REVIEW

by Prof. Lilia Alexandrova Gurova, PhD,
New Bulgarian University, professional field 2.3. Philosophy,
for the thesis "***Critical conditions for observing the inverse base-rate effect***",
submitted by Yolina Atanassova Petrova in partial fulfillment of the requirements for the degree of Ph.D. in Psychology

The presented text is 132 pages long and contains an introduction, 11 chapters, a general discussion and conclusions, thesis contributions, a list of references (120 titles) and 8 appendices.

1. Significance of the research problem

In 2015, Benishek and colleagues announced that nearly 80% of diagnostic errors in medicine are due to cognitive biases, the most common of which are those associated with ignoring or misinterpreting the base-rate (see p. 15 of the dissertation). The inverse base-rate effect, which is investigated in the presented dissertation, is one of the most interesting of this group of biases (although it is debated whether it should be qualified as a bias). The effect could be introduced in the following way: when people are asked to decide to which of two categories (A or B) having a common feature but also one that is unique to the category, they must assign an object that has the features unique to both categories but does not possess their common feature, they systematically assign the ambiguous object to the less frequent category. This effect, described 35 years ago by Medin and Edelson (Medin & Edelson, 1988), has since been the subject of intense discussions, and the reason for this is not only the awareness that certain unwanted and potentially dangerous consequences of the effect might affect negatively the diagnostic practice. The inverse base-rate effect also poses a serious theoretical challenge, since none of the major theories of categorization predicts or explains this effect. The exemplar models, for example, predict that in situations where the inverse base-rate effect is observed in people, they should respond in exactly the opposite way—i.e. to assign the ambiguous object to the more frequent category. The prototypical models on the other hand predict an absence of preference, i.e. random assignment of the unclear case to the one category or to the other. The two most popular specific explanations of the inverse base-rate effect, based on associative

learning and rule-based inference, respectively, also do not receive unequivocal support. These challenges have motivated Yolina Petrova to seek answers to some of the controversial questions on which the contemporary discussion is focused.

2. Aims and tasks of the dissertation, results

The research presented has a three-fold aim (see Chapter 4, p. 32): (1) to investigate the role of learning in the emergence of the inverse base-rate effect (IBRE); (2) to explore the main alternative explanations for this effect; and, in particular, (3) to test the currently dominant explanation representing the effect as a product of associative learning leading to the formation of asymmetric representations of the more frequent and the less frequently occurring category. To achieve this aim, 6 experiments and 1 simulation have been carefully planned and conducted. The contribution of each experiment and the simulation to the realization of the general aim of the dissertation is described below.

Experiment 1 (Chapter 5) replicates the IBRE in the classical classification learning paradigm, using simple visual stimuli instead of the traditional verbal ones. What was new (besides the stimuli), compared to Kruschke's (1996) classical experiment, were the verbal protocols containing the definitions that the participants in the experiment had to give to the categories that were shown to them. The verbal protocols clearly show an asymmetry in the representations of the more frequent and less frequent categories: the more frequent categories are predominantly defined by their two features (the common feature to both categories and the unique feature to the given category), while the less frequent categories are defined mostly by their unique features. Yolina Petrova has interpreted this result as compatible with the association-based approach to IBRE, which assumes a crucial role of asymmetric representations in the formation of IBRE.

Experiment 2 (Chapter 6) is highly original insofar as it replicates IBRE for the first time in an inferential learning situation. This type of learning usually suppresses the formation of asymmetric representations, and thus the experiment is important for resolving the debate as to whether asymmetric representations are a necessary condition for the observation of IBRE. The experiment shows that asymmetric representations are not a necessary condition for IBRE. The verbal protocols show that the participants have not indeed formed asymmetric representations and yet they demonstrated the inverse base-rate effect.

Experiment 3 (Chapter 7) and Experiment 4 (Chapter 8) are designed to test a possible influence of motivation on the inverse base-rate effect when the incentive is introduced before the learning phase (Experiment 3) and before the testing phase (Experiment 4), respectively.

