# Signs of Non-Linearity in Base-Rate Neglect

**Christopher D. Erb (erbcd@mail.uc.edu)**
University of Cincinnati, Department of Psychology,
Dyer Hall, Cincinnati, OH 45221-0376, USA

**Heidi Kloos (heidi.kloos@uc.edu)**
University of Cincinnati, Department of Psychology,
230 Dyer Hall, Cincinnati, OH 45221-0376, USA

## Abstract

Base-rate neglect, the tendency of adults to ignore the prior probability of an event, has been well-studied over the past decades. However, the evidence for base-rate neglect and its theoretical implications are still debated. We argue that such lack of agreement comes from the mistaken assumption that performance unequivocally reflects cognitive processes. We adopt a different viewpoint, namely that performance reflects existing constraints in the person-task relation. To test whether this viewpoint is appropriate for performance in base-rate problems we manipulated the constraints available in the task's response options. With a highly constraining response mode adults are expected to exhibit the classic base-rate neglect, with little variability in their performance as procedural factors are manipulated. However, with a less constraining response mode performance is expected to be more variable and more susceptible to subtle changes in the task procedure. Results support this view, demonstrating non-linear context effects in decision making.

**Keywords:** rationality; adult reasoning; context effects.

## Introduction

Consider the following problem:

> "In a study 1000 people were tested. Among the participants there were 5 engineers and 995 lawyers. Jack is a randomly chosen participant of this study. Jack is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs. What is most likely? (a) Jack is a lawyer (b) Jack is an engineer."

Based on the description of Jack, you may be tempted to think that he is an engineer; after all, he is introverted, he enjoys reading science fiction, and he writes computer programs. Indeed, a vast majority of adults would agree with you (De Neys & Glumicic, 2008). However, the statistical information provided in the problem indicates otherwise. Given that Jack was randomly selected from a study consisting of far more lawyers than engineers (995 vs. 5), it follows that Jack is most likely a lawyer.

This type of decision-making problem has a long history in the literature on reasoning, stretching back to the classic studies of Daniel Kahneman and Amos Tversky (1973). Despite nearly four decades of research featuring these base-rate problems, discussions concerning the task are still going strong. Take, for example, the disagreement about the influence of presenting statistical information as frequencies rather than one-case probabilities. Some argue that presenting problems in terms of frequencies has more ecological validity and, therefore, improves performance on the task (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Hoffrage, Gigerenzer, Krauss & Martignon, 2002). Others have suggested that gains in performance are not attributable to the frequency format alone, but to presentation formats that encourage the formation of a set inclusion mental model (Evans, Handley, Perham, Over & Thompson, 2000).

Or consider the discussion about how to characterize the cognitive processes that take place as the task is solved. Is human reasoning subserved by two distinct processes (c.f., Evans, 1984; 2007; Kahneman & Frederick, 2002; Stanovich & West, 2000), one being heuristic while the other is analytic? If so, how do these processes function in relation to one another? Is one of these processes the default, aided by the other process only in the case of conflict? Or is the dual-process approach presenting a false dichotomy altogether?

There is even disagreement about whether base-rate neglect implies a shortcoming of the human mind or a sophisticated adaptation. Some suggest that neglect of base-rate information is an indication of humanity's underlying irrationality (Nisbett & Borgida, 1975), while others argue that the exact same performance reflects adaptive processes that maximize efficiency in decision making (Gigerenzer & Brighton, 2009).

Similar disagreements in how to interpret performance have been documented in virtually all of adult cognition, including memory, attention, decision making, and learning (for a review see Van Orden, Pennington, & Stone, 2001). It is rather the norm than the exception to disagree about which task might best reflect natural reasoning, or how to best characterize underlying cognitive processes. These disagreements are symptomatic of an assumption that performance in a task allows direct inferences about cognitive structures or cognitive processes that are at work (for full arguments, see Kloos & Van Orden, 2009; Van Orden & Kloos, 2003). Only if performance is thought to be transparent to the underlying cognitive architecture can details of the task context be argued about. But this assumption, known also as the 'effect = structure' fallacy, has been shown to be faulty (e.g., Gibbs, 1994; Lakoff, 1987). An alternative is the assumption that performance reflects a unique person-task relation, one that cannot be

reduced to the person (or the task) alone (e.g. Gibson, 1979).

