, in the second experiment (Experiment 2: *IBRE* with Inference Learning), an inference learning task was used to enforce the acquisition of symmetric representation of the rare category while maintaining the standard test procedure of the *IBRE* studies. The reasoning was that if the effect is still observed, then there would be a strong motive to revisit the statement that asymmetric representation is a necessary condition for observing the *IBRE* (i.e., Kruschke, 1996).

Two other experiments – Experiment 3: *IBRE* with Pre-Learning Motivation and Experiment 4: *IBRE* with Pre-Testing Motivation – examines the effect in conditions of pre-learning motivation manipulations. As per the motivation literature, there are a number of ways in which motivation can affect the cognitive processes – i.e., 1) accessibility of goal-related concepts, knowledge and individual items; and 2) general performance and learning. Therefore, if the magnitude of the *IBRE* in the first experiment differs from the effect's magnitude in these two experiments, given the direction of the effect, we can infer whether the effect is modulated by learning processes or it is rather unlikely.

The fifth experiment went even further and eliminated the process of gradual associative learning itself and the *IBRE* was tested in a pure decision-making task. The experiment lacked a learning phase but kept the exemplar-based scenario by introducing the relevant information within a single categorization trial (i.e., 4 examples of the target categories were presented simultaneously on the screen, together with the to-be-categorized stimulus). The rationale behind this manipulation was that if the effect is still observed, then there would be a serious motive to revisit the statement that *IBRE* is a learning effect (Kruschke, 1996).

Thirdly, the assumption for effective acquisition of the target categories was examined in a final experiment (Experiment 6: *IBRE* with Control Condition). The data from the final experiment was explored in great detail, offering some insightful results (i.e., the fact that participants failing to meet the learning criterion still demonstrated *IBRE*-like preferences). In addition, the experiment introduced a control condition testing the claim that frequency difference during exemplar-based learning is a necessary condition for observing the *IBRE* (Kruschke, 1996). The participants were required to learn two category pairs – one of the pairs followed the classical 3:1 ratio between the categories, while in the other pair both of the categories appeared the same amount of times. If *IBRE* is observed in the pair with frequency differences, but not in the control one, we would have to attribute the effect to the presence of frequency differences of the instances of the two categories sharing an overlapping feature, rather than to other confounding factors.

In addition, the association-based one and the rule-based explanations were also explored through an explicit measurement of the structure of the acquired categories. In the final phase in Experiment 1: *IBRE* with Classification Learning and Experiment 2: *IBRE* with Inference Learning, the participants were asked to describe what defined each of the categories they have acquired. The collected verbal data allows the exploration whether there is represented asymmetry (as suggested by the association-based approach) and frequent category prioritization (as suggested by the rule-based approach).

Finally, introduced is a computer-based simulation of the *IBRE.* The aim of the simulation was to explore whether the effect can be obtained with a purely associative- and probability-based architecture like the autoregressive language model of transformers type and prompt-based scenario offering no learning whatsoever (i.e., *GPT-3*, Brown et al., 2020). If the effect is observed with such a model, it would be unreasonable to assume the necessity of any additional learning-driven represented asymmetries employed for the solely purpose of explaining the emergence of the effect, as the prompt-based testing does not include representational changes.

5. Experiment 1: *IBRE* with Classification Learning

Rationale behind Experiment 1

Two aims were set behind this experiment – 1) to establish a norm for the magnitude of the *IBRE* with simple visual stimuli, 3:1 base-rate ratio and a set of instructions, which would be used for comparison with the subsequent experiments; and 2) to explore the participants' explicit knowledge

about the structure of the acquired categories and any potential prioritization associated with them. Overall, the first experiment closely followed the procedure of an experiment reported by Kruschke (1996, Experiment 1) – all participants had to learn four categories in a laboratory setting (two pairs of two-featured categories, where each pair shared a common feature and compiles a frequent and a rare category). The experiment differs from Kruschke's study (1996) in the employed stimulus material and the introduced attempt to access the participants' explicit knowledge about the acquired categories. For all of the experiments, presented in the thesis, designed were simple and well controlled visual materials, yet easy to verbalize, so that the acquired representations and the generalization preferences would be affected by prior knowledge, perceptual salience, etc. as least as possible. The experiment also employed a procedure developed to explore the conscious status of the participants' explicit knowledge concerning the structure of the acquired categories. The procedure was used as an indirect measure of the categories' prioritization and the prioritization of the features themselves.

Participants

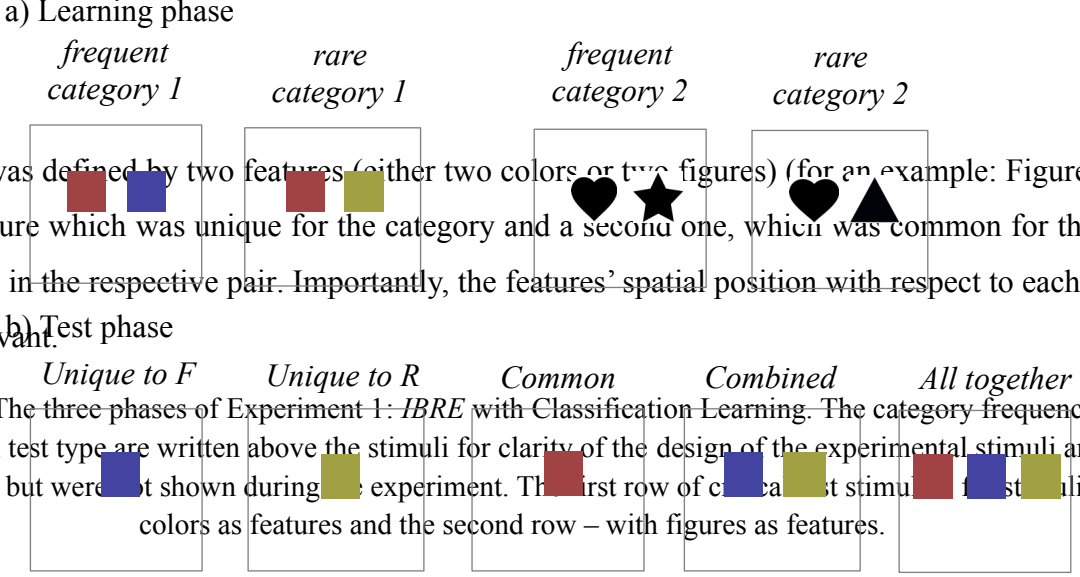
A total of 70 participants took part in the experiment in return for partial course credit. Eight of them were excluded from the analysis as in the third and final block of the learning phase they scored less than 70% correct responses either on the frequent categories or on the rare ones (or both). The final sample consisted of 62 participants (mean age = 23.9 years, SD = 8.8, 48 females). One of those participants was not considered for the verbal part of the experiment due to a technical error in the collection of the verbal report of that participant.

Materials

The stimuli features included 4 colored squares (*red, cyan, blue,* and *yellow*) and 4 black figures (*heart, circle, star,* and *triangle*), Figure 2 for reference. The colors were selected from the so-called Tetradic Colors, distributed evenly around the color wheel, which ensures that there is no clear dominance of any of the colors.

Figure 2. All features for the construction of the categories of Experiment 1 to Experiment 6. From left to right – *blue*, *red*, *cyan*, *yellow*, *circle*, *heart*, *star,* and *triangle*.

Two pairs of categories with overlapping features were designed for each participant (the pairs contained one frequent and one rare category). At random, for each participant one of the category pairs was designed from colored squares as features, and one pair with black figures. Each category



in a pair was defined by two features (either two colors or two figures) (for an example: Figure 3, a) – one feature which was unique for the category and a second one, which was common for the two categories in the respective pair. Importantly, the features' spatial position with respect to each other was irrelevant.

Figure 3. The three phases of Experiment 1: *IBRE* with Classification Learning. The category frequency and the critical test type are written above the stimuli for clarity of the design of the experimental stimuli and the procedure but were not shown during the experiment. The first row of critical test stimuli features stimuli with colors as features and the second row – with figures as features.

Procedure

The experiment consisted of three phases: a learning phase, a training phase and a verbal report phase (Figure 3). Prior to the learning phase all participants received written instructions informing that they would see different images and for each of them they should answer to which of four categories it belongs – "*V*", "*B*", "*N*", or "*M*", by pressing the corresponding QWERTY keyboard key. The grouping of the categories was not explicitly stated to the participants at any point. The participants were only notified that each of their responses would be followed by corrective feedback. Throughout the training phase, all participants saw 120 learning trials in total presented in random order . The learning trials (examples can be seen on Figure 3, panel a) were parted as follows: the frequent category was presented 45 times and its counterpart was displayed 15 times. In other words, in both pairs of categories, one of the categories appeared three times more often than the other (i.e., 3:1 ratio). After each response, the stimulus disappeared from the screen followed by written feedback for 1000 ms ("Correct!" in green or "Wrong!" in red, depending on whether the response was correct or not). Each trial began with a fixation cross presented for 500 ms and ended with 1000 ms inter-trial interval (ITI).

Prior to the test phase, the participants were informed that the task will remain the same (a classification within "*V*", "*B*", "*N*", and "*M*" category), but with new examples of the just learned categories and without corrective feedback. The instructions were followed by 20 test trials (4 per critical test type, Figure 3, panel b). The experiment ended with a request to the participants to list verbally what defined each of the four categories which they have learned in the beginning of the experiment (for the exact formulation of the instruction refer to Figure 3, panel c).

