Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences 107

The Role of High-Level Reasoning and Rule-Based Representations in the Inverse Base-Rate Effect

BY

PIA WENNERHOLM

ACTA UNIVERSITATIS UPSALIENSIS
UPPSALA 2001

Dissertation for the Degree of Doctor of Philosophy in Psychology presented at Uppsala University in 2001

ABSTRACT

Wennerholm, P. 2001. The Role of High-Level Reasoning and Rule-Based Representations in the Inverse Base-Rate Effect. Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 107. 68 pp. Uppsala. ISBN 91-554-5178-0.

The *inverse base-rate effect* is the observation that on certain occasions people classify new objects as belonging to rare categories rather than common ones (e.g., Medin & Edelson, 1988). This finding is inconsistent with normative prescriptions of *rationality*, and provides an anomaly for current theories of *human knowledge representation*, such as the exemplar-based models of *categorization*, which predict a consistent use of base-rates. This thesis presents a novel explanation of the inverse base-rate effect. The proposal is that participants sometimes *eliminate* category options that are inconsistent with well-supported *inference rules*. These assumptions contrast with those of *attentional theory* (e.g., Kruschke, 1996), according to which the inverse base-rate effect is the outcome of *rapid attention shifts* operating on cue-category *associations*. Study I, II, and III verified seven qualitative predictions derived from the eliminative inference idea. None of these phenomena can be explained by attentional theory. The most important of these findings were that elimination of well-known, common categories mediate the inverse base-rate effect rather than the strongest cue-category associations (Study I), that only participants with a rule-based mode of generalization exhibit the inverse base-rate effect (Study II), and that rapid attentional shifts *per se* do not accelerate learning, but rather decelerate it (Study III). In addition, Study I provided a quantitative implementation of the eliminative inference idea, ELMO, that demonstrated that this high-level reasoning process can produce the basic pattern of base-rate effects in the inverse base-rate design. Taken together, the empirical evidence of this thesis suggest that rule-based elimination is a powerful component of the inverse base-rate effect. But previous studies have indicated that attentional shifts affect the inverse base-rate effect, too. Therefore, a complete account of the inverse base-rate effect needs to integrate inductive and eliminative inferences operating on rule-based representations with attentional shifts. The Discussion of this thesis proposes a number of suggestions for such integrative work.

Key words: Inverse base-rate effect, rationality, human knowledge representation, exemplar-based models of categorization, eliminative inference, inference rules, rapid attention shifts, associations.

Pia Wennerholm, Department of Psychology, Uppsala University, Box 1225, SE-751 42, Uppsala, Sweden

ISSN 0282-7492
ISBN 91-554-5178-0

Printed in Sweden by Universitetstryckeriet, Ekonomikum, Uppsala, 2001

To my beloved Patrik

To know that you know,
and to know that you don’t know
- that is the real wisdom.

(Confucius, 551 - 479 B.C)

This thesis is based on the studies listed below, which will be referred to in the text by their Roman numerals:

I. Juslin, P., Wennerholm, P., & Winman, A. (2001). High-level reasoning and base-rate use: Do we need cue-competition to explain the inverse base-rate effect? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*, 849-871.

II. Wennerholm, P., Winman, A, & Shanks, D. R. (2001). *Reasoning or associations? Cognitive representations in the inverse base-rate task.* Manuscript submitted for publication.

III. Winman, A., Wennerholm, P., & Juslin, P. (2001). *Can attentional theory explain the inverse base-rate effect? Comments on Kruschke (in press)*. Manuscript submitted for publication.

Reprints were made with kind permission from the American Psychological Association.

Acknowledgements

Without a doubt, the past five years of training have been the most challenging and rewarding years of my educational life. Most of this training I owe to my supervisor, Professor Peter Juslin, whose enthusiasm and ambitions in research has had an energizing effect on my scientific efforts and achievements. Peter has taught me to stay focused on the ”crucial issues” and not get lost in irrelevent inquiries. No less important, he has been supportive and encouraging in times of frustration and distress. I could not have asked for a better mentor!