The idea that performance reflects non-reducible person-task units has been formalized in the idea of constraints that reduce degrees of freedom for action (Kloos & Van Orden, in press). If a task context is highly constraining (e.g., there are only two answer options, one of which is understood to be correct), then we expect to see formulaic, uniform performance – as if a stable cognitive structure or process is operating. If, however, a task context is less constraining (e.g., the person is presented with many answer options or believes that there is no right or wrong answer), performance is likely to be affected by idiosyncratic aspects of the person's history, miniscule changes in the procedure and seemingly irrelevant aspects of task instructions or stimuli. The resulting difference in performance does not reflect different cognitive processes but rather a different coupling between the person and the task.

In the current paper we investigate whether the idea of constraints could help shed light on performance in a base-rate neglect task (c.f., Kahneman & Tversky, 1973). Adult participants had to determine the likelihood of a certain event, given base-rate information (the a priori statistical probability of a certain event) and individuating information (the stereotypical probability of the event). The crucial manipulation was in the answer options: Participants were presented either with highly constraining multiple-choice answer options (*multiple-choice condition*), or they were presented with less constraining open-ended answer options (*open-ended condition*). We also manipulated a superficially irrelevant factor, namely the order in which information was presented: In *Order 1*, base-rate information appeared first, before the individuating information; and in *Order 2*, base-rate information appeared second, after the individuating information. If constraints, rather than cognitive structure, decide the performance in a task, then our constraints manipulation should matter. In particular, one would expect performance to be more susceptible to order changes in the less constraining task (open-ended response options) than in the more constraining task (multiple-choice response options). A recall task was added at the end of the experiment that had the same response mode across conditions. This allowed us to determine the degree to which conditions differed in how information was encoded.

## Method

### Participants

Participants were 24 undergraduate students from the University of Cincinnati (10 men, 14 women) who volunteered their time in return for course credit. The mean age of participants was 19.25 years (*SD* = 3.43). One additional adult was tested and excluded from the final sample due to apparent confusion with task procedures.

### Materials and Procedure

Participants were tested individually in a quiet room. The testing session consisted of a decision-making task, an unannounced recall task, and a brief exit survey. As was done in a recent study by De Neys and Glumicic (2008), participants were asked to think aloud while solving the decision-making problems. Participants were introduced to the experiment and the thinking-aloud procedure with the following script used by De Neys and Glumicic:

> "In this experiment we try to find out how people solve everyday reasoning problems. Therefore, we ask you to "think aloud" when you're solving the problems. You should start by reading the complete problem aloud. When you're solving the problem you have to say everything that you're thinking about. All of the inferences you're making, all the comments that you're thinking of, basically everything that is going through your mind, you have to say aloud. You should be talking almost continuously up until the point that you have answered the question. Try to keep thinking aloud the whole time. If you are silent for a while I will ask you to continue to voice your thoughts."

Participants were then given the opportunity to ask questions concerning the thinking aloud procedure. Once the participants were ready to move on, the experimenter began the audio recording and presented the decision-making task. Using the same problem set developed by De Neys and Glumicic (2008), the decision-making task consisted of 18 separate decision-making problems, each containing base-rate information and individuating information. The order of problems was randomized, and the problems were organized in booklet form. The first page of each booklet featured a set of instructions which corresponded to the response mode of the featured problems. Participants in the multiple-choice condition received the following instructions, again adapted from De Neys and Glumicic (2008):

> "In a big research project a number of studies were carried out where short personality descriptions of the participants were made. In every study there were participants from two population groups (for example, carpenters and policemen). In each study one participant was drawn at random from the sample. You'll get to see the personality description of this randomly chosen participant. You will also get to see the number of people in each of the two population groups. Finally, you will be asked to indicate which population group the participant most likely belongs to (policemen, for example) by circling a response."

Only the last sentence was modified for participants in the open-ended condition. It read: "Finally, you will be asked to write the probability that the randomly chosen participant belongs to one of the population groups (policemen, for example)." Participants were asked to read the instructions aloud and were given the opportunity to ask questions regarding the task. Participants then began the decision-making task.

The base-rate information featured a brief description of a sample of 1000 people who were said to have taken part of a study. The sample consisted of two groups of people which were grossly disproportionate in number. For example, base-rate information in one problem stated: 'In a study 1000 people were tested. Among the participants there were 5 sixteen-year olds and 995 fifty-year olds. Ellen is a randomly chosen participant of this study.' Other ratios used were 996 to 4 and 997 to 3.