Results and Discussion

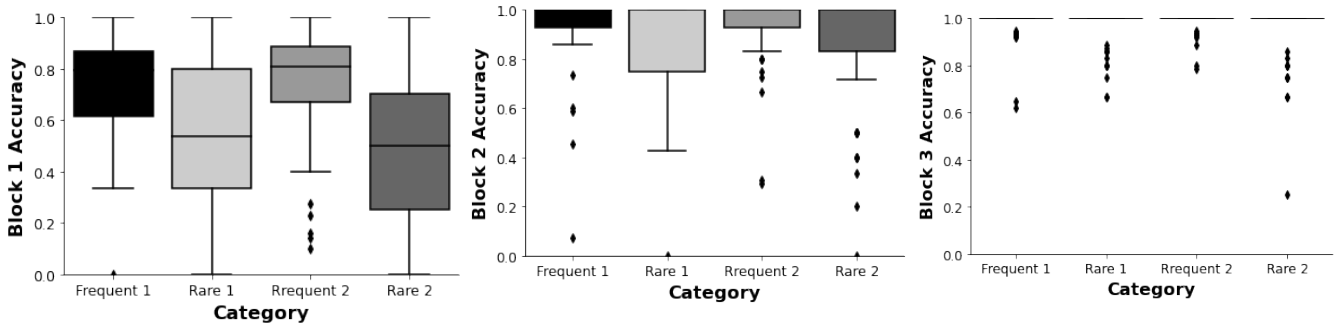


Figure 4. Mean learning accuracy and st. dev. per category in Block 1, Block 2 and Block 3 from left to the right for Experiment 1: *IBRE* with Classification Learning.

Training. Following Kruschke (1996, Experiment 1) the 120 learning trials were divided in three blocks of 40 trials each. The proportion of the correct answers in the first part of the training (first 40 trials out of 120) was higher for the frequent (0.74), compared to the rare (0.50) categories. Clearly, the frequent categories were acquired faster than the rare ones ($t(61) = 6.68$, $p < .001$, $d = 0.864$ with 95% *CI* [0.17, 0.31]), Figure 4, a), as argued by the associative-based explanation of the *IBRE* (Kruschke, 1996). The difference became smaller in the second part of the training ($t(61) = 3.65$, $p < .001$, $d = 0.463$ with 95% *CI* [0.04, 0.12]). By the end of the third and final part of the learning phase, this difference diminished (0.98 for the frequent and 0.96 for the rare category), although it remained significant – ($t(61) = 2.67$, $p = .010$, $d = 0.342$ with 95% *CI* [0.01, 0.04]), Figure 4, c). As can be seen on Figure 4, the frequent categories were learned much earlier than the rare ones. Yet, until the end of the learning trials, all categories were well learned.

Testing. As expected, people correctly choose the frequent category when presented with the frequent unique feature (in 88% of the cases); analogously for the rare unique (choosing the rare outcome in 82% of the cases), refer to Table 2. More importantly, the choice proportions clearly show that the pattern associated with the *IBRE* was successfully replicated. A chi-square goodness of fit analysis computed on the response frequencies for the *Combined* test cases and expected values of 50:50 shows small to medium rare bias – $\chi^2(1, N=237) = 7.09$, $p = .008$, $\varphi = .173$ and 95% *CI* [83, 114] and [124, 154] for the frequent and the rare outcomes respectively. For the *All together* test cases people demonstrated a slight but insignificant preference for the base-rates – $\chi^2(1, N=240) = 3.75$, $p = .053$, $\varphi = .125$ with 50:50 expected values, 95% *CI* [119, 150] and [90, 121] for the frequent and the rare outcomes. Finally, for the *Common* test cases people strongly conformed to the base-rates of the categories – $\chi^2(1, N=238) = 87.13$, $p < .0001$, $\varphi = .605$ again with 50:50 expected values and 95% *CI* [178, 203] and [35, 60] for the frequent and the rare outcomes.

Table 2: Proportion preferred categories per test type for Experiment 1: *IBRE* with Classification Learning.

Test Cases	Choice proportion			
	Frequent	Rare	Frequent O	Rare O
Unique to F	**0.88**	**0.06**	**0.03**	**0.03**
Unique to R	**0.11**	**0.82**	**0.02**	**0.05**
Common	**0.76**	**0.19**	**0.01**	**0.04**
Combined	**0.39**	**0.55**	**0.01**	**0.05**
All together	**0.54**	**0.42**	**0.02**	**0.02**

Verbal reports. The verbal reports concerning the participants' explicit knowledge of the four acquired categories (Figure 3, c) were coded in terms of the type of the reported category definition. After careful qualitative analyzes, all reported definitions were grouped into 6 types – "*complete, starting with common feature*", "*complete, starting with unique feature*", "*common only*", "*unique only*", "*incorrect*", and "*unclassified*". Overall, what is observed in the summary Table 3 is that the frequent categories are represented by both of their features without prioritizing any of them, while the rare categories are represented mainly by their unique features. This observation is in accordance with the association-based approach (Kruschke, 1996) imputing the effect to representational asymmetries.

Table 3: Verbally reported category definitions per definition type and category type in percentages for Experiment 1: *IBRE* with Classification Learning.

Type of category	Reported definition						
	complete, first common	*complete, first unique*	*common only*	*unique only*	*incorrect*	*unclassified*	**total**
frequent	31.97	35.25	7.38	14.75	5.73	4.92	100
rare	18.03	22.95	4.92	33.61	13.93	6.56	100

6. Experiment 2: *IBRE* with Inference Learning (is represented asymmetry necessary for obtaining the *IBRE*)

Rationale behind Experiment 2

In the last two decades, we have witnessed an increasing interest towards the inference–classification learning distinctions (Chin-Parker & Ross, 2004; Sweller & Hayes, 2010; Yamauchi & Markman, 1998). One prominent difference between the two tasks that has been outlined is that classification learning prioritizes the discriminating between the categories features, while the inference learners are forced to distribute their attention towards the defining the category features

more evenly (Sweller & Hayes, 2010). The inference learning requires prediction of a feature (e.g., which is the missing feature – "*skin rash*" or "*back pain"*), based on the presence of a category label (i.e., *category A*) and other features (i.e., "*earaches*"). In other words, the participant is informed that the presented stimuli is an example of *category A* and the example has the feature *X*; the task is to choose which is the missing feature – *Y* or *Z*. Due to the nature of the task, prioritization of the discriminative features is insufficient for optimal performance.

If represented asymmetry is a necessary condition for observing the *IBRE*, as argued by the association-based approach (i.e., Kruschke, 1996), then the effect should not be observed in learning conditions producing symmetric representations – like learning through inference. If the effect is still observed, then one can infer that the classical version of the *IBRE* paradigm (employing classification learning) inherits the represented asymmetry that Kruschke (1996) supposes just as a mere side effect of that learning but it is not a critical condition for obtaining *IBRE.*

Participants

A total of 70 participants took part in the experiment in return for partial course credit. Fourteen of those were excluded from the analysis because they scored less than 70% correct responses either on the frequent categories or on the rare ones (or both). Thus, the final sample consisted of 56 participants (mean age = 23.9 years, SD = 6.4, 39 females).

Materials

The visual features (Figure 2) and the structure of the categories (Figure 3, panel a) were identical to those used in Experiment 1: *IBRE* with Classification Learning.

Procedure

Unlike the previous experiment, the learning phase of this one required inference learning. The participants saw a single feature that belonged to an indicated category, positioned at the center of the screen in a black contoured square near a question mark asking for the missing feature (Figure 7).

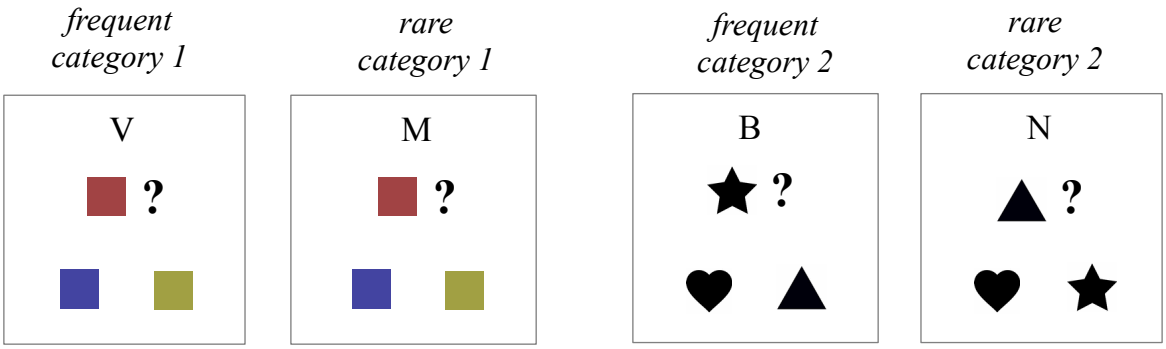


Figure 7. Example stimuli of the learning phase trials in Experiment 2: *IBRE* with Inference Learning. In those cases, the missing feature from the two trials of the color set is the unique one; the missing feature from the two trials of the figure set is the common one. For the full structure of the categories, refer to Figure 3, panel a).

Below each stimulus there were two features presented next to each other. One of the features was always the correct missing one and the other was the unique feature of the other category from the same pair. The two were always positioned randomly (relative to each other) at the bottom of the screen. Importantly, a missing feature could have been both a common feature and a unique one. This assured that the participants attended both features of each category. The participants' task was to press the corresponding button to the missing feature: '*Z*' for the left and '*X*' for the option presented on the right. In all other respects, the procedure of the experiment mimicked Experiment 1: *IBRE* with Classification Learning.