I'm also indebted to Anders Winman, my dear colleague and co-worker, for computer-programming and countless discussions on the subject matter presented in this thesis. Without Peter and Anders this thesis would not be what it has become. I owe you guys particularly many and warm thanks!

I'm also grateful to a number of people who read and commented on a preliminary version of this thesis: Andreas Birgegård, Henrik Olsson, Mats Olsson, and David R. Shanks. I also wish to thank my graduate and senior colleagues at the Department of Psychology, the staff at the Psychology section of the Uppsala University Library and the technical and administrative staff at the Department of Psychology for a friendly and supportive working environment.

For financial support I thank the Swedish Council for Research in the Humanities and Social Sciences, the Bank of Sweden Tercentenary Foundation, the Non-Graduated Researchers Fund at Uppsala University, and the Wallenberg Fund at Uppsala University.

On a more personal note, I'm grateful to my “bosom friend” Sari Jones who has been helpful and supportive in all sorts of ways, both professionaly and privately. She has also helped me to stay in physical shape (which is worthwhile even for academics)! ”I can, I want to, I will” (to quote one of our fitness instructors) is a good thing to remember, not only when you face 60 minutes of fitness training. Åse Haag, Ingrid Israelsson Olsson, and Maria Tillfors have been great friends and supporters too!

I also wish to thank my parents, Hans and Monika; my father for persuading me to start studying at the University in the first place, and my mother for convincing me to follow my heart and study what I was interested in.

Finally, I'm grateful to Patrik, my best friend and beloved partner, who has definitely been the most valuable person during the past five years of my private life, and hopefully much, much longer.

Uppsala 2001-10-29
Pia

Contents

1. Introduction

People make decisions every day, both in their professional and private lives. Sometimes the decisions are not as good as one could hope for, such as when a patient's condition progressively gets worse because the attending physician mistakenly made the wrong diagnosis, or the next-door neighbor erroneously puts the blame on "those immigrants" upon discovering that his or her house has been burgled. Irrespective of the particular domain of interest, almost all decisions people make are based on uncertain or ambiguous information; several potential diseases may be associated with a particular set of symptoms; burglaries in a particular neighborhood may be carried out by mainly native and/or immigrant perpetrators. These "background" data, which define the *base-rates* or relative frequencies of events, and somehow are encoded and processed for retrieval in people's minds, have important implications for the likely success of their judgments and decisions.

Early research on people's ability to incorporate base-rates into their judgments suggested that they are not very good at this task. People's judgments were compared to normative, Bayesian prescriptions (detailed in Section 1.2.1), and observed deviations were seen as evidence of irrational judgments (for a review, see Koehler, 1996). Together with a number of other so-called judgmental biases in the literature, these findings were interpreted as suggesting a "blemished portrait of human capabilities" (backcover of the anthology by Kahneman, Slovic, & Tversky, 1982). For example, in the classic "lawyer-engineer" experiment by Kahneman and Tversky (1973), participants were told that five personality descriptions, based on psychologists' personal interviews and personality tests, had been randomly drawn from a pool of descriptions of 100 people. Half of the participants were then told that the pool consisted of 30 lawyers and 70 engineers, whereas the other half received the reverse base-rate information; that the pool contained 70 lawyers and 30 engineers. The participants' task was to assess the probability that each of the five persons either was an engineer or a lawyer. The results demonstrated that the mean probability judgments only differed with 5% between the two groups, which was interpreted as implying that the participants had failed to incorporate the base-rate information concerning the professions into their judgments. By the beginning of the 1980's, the empirical message concerning people's ability to use base-rates seemed clear: "The genuineness, the robustness, and the generality of the base-rate fallacy[1] are matters of established fact" (Bar-Hillell, 1980, p. 215).