The individuating information provided a description of an individual who was randomly selected from the featured sample of 1000 people. For instance, given the base-rate example provided above, the individuating information was described as: 'Ellen likes to listen to hip hop and rap music. She enjoys wearing tight shirts and jeans. She's fond of dancing and has a small nose piercing.'

In a third of the trials, base-rate information was pitted against individuating information; the description of the selected person was stereotypic of an individual from the smaller group of the sample (like in the example above). These trials were *incongruent* because the stereotypic associations did not match the most probable option according to the base-rate information.

Alternatively, in another third of the trials, base-rate information matched with the individuating information. That is to say, individuating information was stereotypical of an individual from the larger group of the sample. These trials were considered *congruent*. Finally, the remaining six problems did not feature stereotypes of either population group and, therefore, were considered *neutral* problems.

In order to determine how adults combine base-rate with individuating information, each set of information was followed by a question. In the multiple-choice condition, participants had to select the most probable event out of two options. For example, given the information provide above, the test question was: 'What is most likely? (a) Ellen is sixteen (b) Ellen is fifty'. The answer options (a) and (b) were counterbalanced, such that answer option (a) matched with the base-rate information in half of the trials, while answer option (b) matched with the base-rate information in the other half of the trials.

In the open-ended condition participants were asked to write the probability of the event. For example, the question from the base-rate and individuating information above was: 'What is the probability that Ellen is sixteen?' or 'What is the probability that Ellen is fifty?' Half of the questions inquired about the smaller sub-group of the sample and the other half inquired about the larger sub-group of the sample.

In the open-ended response mode additional instructions were occasionally provided. For example, if participants were unsure of how to express their answers, the experimenter explained that probabilities are typically expressed as fractions, decimals or percentages. If participants wrote responses such as "the probability is high," the experimenter requested a more specific, numerical response. Finally, in instances where participants responded with ranges such as "50-70%," the experimenter instructed participants to provide a more precise response.

At the conclusion of the task the audio recording was stopped. The experimenter then checked the decision-making task to ensure that none of the problems were overlooked. After a short break of about a minute, participants were presented with an unannounced recall task and were instructed to answer the questions to the best of their ability. As was done in the De Neys and Glumicic (2008) study, participants solved four recall questions for each corresponding decision-making problem. The first two questions tested recall of base-rate information, and the second two tested recall of individuating information. All four questions were printed on one page. The pages were once again stapled into a booklet and followed the same order with which the decision-making problems were presented. The following is an example of the recall task:

> One of the problems you just solved concerned Ellen whose description was drawn at random from a sample of fifty-year olds and sixteen-year olds. Try to answer the following questions.
>
> Exactly how many sixteen-year olds were there in the study?
>
> _____
>
> Exactly how many fifty-year olds were there in the study?
>
> _____
>
> Circle the correct statement:
>
> a. Ellen likes to knit
> b. Ellen listens to hip hop
> c. Ellen shops at thrift stores
> d. Ellen drives a truck
>
> Circle the correct statement:
>
> a. Ellen speaks German
> b. Ellen plays the trumpet
> c. Ellen does not have a job
> d. Ellen has a small nose piercing

After completing the recall task participants were presented with an exit survey that measured the participants' perceptions of the task.

## Design

There were two different orders (Order 1: Base-rate First; Order 2: Base-rate Second) and two answer modes (Multiple-choice condition; Open-ended condition). Participants were randomly assigned to one of the four resulting experimental groups. Each participant solved six incongruent problems, six congruent problems, and six neutral problems. Recall was identical across groups.

# Results

Our first analysis pertains to participants' performance in the multiple-choice condition. It was scored according to whether the normatively correct option was selected (i.e., the answer option that corresponded to the largest

population group). Responses were collapsed across trials within problem type (incongruent, congruent, neutral), yielding three proportion-correct scores for each participant. Figure 1A displays the means of these scores for the multiple-choice condition as a function of problem type and order. A mixed-design 2 x 3 ANOVA (with problem type as the within-subject factor and order as the between-subject factor) revealed a significant effect of problem type, $F(2, 20) = 101.42$, $p < .001$, but no significant effects of order or order interaction, $p$s $> .4$. As expected, below-chance performance was obtained for the incongruent problems ($M = 0.19$, $SE = 0.05$), while performance was at ceiling (or above chance) for congruent problems ($M = 1.00$) and neutral problems ($M = 0.83$, $SE = 0.11$). The order in which information was presented had no effect on performance in this condition.