Results and Discussion

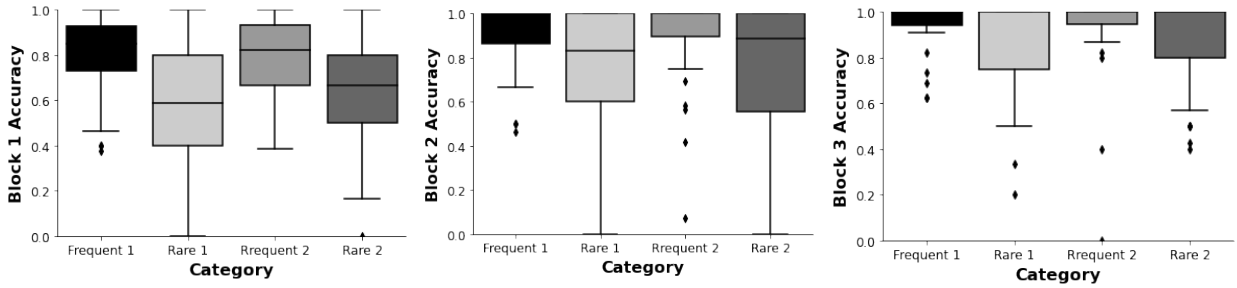


Figure 8. Mean learning accuracy and st. dev. per category in Block 1, Block 2 and Block 3 from left to the right for Experiment 2: *IBRE* with Inference Learning.

Training. As in Experiment 1: *IBRE* with Classification Learning, the frequent categories were acquired much earlier than the rare ones (Figure 8). The proportion of the correct responses for the frequent category (0.80) was significantly higher than the proportion of the correct responses for the rare category (0.62) in the first training block: $t(55) = 6.83$, $p < .0001$, $d = 0.912$ and 95% *CI* [0.13, 0.24], Figure 8, a). This difference decreased but remained significant over the third part of training

(0.95 and 0.88, for the frequent and the rare category correspondingly, $t(55) = 4.57$, $p < .001$, $d = 0.616$ with 95% *CI* [0.04, 0.11]), Figure 8, c). Yet, until the end of the learning trials, the categories were well learned.

Table 4: Proportion preferred categories per frequency and test type for Experiment 2: *IBRE* with Inference Learning.

Test Cases	Choice proportion			
	Frequent	Rare	Frequent O	Rare O
Unique to F	**0.78**	**0.12**	**0.07**	**0.03**
Unique to R	**0.17**	**0.74**	**0.05**	**0.04**
Common	**0.61**	**0.27**	**0.06**	**0.06**
Combined	**0.34**	**0.54**	**0.05**	**0.07**
All together	**0.5**	**0.42**	**0.05**	**0.03**

Testing. The obtained preference pattern was consistent with the *IBRE*. This is despite that the inference learning task enhances symmetric feature representation of the acquired categories (Sweller & Hayes, 2010; Yamauchi & Markman, 1998). Table 4 shows the choice proportion for each test type trial. A chi square analysis computed on the response frequencies for the *Combined* test cases again shows small to moderate rare preference – $\chi^2(1, N=198) = 8.91$, $p = .003$, $\varphi = .212$ with 95% *CI* [64, 92] and [106, 134] for the frequent and the rare outcomes. For the *All together* test cases, people showed numerical base-rate preference but it was not significant – $\chi^2(1, N=205) = 1.76$, $p = .185$, $\varphi = .093$ and 95% *CI* [98, 126] and [79, 108] respectively for the frequent and rare outcomes. Finally, for the *Common* test cases people demonstrate moderate to strong preference for the base-rate outcomes – $\chi^2(1, N=198) = 30.73$, $p < .0001$, $\varphi = .394$ with 95% *CI* [124, 151] and [48, 74] for the frequent and the rare outcomes respectively.

Verbal reports. The reported category definitions were grouped into the same 6 types, as in Experiment 1: *IBRE* with Classification Learning) – "*complete, starting with common feature*", "*complete, starting with unique feature*", "*common only*", "*unique only*", "*incorrect*", and "*unclassified*", allowing the exploration of the conscious status of the participants' explicit knowledge concerning the definitions of the acquired categories.

Table 5: Verbally reported category definitions per definition type and category type in percents for Experiment 2: *IBRE* with Inference Learning.

Type of category	Reported definition						
	complete, first common	*complete, first unique*	*common only*	*unique only*	*incorrect*	*unclassified*	**total**
frequent	39.29	38.39	2.68	8.93	9.82	0.89	100
rare	39.29	32.14	0.89	15.18	11.61	0.89	100

As expected and shown in Table 5, there were no significant differences between the definitions of the two types of categories (frequent and rare ones) – $\chi^2(5, N=224) = 3.6$, $p = .608$, $w = .13$. Thus, it seems that the inference learning task indeed leads to more symmetric representations (as argued by Sweller & Hayes, 2010; Yamauchi & Markman, 1998, etc.). Taken together, the results stay in contrast to the verbal reports of the classification learners (from Experiment 1: *IBRE* with Classification Learning) and to the associative-based explanation of the *IBRE* issuing the effect to asymmetric representation of the two types of categories. Rather, the inference learners seem to represent both categories in a symmetric way. Yet, they are still subjected to the *inverse base-rate effect*. This very well could mean that the classification learners in the *IBRE* paradigm indeed form asymmetric representations with an overall prioritization of the rare feature, but this is not a critical and necessary condition for obtaining the *IBRE*. It might be that the represented asymmetry is a mere side effect of the type of learning, yet not the cause for the *IBRE* itself.

The IBRE across two learning tasks – comparison between Experiment 1: IBRE with Classification Learning and Experiment 2: IBRE with Inference Learning

Manipulating the learning task within the *IBRE* procedure (by using classification learning in the first experiment and through inference learning in the second one) did not affect the magnitude of the *IBRE*, measured through the response preferences in the *Combined* test cases, $\chi^2(1, N=435) = 0.17$, $p = .679$, $w = .02$.

On another note, dividing the participants according to the way they respond to the *Combined* test trials (into *inverse, base-rate* or *mixed* generalizers) strongly indicated that the effect is not a unitary one, i.e., not all participants are subjected to it (also suggested by Winman et al., 2005). This held through for the participants in both of the experiments (refer to Figure 11).

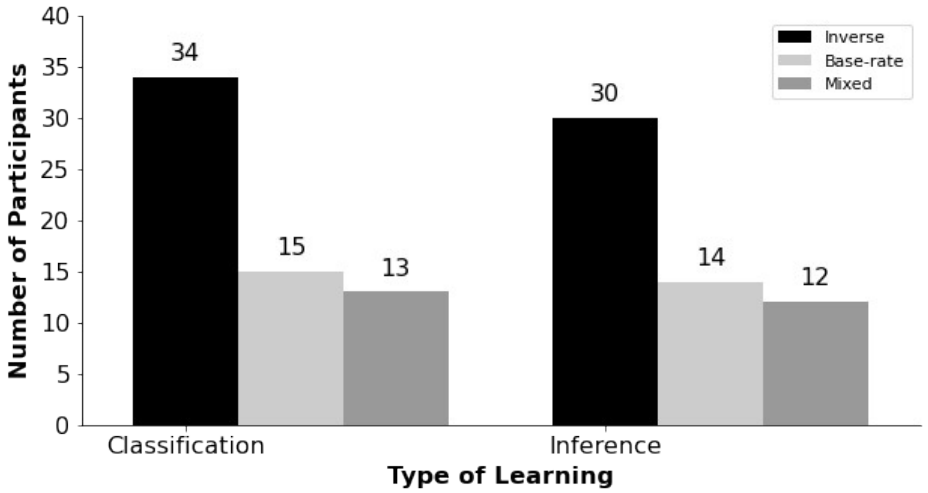


Figure 11. The figure contains the number of people demonstrating each of the generalization modes (*inverse*, *base-rate* and *mixed*) per experiment: Experiment 1: *IBRE* with Classification Learning and Experiment 2: *IBRE* with Inference Learning.

Interim discussion

All in all, on the one hand, the verbal data supported the expectation that the representations of the participants in the classification learning indeed possess representational asymmetry. On the other hand, inference learning – unlike classification learning – resulted in more balanced feature representation of the learned categories, based on the verbal reports of the participants. Yet, the *IBRE* was still obtained, besides the reported representational differences between the two types of learners. Overall, those results go against the statement that asymmetric representation is necessary to produce the *IBRE* (Johansen et al., 2007; Kruschke, 1996), as the inference learners report symmetric category definitions and still generalize in the manner of *IBRE*. It is much more probable that the asymmetric representation in the classical procedure (as assumed by Kruschke's view (1996) and observed through the verbally reported category definitions by the participants) for obtaining the *IBRE* is a mere side effect of that type of classification learning itself. More importantly, the empirical data is both consistent and inconsistent with the association-based explanation of the effect (Kruschke, 1996). On the one hand, frequent categories are, indeed, learned first and participants more or less do report the content assumed by Kruschke (1996). On the other hand, even though the two tasks differ in their attentional demands and produced representations (as also demonstrated by the differences in the reported category definitions), the *IBRE* is preserved across both. Interestingly, the results are also consistent and inconsistent with the rule-based explanation of the effect (Juslin et al., 2001). On the one side, the rule-based explanation expects an *IBRE* with both learning conditions. On the other side, this explanation assumes detailed and symmetric representations for all learners (classification and inference ones), contradicted by

the participants' verbal reports showing more representational asymmetries for the classification learners.

7. Experiment 3: *IBRE* with Pre-Learning Motivation

Rationale behind Experiment 3 and 4

Melchers et al. (2008) offer an extensive review of empirical data showing that manipulations like pretraining, task instructions, motivation among others have an effect on the encoding strategies of the participants. As the encoding strategies during learning change, the mental representations of the acquired information are affected as well and, thus, the further generalization of the participants. Thus, the general expectation is that if *IBRE* is indeed a learning effect, pre-learning additional motivation should modulate the effect. Thus, the third and the fourth experiments tested the effect in conditions of pre-learning and pre-testing motivation manipulations. In case the magnitude of the *IBRE* between those two experiments (and the first one) differs, depending on the direction of the effect, we can infer whether the effect is modulated by learning processes or by testing ones.

Participants

A total of 64 participants took part in the experiment in return for partial course credit. Eleven of them were excluded from the analysis because of not surpassing the learning threshold of at least 70% correct responses on both the frequent and the rare categories in the third block of the learning phase. Thus, the final sample consisted of 53 participants (mean age = 29.3 years, SD = 9.9, 31 females).