Since no significant difference in the size of the effect was observed between the two conditions containing motivation, but there was one compared to the effect obtained in experiment 1 (in the absence of motivation), Yolina Petrova concluded that the effect of motivation is not restricted to learning alone when it precedes the learning phase, therefore, the enhancement of IBRE that it leads to is most likely due to some type of rational reasoning.

Experiment 5 (Chapter 9) aims to directly test the assumption that IBRE is due to processes occurring during the learning phase. The learning in this experiment is highly limited as each participant rarely sees the same categorization task more than twice. Although a true IBRE (a reversal of categorization preferences toward the rarer category when an ambiguous object possessing both unique features of the alternative categories is present) was not observed, participants were still significantly less likely to choose the more frequent category, compared to the other two ambiguous conditions (when an object possessing only the common feature or possessing all features of both categories is present).

In Experiment 6 (Chapter 10) a control condition is introduced where both categories of the same pair appear with equal frequency. The aim is to check whether the difference in the frequency of occurrence of the categories is indeed a necessary (Yolina calls it "critical") condition for the generation of IBRE. The experiment shows that the difference in frequency of occurrence is a necessary/critical condition, as in the condition with equal frequencies of the two categories, the effect is not observed.

In the simulation of IBRE with the GPT-3 language model (Chapter 11), the model is placed in a situation similar to that in the classical IBRE classification learning paradigm, with the caveat that learning in this case is excluded because this model does not change its knowledge as a result of solving a series of categorization tasks. It turns out that, despite the absence of learning in the categorization process, the GPT-3 language model, like humans, shows a noticeable preference for categorization in the rarer category of the ambiguous object possessing the unique features of the two alternative categories. This result is particularly curious given that the decision-making processes of models such as GPT-3 are very different from those that occur in the minds of people placed in similar situations. We can only assert with high certainty (which Yolina Petrova does) that category learning, which is absent in GPT-3, is not a necessary condition for the appearance of IBRE.

3. Knowledge of state-of-the art and relevant literature

Yolina Petrova demonstrates excellent knowledge of the research on the inverse base-rate effect, as well as of the two main approaches to explaining the effect – the association-

based and the rule-based approach. She shows not only a high level of knowledge (demonstrated mainly in Chapters 2 and 3) but also a deep understanding of the literature she refers to, which is evident in the way her research is planned to address contentious issues in current discussions, but mostly in the interpretation of the obtained results.

4. Relevance of the chosen research methodology to the aims of the dissertation

Yolina Petrova successfully combines classical experimental research, qualitative research (verbal protocol analysis) and simulations with a transformer-type language model to achieve the main goal she set herself – to test the dominant association-based explanation of the inverse base-rate effect and the more general assumption that the processes leading to the inverse base-rate effect are critically related to learning.

5. Personal contribution in the collection and analysis of empirical data

The 6 experiments planned by Yolina Petrova were conducted by experimenters trained by her. The received raw data were processed and analyzed by Yolina Petrova. The computer simulation based on the language model GPT-3 was fully implemented by Yolina Petrova.

6. Evaluation of thesis contributions

A self-assessment of the contributions of the dissertation is found in the dissertation itself (pp. 105-106) and in the dissertation extended abstract. Yolina Petrova has summarized her contributions in three groups: methodological, empirical and theoretical contributions. Without undervaluing the methodological and theoretical contributions, for me the most interesting and important in view of the current discussion are the following empirical findings: (a) the inverse base-rate effect is also observed in inference learning tasks, not only in classification learning; (b) the formation of asymmetric representations is not a necessary condition for the emergence of this effect; (c) the presence of a learning phase is also not a necessary condition for the generation of the effect. These experimental findings (the latter also confirmed by a simulation involving GPT-3), in addition to casting doubt on the dominant association-based explanation of IBRE, expand our understanding of the scope of this effect and suggest that IBRE could be a product of different mechanisms that manifest themselves to a different extent in different situations, much like the categorization itself in the context of which this effect is observed.