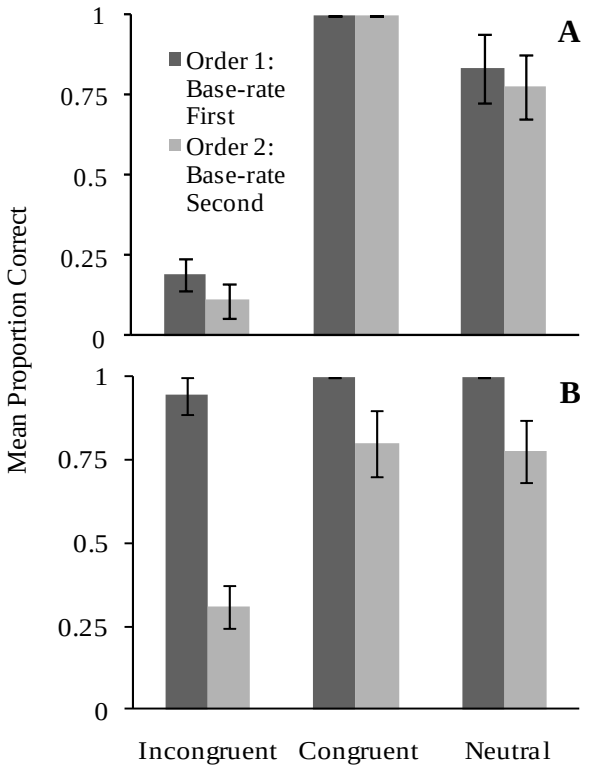


Figure 1: Mean proportion of correctly answered problems as a function of problem type and order. A: multiple-choice condition. B: open-ended condition. Error bars display standard errors.

A very different picture emerges when adults were given a continuum of response options (opened-ended condition). Responses to the prompt for each trial were first scored to match the multiple-choice scoring system. Probabilities below 50% were scored as correct for questions that pertained to providing the probability that the individual is a member of the smaller population group. Alternatively, probabilities above 50% were scored as correct for questions that pertained to providing the probability that the individual is a member of the larger population group. Responses of 50% were not included in the following analysis. This resulted in the exclusion of three of 108 responses in Order 1 and six of 108 in Order 2.

Figure 1B displays the mean proportion of correct responses in the open-ended condition as a function of problem type and order. A mixed-design 2 x 3 ANOVA (with problem type as the within-subject factor and order as the between-subject factor) revealed not only the expected significant effect of problem type, $F(2, 20) = 10.04$, $p < .001$, but also a significant effect of order, $F(1, 10) = 52.90$, $p < .001$, and a significant interaction, $F(2, 20) = 6.30$, $p < .01$. Problem type affected performance only in Order 2, $F(2, 10) = 9.05$, $p < .01$, with below-chance performance on incongruent problems ($M = 0.31$, $SE = 0.06$), and above-chance performance on congruent problems ($M = 0.80$, $SE = 0.10$) and neutral problems ($M = 0.78$, $SE = 0.09$). In Order 1, however, problem type did not affect performance, $F < 1.0$, $p < .4$, with participants performing at or near ceiling on all problem types ($M = .98$, $SE = .02$).

One critique of the above analysis is that the assumption of homogeneity of variance was not met across problem types. A Levene's test of equality of error variances revealed significant difference in variance for the congruent and neutral problem types ($p$s $< .01$), undermining the results of the parametric tests for these problem types. For this reason, we focus only on the incongruent problem type in the next analysis. Recall that this problem type is the more relevant problem type in the base-rate literature because it demonstrates the base-rate neglect. A 2 x 2 between-subjects ANOVA, with response mode (multiple-choice; open-ended) and order (Order 1, Order 2) as between-subject factors, replicates the results of our previous analyses. It revealed a significant effect of response mode, $F(1, 20) = 69.45$, $p < .001$, a significant effect of order, $F(1, 20) = 39.53$, $p < .001$, and a significant interaction effect, $F(1, 20) = 23.28$, $p < .001$.

To account for performance in the open-ended condition on a continuum, and thus to get a more accurate sense of the data, we computed the distance of responses from the normatively correct probability. For example, if a participant responded with "30%" when the normatively correct response 0.5% or lower, the resulting score would be 29.5%. These scores were once again collapsed across trials within a problem type, yielding three mean distance scores for each participant. Figure 2 shows the mean scores as a function of problem type and order.