Materials

The stimuli materials mimicked the ones in Experiment 1: *IBRE* with Classification Learning (Figure 2).

Procedure

In addition, and in difference to the prior experiments, just **before the learning phase** the participants were also notified that they should try to perform as good as possible, because the top 3 performers in the experiment will receive a voucher for the Orange bookstore in the amount of 50 BGN. In all other respects the experiment consisted of the same learning and testing phases described in regards to Experiment 1: *IBRE* with Classification Learning.

Results and Discussion

Training. The proportion of the correct answers in the first part of the training was higher for the frequent (.72), compared to the rare (.52) categories ($t(52) = 6.22$, $p < .0001$, $d = 0.854$ with 95% *CI* [0.14, 0.27]). This difference diminished by the end of the third and final part of the learning phase (0.98 for the frequent and 0.96 for the rare category) but stayed significant– ($t(52) = 2.39$, $p = .02$, $d = 0.328$ and 95% *CI* [0.00, 0.04]). As in Experiment 1: *IBRE* with Classification Learning, the frequent categories were learned much earlier than the rare ones (Figure 12). Nevertheless, until the end of the learning trials, the categories were well learned.

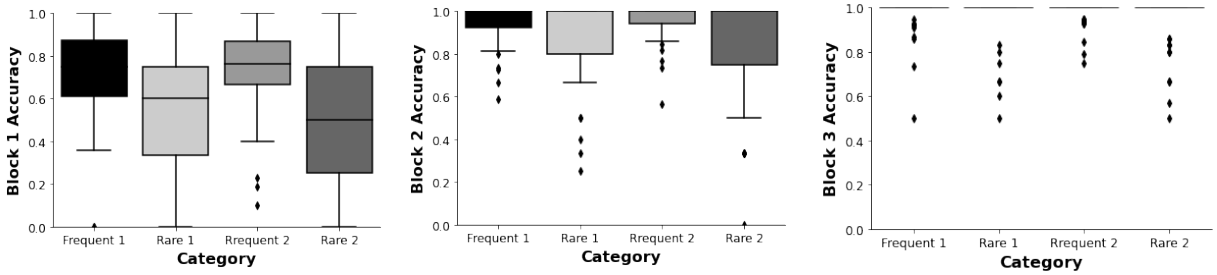


Figure 12. Mean learning accuracy and st. dev. per category in Block 1, Block 2 and Block 3 from left to the right for Experiment 3: *IBRE* with Pre-Learning Motivation.

Testing. The generalization preferences (Table 6) clearly showed that the pattern associated with the *IBRE* was successfully replicated. Chi square analysis (all with 50:50 expected values) computed on the response frequencies were as follows: the *Combined* test cases shows rare bias – $\chi^2(1, N=195) = 17.85$, $p < .0001$, $\varphi = .3$ and 95% *CI* [55, 82] and [113, 140] for the frequent and the rare outcomes respectively; for the *Common* test cases people strongly conformed to the base-rates of the categories – $\chi^2(1, N=199) = 51.26$, $p < .0001$, $\varphi = .51$ and 95% *CI* [137, 162] and [37, 62] for the frequent and the rare categories. For the *All together* test cases, people demonstrated slight but not significant base-rate preference – $\chi^2(1, N=197) = 1.14$, $p = .285$, $\varphi = .07$, 95% *CI* [92, 120] and [77, 105] for the frequent and the rare outcomes respectively.

Table 6: Proportion preferred categories per test type for Experiment 3: *IBRE* with Pre-Learning Motivation.

Test Cases	Choice proportion			
	Frequent	Rare	Frequent O	Rare O
Unique to F	**0.77**	**0.14**	**0.06**	**0.03**
Unique to R	**0.1**	**0.85**	**0.02**	**0.03**
Common	**0.71**	**0.23**	**0.03**	**0.03**
Combined	**0.32**	**0.6**	**0.03**	**0.05**
All together	**0.5**	**0.43**	**0.04**	**0.03**

8. Experiment 4: *IBRE* with Pre-Testing Motivation

Participants

A total of 64 participants took part in the experiment in return for partial course credit. Three of them were excluded from the analysis because of not surpassing the learning threshold of at least 70% correct responses on both the frequent and the rare categories in the third and final block of the learning phase. The final sample consisted of 61 participants (mean age = 26.2 years, SD = 8.4, 43 females).

Materials

The stimuli materials mimicked the description already provided for Experiment 1: *IBRE* with Classification Learning.

Procedure

As in Experiment 3: *IBRE* with Pre-Learning Motivation, the experiment consisted of two phases: a learning phase and a testing phase. Importantly, no pre-learning motivation was provided whatsoever. Contrary to Experiment 3: *IBRE* with Pre-Learning Motivation, the additional monetary incentive was provided just **before the test phase**, where the participants were notified that they should try to perform as good as possible, because the top 3 performers in the experiment will receive a voucher for the Orange bookstore in the amount of 50 BGN. In all other respects the learning and the testing phases of the experiment mimicked Experiment 1: *IBRE* with Classification Learning.

Results and Discussion

Training. The proportion of the correct answers in the first part of the training was higher for the frequent (.73), compared to the rare (.53) categories ($t(60) = 5.72$, $p < .0001$, $d = 0.732$ with 95% *CI* of [0.13, 0.27]). By the end of the third and final part of the learning phase this difference

diminished – ($t(60) = 1.73$, $p = .089$, $d = 0.222$ with 95% *CI* of [-0.003, 0.035]). As in Experiment 1: *IBRE* with Classification Learning and Experiment 2: *IBRE* with Inference Learning, the frequent categories were learned much earlier than the rare ones (Figure 15). Nevertheless, until the end of the learning trials, the categories were well learned.

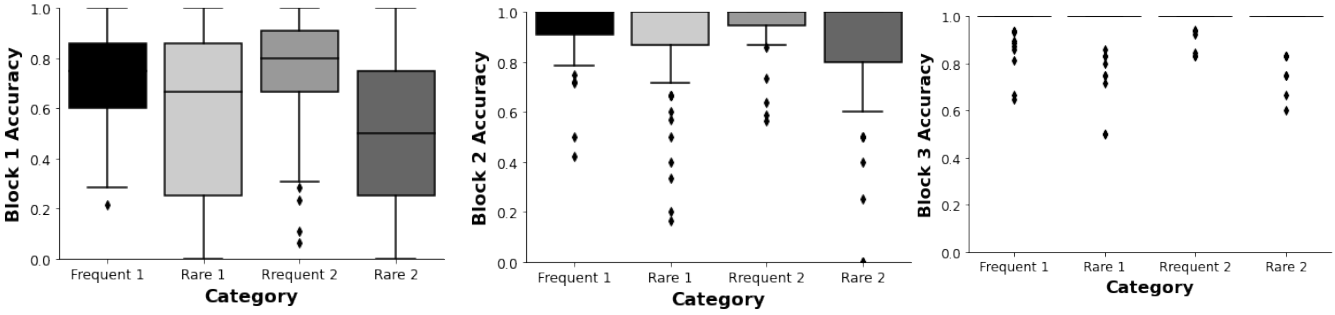


Figure 15. Mean learning accuracy and st. dev. per category in Block 1, Block 2 and Block 3 from left to the right for Experiment 4: *IBRE* with Pre-Testing Motivation.

Testing. As shown on Table 7, the pattern associated with the *IBRE* was successfully replicated. A chi square analysis computed on the response frequencies for the *Combined* test cases shows strong rare bias – $\chi^2(1, N=234) = 36.17$, $p < .0001$, $\varphi = .393$ and 95% *CI* [57, 86] and [148, 177] for the frequent and the rare outcomes respectively. For the *Common* test cases people strongly conformed to the base-rates of the categories – $\chi^2(1, N=234) = 63.61$, $p < .0001$, $\varphi = .521$ and 95% *CI* [164, 191] and [44, 70] for the frequent and the rare categories. A small but significant base-rate preference – $\chi^2(1, N=237) = 7.8$, $p = .0052$, $\varphi = .181$ and 95% *CI* [125, 155] and [82, 113] for the frequent and the rare was observed for the *All together* test cases.

Table 7: Proportion preferred categories per test type for Experiment 4: *IBRE* with Pre-Testing Motivation.

Test Cases	Choice proportion			
	Frequent	Rare	Frequent O	Rare O
Unique to F	**0.89**	**0.07**	**0.03**	**0.01**
Unique to R	**0.08**	**0.86**	**0.03**	**0.03**
Common	**0.72**	**0.23**	**0.02**	**0.03**
Combined	**0.29**	**0.69**	**0.01**	**0.01**
All together	**0.57**	**0.39**	**0.01**	**0.03**

The IBRE under different motivation conditions (no additional motivation vs. motivation before learning vs. motivation before testing)

Importantly, manipulating the motivation incentive within the *IBRE* procedure (by additionally motivating the participants with vouchers before the learning phase in Experiment 3: *IBRE* with Pre-Learning Motivation and before the test phase in Experiment 4: *IBRE* with Pre-Testing Motivation) did not affect the magnitude of the effect, as measured through the response preferences in the *Combined* test cases, $\chi^2(1, N=428) = 1.13$, $p = .287$, $w = .05$. Interestingly, there was a small but significant difference in the magnitude of the effect between the two motivation incentives and Experiment 1: *IBRE* with Classification Learning – $\chi^2(1, N=661) = 6.89$, $p = .032$, $w = .10$. It seems that the effect is stronger with additional motivation incentive. As no difference is observed between the two motivational manipulations, the motivational moderation should be taking effect mainly during the test phase (as the motivation received before the learning phase (i.e., Experiment 3) is present during the testing as well, while the motivation received before the testing phase (i.e., Experiment 4) does not include additional motivation before the learning). Therefore, the difference in the magnitude of the effect (compared to the effect in Experiment 1: *IBRE* with Classification Learning) can be viewed as kind of a support for approaches imputing the *IBRE* to some rational basis.