Eventuelly, however, methodological objections toward these studies were raised, primarily because people's base-rates were equated with the verbally presented "summary statistics" given to them when making the decisions. The question was whether this information actually could be said to represent people's base-rate knowledge. In many contexts base-rate information is conveyed through experience, for example a physician acquires it to a large extent through his or her practice with patients. Thus, knowledge about category base-rates might be implicitly gained through experience with exemplars of categories (e.g., Hasher & Zacks, 1984). Similarly, directly

[1] In the literature, the concepts "base-rate fallacy" and "base-rate neglect" are used synonumously and refer to the same empirical phenomenon, namely that people underemphasize or fail to use the base-rate information available to them when making decisions.

experienced base-rates may be accorded more weight than indirectly experienced base-rates because they invoke an implicit rather than an explicit learning system (e.g., Holyoak & Spellman, 1993; Shanks & St John, 1994). When the implicit learning experience comes in the form of trial-by-trial learning, the information at each trial may be encoded as a separate memory trace, as is assumed in many exemplar-based models of categorization (detailed in Section 1.2.2).

In line with this hypothesis, some researchers reported experiments showing appropriate use of base-rates. For example, a study by Christensen-Szalanski and Beach (1982) reported that physicians who learned the low base-rate for pneumonia from their clinical experience relied heavily on this base-rate information when making their diagnosis (for similar results, see e.g., Butt, 1988; Nelson, Biernat, & Manis, 1990). In a similar vein, some researchers demonstrated that people can use base-rates if they are reframed in terms of frequencies (i.e., natural numbers) instead of probabilities (e.g., Cosmides & Tooby, 1996; Gigerenzer, 1994; Gigerenzer & Hoffrage, 1996). However, a third line of research on the role of base-rates in category learning demonstrated that participants fail to use base-rates (e.g., Gluck & Bower, 1988), or even use them in an inverse manner (Medin & Edelson, 1988), despite the fact that they are experienced in a trial-by-trial manner. These latter findings raised serious concerns because if the encoding of frequency is automatic, then the encoding of base-rates which are simply relative frequencies should be automatic too.

Suppose, for example, in the case of a physician, that he or she knows that the patient's symptoms are associated with two diseases, one of which appears a hundred times more often than the other (i.e., the diseases have different base-rates). Under these circumstances it would be quite surprising if the physician thought that the rare disease would be the more likely diagnosis. Similarly, in the case with the neighbor, suppose that he or she knows that most burglaries in the neighborhood are perpetrated by natives and *not* by immigrants, would it not be surprising then if your neighbor would insist on an immigrant perpetrator? Nevertheless, in some situations this is precisely what research on the experiential impact of base-rates suggests; a counter-intuitive finding known as the *inverse base-rate effect* (Fagot, Kruschke, Depy, & Vauclair, 1998; Kruschke, 1996; in press; Medin & Bettger, 1991; Medin & Edelson, 1988; Shanks, 1992). More specifically, simultaneously presenting participants in an experimental paradigm with a conflicting compound consisting of a Perfect predictor of a Rare outcome, *PR*, and a Perfect predictor of a Common outcome, *PC* (i.e., *PC.PR*) (the dot indicates co-occurrence), this effect is evidenced by participants' preference of the rare outcome, *R* (see Table 1).

This result is incompatible with normative prescriptions of rationality and has theoretical implications for current theories of categorization (detailed in Section 1.2). Therefore, a number of researchers have developed explanations for it (Anderson, 1990; Gluck, 1992; Gluck & Bower, 1988; Kruschke, 1996; in press; Medin & Bettger, 1991; Medin & Edelson, 1988; Shanks, 1992). Most of these accounts revolve around the associationist concept of *cue-competition*; cues with the same objective contingency become differently associated with the outcomes (but see Anderson, 1990). In the inverse base-rate paradigm (detailed in Section 1.1 below), this implies that the Perfect predictor of the Rare outcome, *PR*, gains more associative strength than the Perfect predictor of the Common outcome, *PC*, during the learning phase (cf. Table 1).

In attentional theory, the most successful of these associationist accounts, cue-competition is formalized as a *rapid attention shifting mechanism*, an idea first formalized by Mackintosh (1975), but which has more recently been implemented in the connectionist models ADIT (Kruschke, 1996) and EXIT (Kruschke, in press). According to this theory, rapid attention shifts produce a stronger encoding of the *PR-Rare outcome* association than the *PC-Common outcome* association. The consequence of this encoding asymmetry is evidenced by participants responding on the rare outcome, *R*, when presented with the conflicting compound, *PC.PR*.