A mixed-design 2 x 3 ANOVA (with problem type as the within-subject factor and order as the between-subject factor) revealed a significant effect of problem type, $F(2, 20) = 11.17$, $p < .001$, a significant effect of order, $F(1, 10) = 33.41$, $p < .001$, and a marginally significant interaction, $F(2, 20) = 3.04$, $p < .07$[1]. Once again, a significant effect of

[1] The interaction might not have reached significance due to unequal variances between the two orders, found for each of the problem types (Levene's Test: $F$s$(1,10) > 11.8$, $p$s $< .01$).

problem type was found for Order 2, $F(2, 10) = 8.81$, $p < .01$, but not for Order 1 ($p > .13$).

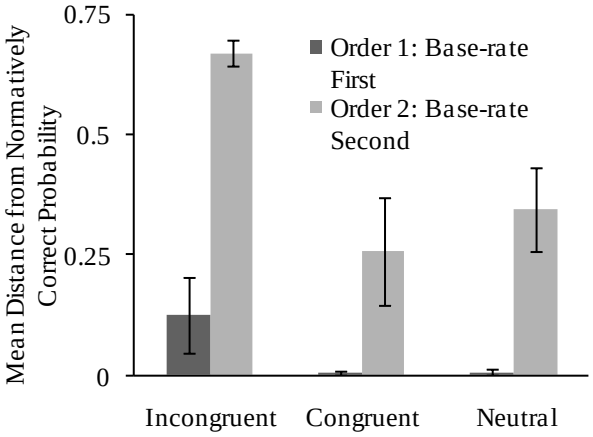


Figure 2: Mean distance from normatively correct probability as a function of problem type and order in the open-ended condition. Error bars display standard errors.

Finally, performance in the open-ended condition was scored in a third way, this time according to whether the response violated the rules of normative probability. For example, if the base-rate information listed a ratio of 3 to 997, probability judgments above 0.3% were scored as incorrect (assuming the question pertained to providing the probability that the individual is a member of the smaller population group). A mixed-design 2 x 3 ANOVA (with problem type as the within-subject factor and order as the between-subject factor) revealed a significant effect of order $F(1, 10) = 26.35$, $p < .001$, no significant effect of problem type, $p > .4$, and no significant interaction, $p > .4$. As Figure 3 illustrates, average performance across problem types was higher for Order 1 ($M = .81$, $SE = .12$) than for Order 2 ($M = .06$, $SE = .12$).

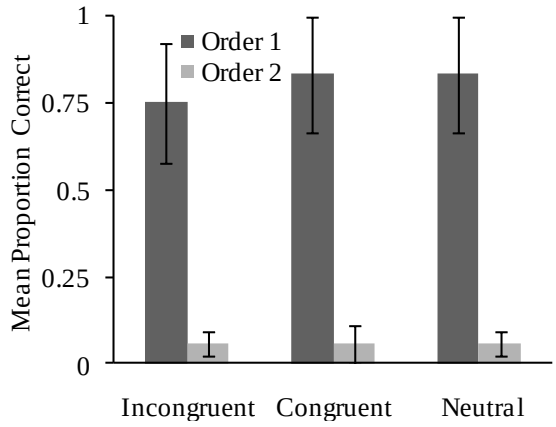


Figure 3: Mean proportion of normatively correct answers as a function of problem type and order (Order 1: Base-rate First, Order 2: Base-rate Second) in the open-ended condition. Error bars display standard error.

Thus far we have shown that the pattern of responses on the decision-making task varied with response mode (multiple-choice vs. open-ended). In the multiple-choice condition order had no effect on performance. But in the open-ended condition, no matter how data was scored, order had a highly significant impact.

One could argue that the difference between conditions is spurious, due to perhaps extraneous factors pertaining to small sample size. Our analysis of participants' base-rate recall provides reason to doubt these possible objections. Bear in mind that recall took place at the end of the experimental session, and the task employed the same response mode for all participants. Thus, if the effect of response mode in base-rate problems was spurious due to small sample size, then we would expect to see differences among conditions in the recall task as well.