9. Experiment 5: *IBRE* without Learning (is learning necessary for obtaining the *IBRE*)

Rationale behind Experiment 5

A logical next question is whether the *IBRE* is a learning phenomenon at all or it can occur in a pure decision-making task as well. In an attempt to find the minimal necessary conditions for observing the *IBRE*, Johansen et al. (2007) already tried asking this question and showed that a pure decision-making task offering explicit summary of the categories and their base-rates do not result in an *IBRE*, inferring that for the *IBRE* to be observed, a base-rate neglect is also needed (Johansen et al., 2007). However, Johansen et al. (2007) presented all the instructions, the category examples (introducing the categories' frequencies) and, most importantly, all the test trials on the same page, allowing the appearance of specific alternative strategies that, probably, do not appear in the classical *IBRE* (i.e., explicit comparisons of the differences between the test examples). More importantly, their instructions to the participants included phrases like *"... you are a medical doctor in training..."* and *"... once you have read this carefully..."* (referring to the training category examples), both implying that something needs to be learned before moving on to the test cases),

which takes away from the initial idea for a task with a pure decision-making accent. Thus, put at a test next was whether the *IBRE* occurs in a pure decision-making task, positing conditions that are unlikely to allow base-rate neglect. The aim of this experiment was to test further the association-based explanation of the *IBRE* (Kruschke, 1996) by reducing any possible learning while also removing the possibility of reducing neglecting the base-rates of the categories claimed by Juslin et al., 2001 to be critical for obtaining the effect.

Participants

A total of 75 participants took part in the experiment in return for partial course credit. Twelve were below the criteria of no more than 5 mistakes on all identical control test types (and no more than 4 on a single control test type). Thus, the final sample resulted in 63 participants (mean age = 24.5 years, SD = 6.4, 52 females).

Materials

The materials consisted of the four colored squares from the first two experiments (Figure 2). However, additional variability between the trials was introduced. This was realized through presenting the stimuli in 4 (rather than 2) possible positions for each of the features. Figure 19, a) offers a visualization of a single trial.

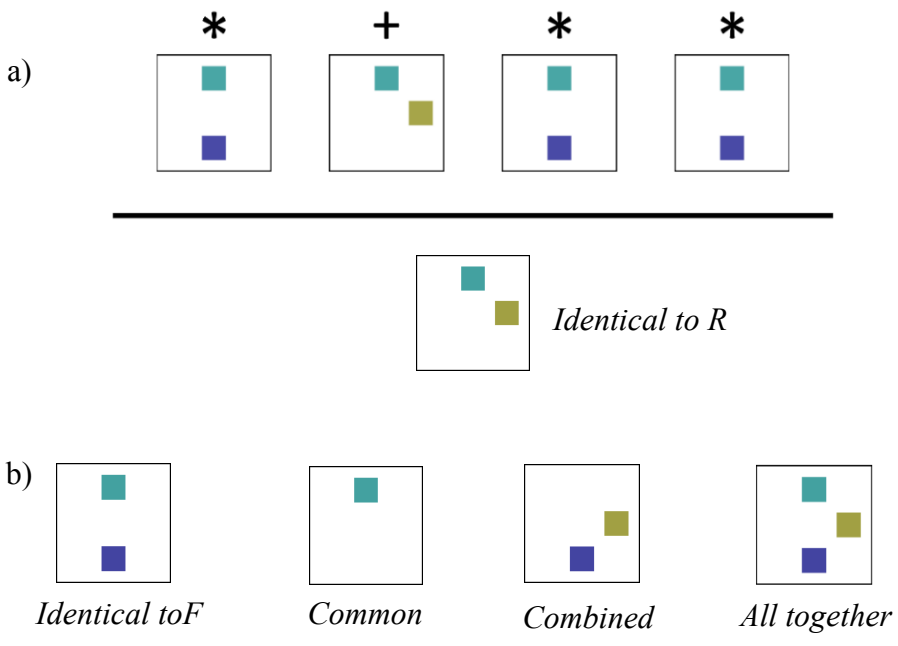


Figure 19. Panel a) shows a single (*Identical to Rare*) trial in Experiment 5: *IBRE* without Learning. (*Note*. In a real trial the label "*Identical to R*" would not appear on the screen). Panel b) visualizes all other test stimuli that could be in the place of the stimuli below the line.

Procedure

In each trial of the experiment the participants were presented with a setting like the one presented on Figure 19, a) (without the sign "*Identical to R*"). The task of the participants was to decide "*What is the stimulus below the line – '*' or '+'*". The responses were collected through key presses (B for '*' and M for '+'). As in Experiments 1: *IBRE* with Classification Learning and Experiments 2: *IBRE* with Inference Learning, all trials began with a fixation cross presented for 500 ms and ended with 1000 ms inter-trial interval (ITI). No corrective feedback was provided after the responses. Importantly, all of the trials were unrelated to each other. Each and every trial was created on-line – with the features (the colors), the positions of the colors, the position of the examples of the two categories, etc. generated at random.

Of extreme importance here is that this experimental setting allows the base-rates of the category examples (the stimuli above the horizontal line) to be presented simultaneously and each trial to act as a self sufficient test case, which is independent of the rest. In other words, the experimental setting should minimize all learning influences whatsoever (the number of identical trials across the experiment rarely exceeded two per participant). From one side this setting allowed the incorporation of all test cases usually tested in the *IBRE* paradigm. Figure 19, a) presents one example trial (in this case, target example that needs classification is identical to one of the categories). Figure 19, b) contains a more exhaustive list of the critical test types. From another side, the setting allowed the use of frequency differences between the category examples (presented above the line). Used were both 3:1 ratios (as the example in Figure 19, a) and also control cases with 2:2 ratios in which each of the two categories was represented by 2 examples. The expectation was that in the control cases each of the critical test trials would be answered in random. The experiment contained 100 trials per participants – for the 2:2 frequency test condition there were 5 trials per test type; for the 3:1 test condition there were 10 trials per *Identical to F* and *Identical to R* tests and 20 trials per critical for the *IBRE* tests (*Common*, *Combined* and *All together*). The reason behind this difference between the number of trials across the different ratios and test types is entirely practical (so that the experiment is kept to a reasonable length).

Results and Discussion

Testing. First, addressed was the question whether the 2:2 condition indeed served as a control one and there were no certain biasing preferences for some of the test types (except for the *Identical to Control 1* and *Identical to Control 2* tests). As expected, there was no preference for any of the categories: for the *Common* tests the results stood at $\chi^2(1, N=315) = 0.03$, $p = .866$, $\varphi = .05$ with

95% *CI* of [138, 174] and [141, 177] for each of the categories); there was no preference for the *Combined* test type (χ^2(1, N=315) = 0.92, p = .338, φ = .01 with 95% *CI* of [148, 184] and [131, 167] for the two categories), nor for the *All together* tests (χ^2(1, N=315) = 1.68, p = .195, φ = .05 with 95% *CI* of [128, 164] and [151, 187]).

Table 8: Proportion preferred categories per frequency condition and test type in Experiment 5: *IBRE* without Learning.

Test Cases	Choice proportion			
	Frequent	**Rare**	**Control 1**	**Control 2**
Unique to F / Control 1	**0.97**	**0.03**	**0.95**	**0.05**
Unique to R / Control 2	**0.1**	**0.9**	**0.04**	**0.96**
Common	**0.74**	**0.26**	**0.5**	**0.5**
Combined	**0.54**	**0.46**	**0.53**	**0.47**
All together	**0.66**	**0.34**	**0.46**	**0.54**

The results of higher interest were the ones that come from the preferences in the 3:1 condition. The chi-square tests for the *Common* test cases show a strong frequent bias preference – χ^2(1, N=1260) = 287.62, p < .0001, φ = .478 and 95% *CI* of [899, 996] and [299, 361] for the frequent and the rare preference. The same was observed for the *All together* test cases – χ^2(1, N=1260) = 125.72, p < .0001, φ = .316, and 95% *CI* of [651, 721] and [539, 609]. Finally, in the *Combined* test cases people demonstrated a slight but still significant preference for the base-rates – χ^2(1, N=1260) = 9.96, p = .002, φ = .089, and 95% *CI* of [651, 721] and [539, 609]. For summary results of the preference proportions refer to Table 8.

In the classical version of the *IBRE*, when it comes to the *Combined* test trials, there is an inversion of the responses; meaning that the rare responses are more common than the frequent ones. Despite the lack of complete reversal in this case, the preference pattern clearly shows the pattern associated with the *IBRE* – the rate of frequent responses was highest for the *Common* test type (up to 74%), followed by the *All together* test type (66%) and the *Combined* test type with the least frequent responses (54%). Even more, a chi-square test of association showed that there was significant association between type of critical test and category preference for the 3:1 ratio differences, χ^2(2, N=3780) = 105.28, p < .0001, w = .167. In other words, it is not only that in this setting people prefer to answer with the frequent category more often, but something made them choose the frequent category more when seeing a *Common* test type; and something made them choose the frequent category way less when they saw a *Combined* test type. Of course, there is a possibility that the reason for not witnessing the full version of the classical pattern here (which includes preference reversal when it comes to the *Combined* tests) is a procedural one, i.e., the need for more

distinct differences in terms of ratios (like 7:1). Shanks (1992) reports *IBRE* in its classical paradigm with 7:1 ratios, but not with 3:1.