7. Evaluation of thesis publications

In her extended abstract, Yolina Petrova indicates 3 publications (co-authored) related to the topic of the dissertation. Two of these publications are in editions indexed in SCOPUS. One of these publications (Petkov & Petrova, 2019) has 1 citation in a journal indexed in SCOPUS. In total, for the period from enrollment in the doctoral program in 2018 until now, Yolina Petrova has published 5 papers (3 of them in journals indexed in SCOPUS) and one (on experiment 1 and experiment 2 of the dissertation) is going to be submitted for publication in the journal *Memory & Cognition* (indexed in SCOPUS and Web of Science). The number and quality of Yolina Petrova's publications speak about her outstanding abilities to do excellent research and her skills for working in an interdisciplinary team.

8. Personal qualities of the doctoral candidate

I know Yolina Petrova from the time when she was a student in the master's program in cognitive science at NBU, but I have lasting impressions from her work as a full-time doctoral student at the Department of Cognitive Science and Psychology in the period 2018-2020. Although a change of the supervisor and the topic of the dissertation had to be made in the second year of her doctoral studies, Yolina was able to organize herself and collect the necessary credits for obtaining the right of early defense: for two years, instead of the 3 years provided for in the Bulgarian legislation. In addition, Yolina impresses with her motivation to acquire new knowledge and skills, which she quickly learns to a degree that allows her to teach these knowledge and skills to other students. Until now, we have not had another PhD student who, without prior mathematical or technical training, mastered certain computer modeling methods to the extent that made it possible to use them to solve research problems. Last but not least, I have excellent impressions of Yolina Petrova as a teacher. While still a PhD student, she independently developed and delivered several lectures in the "Concepts and Categorization" course in the Master's program in Cognitive Science, and after her appointment as an assistant professor in the department, she took over the course in its entirety. The feedback from students about Yolina's teaching is highly positive.

9. General opinion, recommendations and notes

My overall assessment of the presented doctoral thesis is very high. Interesting results have been obtained, which may contribute significantly to the development of discussions on the inverse base-rate effect. The text is professionally written, in a concise manner, while at the same time being clear and readable. Once again, I would like to draw attention to the fact that

in order to achieve her goals, the doctoral student skillfully combined different methods – classical behavioral experiments, exploratory analyses, analyzes of verbal protocols, computer simulations. This is rarely found in dissertations in the field of psychology defended in our country.

My questions and comments are mainly related to the interpretation of the most important results. We can accept with a high degree of confidence that these results show that the formation of asymmetric representations of the categories presented at different frequencies is not a necessary condition for the appearance of the inverse base-rate effect. However, the conclusion that this effect is not related to learning needs a more careful consideration. Yolina herself, both in the discussion of the results of Experiment 5 (in which there was no explicit learning phase) and in the general discussion (p. 100), allows for the possibility of some kind of exemplar-based learning in the testing phase. If she decides to publish the results of Experiment 5 (which I recommend), possibly together with the results of some additional research, it would be good to consider how the hypothesis of the presence of latent learning during testing might be confirmed or rejected. If, for example, it turns out that the size of the observed inverse base-rate effect depends on the duration of the testing phase (the number of cases that the participants have to categorize), then it is very likely that a kind of latent learning is involved in the formation of this effect.

I would recommend Yolina to continue her research with language model-based simulations, because this type of research does allow for precise exclusion or manipulation of factors of interest. I would also recommend her to publish the results of these simulations, drawing attention in the discussion to the fact that the decision-making processes in language models are fundamentally different from the processes we assume to be in place in humans.

10. Conclusion

The presented dissertation on the topic "***Critical conditions for observing the inverse base-rate effect***" meets all the requirements of Bulgarian legislation for awarding the educational and scientific degree "Doctor". Based on this and the above-mentioned merits of the submitted dissertation, I will vote "for" the awarding of the educational and scientific degree "Doctor" in professional field 3.2. Psychology to Yolina Atanassova Petrova.

11.08.2023 ……..…….

/Prof. Lilia Gurova/