Performance on recall of the base-rate information was scored according to whether participants correctly identified the relative size of each group (i.e., which group was larger and which group was smaller). A 2 x 2 x 3 mixed-design ANOVA was conducted, with condition and order as the between-subject factors and problem type as the within-subject factor. Importantly, there was no significant difference and no significant interaction ($F$s < 2.47, $p$s > .13), with above-chance performance for each group (assuming a chance probability of 0.5), single-sample $t$s > 2.2, $p$s < 0.05. Figure 4 provides the individual means for response mode and problem type, collapsed across order.

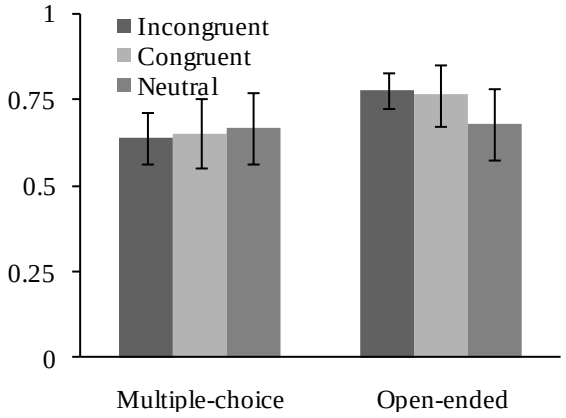


Figure 4: Mean proportion of correct recall as a function of response mode and problem type. Error bars display standard error.

## Summary and Discussion

A commonly used base-rate problem was adapted in the current experiment to manipulate the constraints of the task context. Adults participated in one of two conditions that differed only in whether the base-rate problems had a constraining multiple-choice response mode, or a less constraining, opened-ended response mode. Patterns of performance across base-rate problems differed markedly as a function of our manipulation.

In the multiple-choice response mode participants demonstrated the classical base-rate neglect without being affected by superficial changes in the order in which the information was presented to them. Conversely, in the open-ended response mode participants neglected base-rate information only in one of the order conditions, when base-rate information was presented after individuating information (Order 2). In the reverse order, when base-rate information was presented before individuating information (Order 1), participants took base-rate information into account.

Note that Order 1 is the common way in which information was presented to participants in previous research (e.g., De Neys & Glumicic, 2008; Kahneman & Tversky, 1973). Accordingly, we did indeed replicate the previous findings when the multiple-choice response mode was used. But when the response mode was less constraining, the superficial changes in order made a difference in performance. Performance in the recall task provides reason to doubt the possibility that these differences are spurious effects of some sort. Participants in all groups performed above chance on the recall task, independently of how the information was presented in the decision-making problems.

The results of the present investigation underscore the idea that performance cannot be uniquely attributed to cognitive structures or processes. Any plausible cognitive structure that could be responsible for the current findings would be post-hoc and rather complex, given that even irrelevant changes in order affected performance. A constraints view, in contrast, could readily explain our results. It predicts, a priori, that the tightening of degrees of freedom cuts down on idiosyncratic variability in performance and the impact of seemingly superficial factors. Our findings suggest that adults are neither rational nor irrational reasoners. Instead, their performance reflects a coupling with the task, and thus says as much about the task as about the reasoner.

## Acknowledgments

The authors thank Sue Collins for her help with scoring the data. Writing of the manuscript was supported by a grant from the National Science Foundation (DRL # 723638) to Heidi Kloos.

## References

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248-1299.

Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, *75*(4), 451-468.

Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321-339.

Evans, J. St. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*(3), 197-213.

Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language, and understanding.* New York: Cambridge University Press.

Gibson, J. (1979). *The ecological approach to visual perception.* Dallas: Houghton Mifflin.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107-143.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction. *Psychological Review*, *102*(4), 684-704.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, *84*(3), 343-352.

Kahneman, D., & Frederick, S. (2002). Representativeness revisted: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*. U.S.A.: Cambridge University Press

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251.

Kloos, H. & Van Orden, G. C. (2009). Soft-assembled mechanisms for the unified theory. In J.P. Spencer, M. Thomas, & J. McClelland (Eds.), *Toward a New Grand Theory of Development? Connectionism and Dynamics Systems Theory Reconsidered*. Oxford: Oxford University Press.

Kloos, H., & Van Orden, G. C. (in press). Voluntary performance on cognitive and motor tasks. *Mind and Matter*.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, *32*(5), 932-943.

Stanovich, K. E., & West, F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645-665.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? *Cognitive Science, 25*, 111–172.

Van Orden, G. C., & Kloos, H. (2003). The module mistake. *Cortex, 39*, 164–166.