10. Experiment 6: *IBRE* with Control Condition (is frequency difference necessary for observing the *IBRE*)

Rationale behind Experiment 6

A seemingly ignored detail is that the *inverse base-rate effect* is rarely tested in control conditions, i.e., when both of the categories in the category pair appear the same amount of times. To my knowledge, the effect was never tested and observed in conditions where the structure of the categories in the control pairs is the same (each category has a common and a unique feature), but there are no frequency differences between the to-be-learned target categories of the category pair. If *IBRE* is obtained for the category pair with frequency differences (but not in the pair without frequency differences), it would mean that the effect cannot be imputed to confounding stimuli characteristics, as the only difference between the two category pairs would be the within-pair frequency difference. Thus, the experiment aimed to test one of the critical conditions for observing the *inverse base-rate effect* – namely, the need of a frequent and a rare category that needs to be learned. Moreover, the data from this experiment was subjected to additional exploratory analysis. More specifically, among the main aims was to explore whether the *IBRE* appears with participants who failed to reach the learning criteria. The rational here is that, if *IBRE* indeed relies on the acquiring of asymmetric representations, then the participants who fail to satisfactorily learn the categories will not be subjected to the *IBRE*.

Participants

A total of 170 participants took part in the experiment in return for partial course credit. Fifty-five participants were removed for failing to meet the learning criterion (70% correct responses in the third and final part of the training for either of the categories). Thus, the final sample consisted of 115 participants (mean age = 26.72 years, SD = 9.42, 95 females).

Materials

The stimuli features mimicked the explanation already provided for Experiment 1: *IBRE* with Classification Learning. The difference consisted in the category structure of one of the category

pairs – while one of the category pairs consisted of a frequent and a rare category, the categories in the other pair appeared the same amount of times – 30 learning trials per category.

Procedure

The experiment mimicked the procedure of Experiment 1: *IBRE* with Classification Learning.

Results and Discussion

Training. The proportion of the correct answers in the first 40 learning trials was significantly different between the categories – ($F(3, 111) = 11.3$, $p < .001$, $\eta_p^2 = 0.069$) with an overall average of 0.63 (more specifically, 0.72 for the frequent category, 0.52 for the rare one and per 0.63 for both of the control categories). The Bonferroni pairwise post-hoc comparisons showed that the difference is in regards to the rare category only. In the first part of the learning, the participants responded with more mistakes on the rare trials compared to the frequent ones (0.77) – $t(114) = 5.79$, $p = < .001$, $d = 0.763$ and 95% *CI* [0.11, 0.30]; and to the controls – $t(114) = 3.38$, $p = .005$, $d = 0.446$ and 95% *CI* [0.03, 0.21] for one of the control categories and $t(114) = 3.32$, $p = < .006$, $d = 0.438$ and 95% *CI* [0.03, 0.21] for the other category in the control pair. The accuracy difference between the categories dropped in the second block ($F(3, 111) = 2.75$, $p < .042$, $\eta_p^2 = 0.018$). The following Bonferroni pairwise post-hoc comparisons showed that the only significance comes from a small difference between the frequent (0.93) and the rare (0.86) category – $t(114) = 2.83$, $p = .03$, $d = 0.02$ and 95% *CI* [0.01, 0.13]. Clearly, and as expected, the frequent category was acquired much earlier, followed by the control pair and the rare one acquired the latest (Figure 22). By the end of the third and final part of the learning phase, this difference diminished completely (0.98 for both the frequent and the rare category; 0.97 for both of the control categories) – $F(3, 111) = 0.89$, $p < .449$, $\eta_p^2 = 0.006$. In other words, by the end of the learning phase, all four categories were learned equally well.

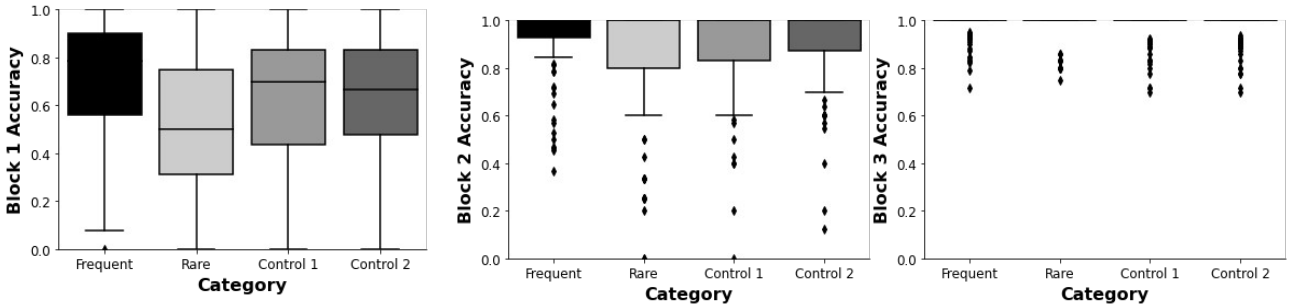


Figure 22. Mean learning accuracy and st. dev. per category in Block 1, Block 2 and Block 3 from left to the right in Experiment 6: *IBRE* with Control Condition.

Testing. As expected (Table 9), when it comes to the frequent-rare pair of categories, people correctly choose the frequent category when presented with the frequent unique feature (in 88% of the cases) and the rare outcome when presented with the rare unique feature (92% of the cases). The same preferences are observed regarding the control pair (90% correct classifications for one of the control categories and 96% for the other).

Table 9: Proportion preferred categories per test type for Experiment 6: *IBRE* with Control Condition.

Test Cases	Choice proportion			
	Frequent	Rare	Control 1	Control 2
Unique to F / Control 1	**0.88**	**0.12**	**0.9**	**0.1**
Unique to R / Control 2	**0.08**	**0.92**	**0.04**	**0.96**
Common	**0.72**	**0.28**	**0.48**	**0.52**
Combined	**0.36**	**0.64**	**0.44**	**0.56**
All together	**0.58**	**0.42**	**0.53**	**0.47**

Note. The proportions are re-calculated to include only the relevant choices (i.e., for the *Unique to R* test cases considered were only the frequent and the rare outcomes). Thus, the frequent and rare proportions sum up to one, and the two controls sum up to one as well.

More importantly, the generalization preferences regarding the critical test trials clearly showed that the pattern associated with the *IBRE* was successfully obtained in the 3:1 pair of categories. A chi square analysis computed on the response frequencies for the target *Combined* test cases and expected values of 50:50 shows significant rare-bias $\chi^2(1, N=217) = 16.04$, $p < .001$, $\varphi = 0.27$ with 95% *CI* of [65, 94] for the frequent and [123, 152] for the rare outcome respectively. For the *Common* test cases people strongly conformed to the base-rates of the categories – $\chi^2(1, N=223) = 42.19$, $p < .001$, $\varphi = 0.44$ and 95% *CI* of [146, 173] and [50, 77] for the frequent and the rare outcomes. Finally, for the *All together* test cases people demonstrated a slight base-rate preference – $\chi^2(1, N=219) = 5.59$, $p = .018$, $\varphi = 0.16$ and 95% CI of [112, 142] and [78, 107] for the frequent and the rare categories. As assumed, those biases are not observed when it comes to the control pair. None of the chi squares showed significant bias for any of the control categories – $\chi^2(1, N=218) = 0.29$, $p = .588$, $\varphi = 0.04$ (for the *Common* test trial); $\chi^2(1, N=224) = 3.5$, $p = .061$, $\varphi = 0.13$ (for the *Combined* test trial) and $\chi^2(1, N=212) = .68$, $p = .41$, $\varphi = 0.06$ (for the *All together* test case). In other words, no *inverse base-rate effect* was observed for the control categories. From this alone, we can conclude at least two things: 1) that the frequency difference between the categories in the learning phase is indeed a critical condition for observing the *IBRE*; 2) the obtained *IBRE* cannot be due to confounding stimuli characteristics, as the only difference between the two category pairs were the frequencies between the categories forming each pair.

Additional Exploratory Analysis

To check the possibility that the effect is a side effect of some type of response bias, explored was the ratio between the frequent and rare test choices (i.e., button presses) of the participants. It seems that people are quite balanced in their choices – a frequent generalization was made on 51.78% of the target trials compared to 48.22% of rare choices.

Table 11: Proportion preferred categories per test type for Experiment 6: *IBRE* with Control Condition of the participants that did not pass the learning threshold.

Test Cases	Choice proportion	
	Frequent	Rare
Unique to F	**0.74**	**0.26**
Unique to R	**0.4**	**0.6**
Common	**0.69**	**0.31**
Combined	**0.42**	**0.58**
All together	**0.55**	**0.45**

An exploratory analysis was also done on the preferences of the participants who did not surpass the learning threshold and, thus, were previously not included in the analysis. Their choices for the critical test trials (Table 11) showed a preference pattern very closely associated with the *IBRE* – that is, the rate of frequent responses was highest for the *Common* test type (up to 69%), followed by the *All together* test type (55%) and the *Combined* test type with the least frequent responses (42%). The frequent preference on the *Common* test cases showed a strong confirmation to the base-rates of the categories – $\chi^2(1, N=91) = 13.46$, $p = < .001$, $\varphi = 0.62$. Even though the proportion choices show the numeric preference pattern that we usually observe with the *IBRE*, neither the rare preference on the *Combined* tests nor the frequent preference on the *All together* test cases showed significance (respectively, $\chi^2(1, N=89) = 2.53$, $p = .112$, $\varphi = 0.17$ and $\chi^2(1, N=98) = 1.02$, $p = .117$, $\varphi = 0.10$). Yet, there are researchers that that do not report results from statistical analysis, but treat the participants' numeric preferences as good enough of a demonstration of the effect (i.e., Lamberts & Kent, 2007). The lack of significant preferences in the *All together* and the *Combined* tests is not that surprising as the sensitivity of the last two chi-squares comparisons is below 0.4 (in other words, the number of the observations is rather limited). Thus, no major conclusions can be made from that analysis alone. Despite the lack of statistical significance, the numerical preferences associated with *IBRE* are evident. On its own, this result is kind of a question to the statement that the *IBRE* is a purely learning effect. Thus, the difference between the different types of generalizers (*inverse, base-rate* and *mixed* ones), it makes sense to explore the potential differences somewhere else.

11. *IBRE* with a Transformer-based Language Model

Both models offering an explanation of the *IBRE* – *EXIT* (Kruschke, 2001) and *ELMO* (Juslin et al., 2001) – suffer from lack of generalizability. Their exact instantiations contain specific mechanisms (learning mechanisms causing asymmetric representations in the case of *EXIT* and elimination inferences in the case of *ELMO*), designed to address the *IBRE* precisely.

One exception to this trend when it comes to *IBRE* is the *RoleMap* model (Petkov & Petrova, 2019). *RoleMap* is based on the general-purpose architecture *DUAL* (Kokinov, 1988, 1994) which accounts for *IBRE* as a result of the relaxation of a constraint satisfaction network of the links expressing two general-purpose pressures – the tendency to see similar things as corresponding and the tendency to see one thing as corresponding to only one other thing.

Another unexplored endeavor is whether the *IBRE* could still be obtained with a pure association-based general-purpose architecture. The Transformer-based language models (TLMs) (Vaswani et al., 2017) are one such group of models. The specificity of there models consists in their representations, which are extremely complex and can be adapted to a large amount of tasks. This is possible due to the architecture's design – it is adapted to identify relevance of the information despite its location. In other words, it handles long-range correlations between the items in the input text, while attending some words more than others. All in all, the TLMs are models of the statistical distribution of words as extracted from a vast corpus of natural human-generated text. They are generative because they allow to be sampled, i.e., people can "ask" questions (by presenting some text fragment on the model's input), and the models can "answer" by continuing the text fragment with the most likely to follow words (Shanahan, 2022).

With the introduction of the *Generative Pre-trained Transformer 3 model* (*GPT-3)* – a prominent example of this group of models (Brown et al., 2020) – it was demonstrated that the TLMs are able to imitate humans in one particular respect – they can be few-shot learners as well (through the so-called prompt-based engineering, Zhang et al., 2021). The significant difference in the prompt-based few-shot approach is that it is in-context learning from prompting the model with just several examples (without the need of thousands of examples as its predecessors' architectures). This approach was used for the simulation, presented below.

11.1. Simulation: IBRE with GPT-3

Rationale behind the Simulation. All in all, *GPT-3* does nothing more than extracting complex statistical regularities from an enormous amount of written natural language. Importantly, the model

adapts to performing a task without any gradient updates (i.e., without any change in its representations). The prompt-based procedure is seen more as conditioning rather than learning through representation formation and change (Brown et al., 2020; Radford et al., 2019). This allows easy dissociation between the frozen state of a model from changes due to learning. Thus, due to its specificity, the model can act as a testing ground to whether the *IBRE* relies on learning prerequisites (as argued by Kruschke, 1996) or other types of mechanisms need to be considered. If *IBRE*-like "preferences" are observed in *GPT-3*, it would be a clear demonstration of at least two lines of though: 1) association-based processes are enough for the *IBRE* to appear (as the TLMs are based on association-like statistical understandings of language); 2) *IBRE* is not a pure learning-driven effect which relies on asymmetric representations (as the prompt-based approach does not change the model's representation).

Materials. To this purpose, a well established in the *IBRE* literature version of the stimuli material was used. The adopted stimuli closely followed the ones used by Kruschke (Kruschke, 1996). More specifically, employed were 6 words/phrases used as category features – i.e., *ear aches*, *skin rash*, *back pain*, *dizziness*, *sore muscles*, and *stuffy nose*. The features were presented to the model verbally in written format. Two categories with overlapping features were designed for each simulation run. One of the categories was more prevalent than the other (i.e., there was a frequent and a rare category). The two categories were labeled arbitrarily with 2 different Latin letters: "*F*" and "*R*", although for simplicity the frequent category would be referred to as category *F*, and the rare one as category *R*. Each of the two categories was defined by two features – one of the features was unique for the category and the other was shared between the two categories.

Procedure. A single simulation run consisted of the combination of 60 categorized examples with a frequency difference of 3:1 (45 examples of the frequent category and 15 examples of the rare one) with their respective labels and a single not categorized test case for which a response was required. The simulation was run 300 times in total (60 runs for each of the critical test types – *Unique to F*, *Unique to R*, *Common*, *Combined* and *All together*). For every single simulation run constructed were unique random combinations for *the two categories* (i.e., which three of the six possible features will be included in the simulation run); *the features distribution per category* (which would be the common feature, which would be the unique feature for the frequent category, and which would be the unique feature for the rare category); *the order of the examples* (so that the frequent and the rare examples are provided in a mixed order); and *the position of the features* (aiming random spatial distribution, so that the common feature is presented relatively equal number of times on the left and on the right side relative to the unique features). The trials were administered

as natural language prompts and recorded was the completion of the pattern that the model provides. Every response was classified as a frequent or a rare preference (as is done with humans' data).

Results and Discussion. As per the results in Table 12, when presented with a setting mimicking the *IBRE* paradigm, the model demonstrates *IBRE*-like preferences that we observe with humans. More specifically, when presented with a unique feature, the model responds correctly (it chooses with high certainty the frequent category when presented with the feature which is unique for the frequent category and vice versa for the unique feature of the rare category). The model prefers the more frequent option when presented with the *Common* feature and it reverses its preference when presented with a *Combined* test. The preferences demonstrated on the *All together* tests are somewhere in the middle.

Table 12: Proportion preferred categories per test type for Simulation: *IBRE* with *GPT-3* model.

Test Cases	Choice proportion	
	Frequent	Rare
Unique to F	**0.9**	**0.1**
Unique to R	**0.3**	**0.97**
Common	**0.67**	**0.33**
Combined	**0.4**	**0.6**
All together	**0.5**	**0.5**

Importantly, the results are not due to the order of the features in which they are presented during testing (i.e., whether the first presented feature is the unique frequent or the unique rare one) nor any other prior occurrence probability (i.e., the possibility that some of the features are more prevalent in the natural language and, thus, the model exhibits more attention to them). This inference can be drawn from the tokens' probability spectrum[1], which might be seen as a more detailed measure of both the direction (i.e., frequent or rare) and the strength of the generalization preference. It refers to the expectation by the model to "see" exactly the words it is presented with and the probability to continue the word sequence in any specific way (i.e., the categorization preference). For example, on the *Combined* text cases, the model demonstrated the same rare preference – both when the test case starts with the unique feature of the frequent category first and when the test case starts with the unique feature of the rare category (0.5934 in the first case and 0.6071 in the second case). In fact, the generalization probabilities (the probability for classifying

1 As the TLMs have stable representations, *GPT-3* and its probability spectrum(s) can be further used as an exploration tool regarding what could be guiding the effect. The probability spectrum of the model contains information about the top five potential completions to the prompt and their associated probabilities.

the example as an example of the frequent or the rare category) are much more influenced by the order of the examples of the categories. Even though the model has higher preference for the rare category for every of the explored example orderings, it seems that the order of the examples appears to influence the preference strength with a complex preference pattern – from one side, there seems to be a bias towards alternating answers; from another, this bias seems to weaken when there is a repetition in the examples.

It is difficult to unequivocally say that this result directly supports either the associative or the rule-based approach. The rule-based approach simply assumes that the frequent category is better learned and is more or less ignorant to the order of the examples presented in the learning phase (Juslin et al., 2001). Rather, the result contradicts the association-based approach as instantiated by Kruschke (1996), as the asymmetric representation assumption relies on the frequent category to be acquired first (i.e., as from the very beginning the participants see more examples of that category).

11.2. Interim Discussion

All in all, the *GPT-3* model (and the models similar to it) successfully replicate human eye-tracking data, reading times, and other psychological phenomena (Merkx & Frank, 2021; Schrimpf et al., 2020; Marinova et al., 2021). As *GPT-3* has been noted to perform at human level on a number of NLP tasks, it is among the most common transformer models studied by the cognitive psychologists (Binz & Schulz, 2022). As it is trained with human-made text data, it is expected that it has encoded various human-related biases – i.e., it demonstrates gender and representation biases when prompted to generate a story (Lucy & Bamman, 2021); it suggests different occupations depending on gender, race and sexual orientation (Sheng et al., 2020). For example, it is subjected to the same heuristics and biases as people when presented with the canonical "Linda problem" and "hospital problem" (Binz & Schulz, 2022). However, such models for sure do not possess high-level reasoning abilities. For example, *GPT-3* struggles on natural language inference tasks (i.e., the *ANLI* dataset, Brown et al., 2020), and it shows no sign of directed exploration, which is strongly associated with humans (Binz & Schulz, 2022).

The fact that *IBRE*-like preferences are observed with a model like *GPT-3* can be considered as a support for the idea that higher-level reasoning processes are not critical for obtaining the effect. *GPT-3* is just a statistical tool trained to do nothing else but predict the next word(s) given a sequence of words. It works in a purely associative-based manner relying on probabilistic distributions of sequences of words and representations which are simply statistical regularities. Hence, the results from the simulation question both – views attributing *IBRE* to high-level

reasoning processes and Kruschke's (1996) explanation of *IBRE* as resulting from acquired during learning represented asymmetry – and suggest that the effect might be explained with another statistical regularity that *GPT-3* could capture, such as uniform distribution of the available answers.

General Discussion and Conclusions

It has been more than 30 years now since a phenomenon, called the *inverse base-rate effect*, has been pressuring the categorization literature for an explanation. To reiterate, the *IBRE* is associated with a preference for assigning specific ambiguous examples to less prevalent outcomes. Usually, this preference appears together with a preference for more frequent categories when presented only with shared between the two categories feature. As Don et al. (2021) note, the investigation of the generality of the effect has been deeply neglected. For long, the *IBRE*-like preference has been imputed to attention- and associative-related constructs. More specifically, the effect was seen as resulting from the acquired asymmetric representations during the learning of the categories (Kruschke, 1996, 2009). The question whether the acquired represented asymmetry is not a critical condition for the effect, but simply a side-effect of the classification learning itself (i.e., it is not the reason why *IBRE* appears, but rather a parallel characteristic of the human behavior following classification learning), was never explored.

Among the goals of this thesis was to push further the debate regarding the mechanisms behind the *IBRE.* More specifically, the thesis aimed at testing the presupposed role of learning asymmetric representations for obtaining the effect (Experiments 1 to 3) and whether learning is at all required (Experiments 4 to 6). Further on, a simulation with a transformers-based probabilistic model tested in addition whether *learning* is a necessary prerequisite for obtaining the effect or it can be obtained without representational changes/learning whatsoever.

Relating the results from the experimental settings to the association- and rule-based explanations of the IBRE

In six experiments, we put the highly supported explanation – that learning-driven *asymmetric representations* stay behind the *IBRE* (Kruschke, 1996, 2009) – to a test and argued that the effect could be at least partly modulated by other processes as well. The first experiment (Experiment 1: *IBRE* with Classification Learning) reports a replication of the classical paradigm of the *IBRE* and serves as an assurance that the effect could be obtained with novel visual materials and 3:1 ratio difference. The second experiment (Experiment 2: *IBRE* with Inference Learning) introduced an important procedural difference – the generally used classification learning task (which could lead

to represented asymmetry as a side effect) was substituted with inference learning (which was expected to impede the formation of such representational asymmetries). The results from this experiment challenge the association-based hypothesis since it introduces a setting which hinders representational asymmetry and yet the *IBRE* occurs. The results from Experiment 1: *IBRE* with Classification Learning and Experiment 2: *IBRE* with Inference Learning and the potential representation differences that the two tasks lead to are supported by the verbal reports of the participants regarding the acquired category representations. The verbal reports suggest that the participants do form different category representations – the classification learners report the definitions of the categories as asymmetric, while the inference learners are more inclined to report them as defined by both of their features. The *IBRE* is, therefore, obtained despite differences in learned representations and lack of asymmetric representations in the categories sharing an overlapping feature.

In addition, the third and fourth experiments (Experiment 3: *IBRE* with Pre-Learning Motivation and Experiment 4: *IBRE* with Pre-Testing Motivation) tested the *IBRE* in additional motivation conditions offering incentive prior the learning and prior the testing phases. Although the additional incentive enhanced the effect and its magnitude was higher compared to the classical version of the paradigm (Experiment 1: *IBRE* with Classification Learning), it did not differ depending on when the additional monetary stimulus was administered, i.e., before learning (Experiment 3: *IBRE* with Pre-Learning Motivation) or before testing (Experiment 4: *IBRE* with Pre-Testing Motivation).This leads to the conclusion that *IBRE* could be if not driven at least modulated by rule-based processing appearing during testing, as the motivation received before the learning phase might still have an effect as it is present during the testing (leaving the pre-testing motivation the common factor between the two).

For the fifth experiment, we designed a pure-decision-making scenario where the learning was eliminated completely. The employed experimental setting allowed the base-rates of the category examples (the stimuli above the horizontal line) to be presented simultaneously and each trial to act as a self sufficient test case, which is independent of the rest. Yet, the pattern associated with the *IBRE* was obtained. These results question all views seeing the effect as a learning one. The results are much more in line with possible rule-based and/or exemplar-based reasoning processes behind the *IBRE*.

The sixth experiment tested the *IBRE* against a control condition presenting the target categories with equal frequency differences. As the effect was only observed for the categories with frequency

differences (but not with the categories lacking frequency differences), it is clear that we cannot impute the effect to other paradigm and or stimuli specifics. Rather, frequency difference seems as a necessary condition for observing the effect. In addition, the results from this experiment were explored in greater detail. One of the more significant findings in this regard showed that the failed learners – who are usually removed from the data analysis due to failure to surpass the pre-set learning criteria – are, in fact, subjected to the same *IBRE*-related preferences. Therefore, neither learned asymmetry nor learning, in general, is crucial for *IBRE* to appear. Rather *IBRE* relies on some kind of exemplar-based reasoning that can be involved during the testing phase (i.e., Experiment 5: *IBRE* without Learning), or exemplar-based learning (Experiments 2, 3, 4, and 6) that can be used during the test.

The reported results are extended with a simulation with a transformer-based associative architecture (more specifically, *GPT-3*). On one hand, it is notable that the *IBRE* appears with such an architecture relying on nothing more but statistical correlation between natural language. On the other hand, it is important to note that the effect was obtained without any learning and representational change, acting as an argument against all explanations of the *IBRE* imputing it to learning processes. Rather, the simulation and more specifically, the explored probability spectrum of the words and the example order manipulations in the reported simulations point to the idea that the *IBRE* could be due to an interaction between two pressures – an interaction between how long an example of a category has not appeared and a kind of willingness to keep the responses at 50:50.

Overall, the results undermine the specific instantiations of both the association-based (Kruschke, 1996) and the rule-based (Juslin et al., 2001) explanations of the *IBRE*. The dominant association-based account (Kruschke, 1996) cannot explain the reported data from the first three experiments, as it relies on asymmetric representations and learning. The results cannot be interpreted hastily as supporting the alternative rule-based explanation of the *IBRE* as well, issuing the effect to inferential reasoning processes (Juslin et al., 2001). At least this specific instantiation of rule-based explanation of the effect meets some difficulties in explaining the results from the fifth experiment (discarding the learning phase), since it relies on rule-based representations also formed through the category learning phase and assumes memory capacity limitations during the decision-making phase. In the context of the fifth experiment (when all categories are apparent and thus active), the specific rule-based instantiation predicts random choosing between the two categories (in other words, it predicts lack of effect), which was shown that is not the case.

Limitations of the study

The thesis' results undermine the dominant explanations of the *IBRE* (i.e., the association-based account (Kruschke, 1996) and the rule-based account (Juslin et al., 2001)). Nevertheless, the thesis does not offer an exhaustive systematic exploration of any specific alternative mechanisms that can potentially underlie the *inverse base-rate effect*.

Final Conclusions

The main conclusions from the reported experiments and simulations are that neither representational asymmetry acquired through learning (Experiment 2: *IBRE* with Inference Learning) nor learning itself (Experiment 5: *IBRE* without Learning) are critical for observing the *IBRE*. The *IBRE* seems task-immune: the same preference reversal was found for the critical test trials independently of the learning task history. Hence, the *IBRE* may not be easily reconciled just as a side effect of the usually applied classification learning task. It rather tells something about categorization of entities with overlapping features, which can be generalized across tasks (Classification vs. Inference learning) and situations (Learning vs. No-learning).

Given that the *IBRE* appears both after learning and in a situation with no learning conditions, at least some part of the effect should come from processes (or/and strategies) generated in the testing phase. Any explanations that issue the effect to a purely learning processes will have difficulties explaining the results from Experiment 5 (Experiment 5: *IBRE* without Learning).

Thesis Contributions

Methodological Contributions

1) The thesis investigates the inverse base-rate effect in a systematic way across six experiments manipulating different factors (i.e., learning task, motivation incentives, decision-making scenarios and control conditions) while keeping the stimuli materials and test procedure constant. This allows for a clearer view of the possible determinants of the *IBRE* and for better accounting of critical for the effect conditions as all other variables are held constant.

2) The results from the six experiments are analyzed in a consistent manner allowing comparisons across the experiments. Therefore, the observed differences cannot be explained by different data trimming, different criteria for effective learning, or different analyses.

3) Used are both qualitative and quantitative data in order to test the association-based account towards *IBRE* and more specifically the claim that the effect is due to representational asymmetry. This is important since it gives a more complete description of the given behavior.

Empirical Contributions

1) The *inverse base-rate effect* was obtained for the first time through an inference learning task. It seems that learning tasks reducing the representational asymmetries of the obtained categories still produce *IBRE*. This is important as it is a direct challenge for the association-based account of *IBRE* (imputing the effect to asymmetric representations). In addition, it is a test towards the generalizability of the effect.

2) The thesis presents for the first time an indirect measurement of the representations of the target categories as reported by the participants. The verbal reports reveal that, indeed, the classification learners acquire asymmetric representations as opposed to the inference learners. Yet, the *IBRE* appears in both types of learning tasks.

3) *IBRE*-related preferences were obtained in pure exemplar-based conditions without learning whatsoever. This is important as it questions claims arguing that the effect is a learning one (and in particular, that it is due to specific representations acquired during the learning phase).

4) It was demonstrated that *IBRE* occurs in participants who are classically rejected as bad learners. This is important, as it is a further questioning of claims relating the effect to effective learning.

5) The *IBRE* was also tested in the context of additional motivation incentives for the first time. Tested was the monetary influence as introduced either before the learning or before testing.

6) For the first time *IBRE* was tested in the context of a novel control condition – the effect was not obtained in a control condition characterized by a lack of category frequencies.

7) For the first time, it was tested whether the *inverse base-rate effect* can be obtained in a general-purpose association-based architecture. It was demonstrated that the effect occurs in such a model (more specifically, *GPT-3*) without any representation changes/learning. This is important as it stays in contradiction with the currently dominant view that in order for the effect to be obtained, learning (and more specifically, acquiring asymmetric representations of the target categories) is required.

Theoretical Contributions

The thesis presents results that are a theoretical challenge for both explanations of the *IBRE* – the associative-based one and the rule-based one.

Author's Publications

Marinova, I., **Petrova, Y**., Slavcheva, M., Osenova, P., Radev, I., & Simov, K. (2021). Monitoring Fact Preservation, Grammatical Consistency and Ethical Behavior of Abstractive Summarization Neural Models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 901-909).

Petkov, G., & **Petrova, Y.** (2019). Relation-based categorization and category learning as a result from structural alignment. The RoleMap model. *Frontiers in Psychology*, *10*, 563.

Petrova, Y., & Petkov, G. (2018). Role-Governed Categorization and Category Learning as a Result from Structural Alignment: The RoleMap Model. *International Journal of Computer and Information Engineering*, *12*(8), 578-585.