The primary question addressed in this thesis is whether associative principles in general, and the notion of rapid attention shifts operating on cue-category associations in particular, provide the only or most important account of the inverse base-rate effect. An alternative explanation is presented which rests on a higher-order reasoning mechanism, operationalized in terms of *eliminative inference*. According to this idea, people's knowledge is not asymmetric due to attentional shifts causing encodings of differential strengths during learning. Instead it is proposed that people form inference rules during the learning phase of the inverse base-rate task that are used in a flexible and controlled manner to both induce and eliminate category membership in the transfer phase of this task. Thus, when confronted with the conflicting compound, *PC.PR*, participants will eliminate the common outcome, *C*, that has always been associated with $I.PC \rightarrow C$ and choose the more vaguely known rare outcome, *R*, instead.

In the remainder of this thesis, I will present data that are inconsistent with the present formalization of attentional theory, but consistent with the eliminative inference approach. The results indicate that the eliminative inference mechanism may be a significant contributor to the inverse base-rate effect. Before proceeding to the details of the empirical results, however, I will outline the inverse base-rate paradigm in more detail, and explain why research on the inverse base-rate effect is important. I will present the previous theoretical accounts of the phenomenon, as well as the novel account based on the eliminative inference mechanism. I will derive a number of novel qualitative predictions from the eliminative inference mechanism that in critical respects deviate from the predictions by attentional theory (Kruschke, 1996, in press). Then, I will present Study I, II, and III which test the predictions. Finally, I will discuss and evaluate the results in relation to contemporary theorizing on reasoning and categorization.

1.1. The Inverse Base-Rate Paradigm

Each time a decision is made it is based on the decision maker's previously acquired knowledge. The information building up this knowledge is somehow encoded, organized, and processed in his or her mind and subsequently retrieved for thinking, conversation and/or decision-making. The challenge facing cognitive scientists is to figure out *how* these processes take place and *why* professionals and laypeople alike sometimes reach suboptimal decisions (such as the inverse base-rate effect) as a result of these processes. However, because decisions are often affected by other factors besides the purely cognitive ones, such as emotion, a common procedure is to strip away the naturally occurring richness of cues to a few controllable and easily analyzable components, without losing resemblance to the real world. Medin and Edelson (1988) created such a method for research on the use of base-rate information that is derived from experience in classifying examples of a category, a task that can also be thought of as an inductive problem-solving task.

Medin and Edelson (1988) required participants to imagine that they were physicians at a hospital diagnosing hypothetical patients suffering from a number of fictitious diseases. The task was divided into a learning phase and a transfer phase. On each training trial, a pair of symptoms were presented together with six fictitious diseases, and participants were requested to choose which of these diseases each hypothetical patient was suffering from (for simplicity in Table 1 below, one pair of events are presented instead of the original six, cf. Medin & Edelson, Experiment 1, 1988). After each choice the participant was informed about the proper diagnosis (i.e., disease) and then the next training trial followed. The critical manipulation concerned the base-rate (i.e., the relative frequency) of each disease, with the common diseases occurring three times more often than the remaining rare ones.

Specifically, during training every particular instance of a common disease, *C*, occurred in the presence of two symptoms: One imperfect predictor, *I*, and one perfect, *PC*. Similarly, every particular instance of a rare disease, *R*, was paired with two symptoms: One imperfect, *I*, and one perfect, *PR*. Thus, each *imperfect predictor* was associated with both a common and a rare disease, and each *perfect predictor* was uniquely associated with only one disease (see the learning phase in Table 1).

After training, participants were asked to apply their newly acquired knowledge onto a number of novel patients. In contrast to the training phase, in which the patients always exhibited two symptoms, the number of symptoms in the transfer phase could be one to three. Therefore, the symptoms were ambiguous in the sense that they had previously been associated with both a common and a rare disease (see the transfer phase in Table 1). Moreover, no feedback concerning the correctness of the answers was provided.

Table 1
Simplified Abstract Design of a Typical Inverse Base-Rate Experiment

Learning phase:		
Learning Frequency	Symptoms	Disease Category
3	$I_1.PC_1$	C_1
1	$I_1.PR_1$	R_1
Transfer phase: Tests for base-rate information		
1. Imperfect predictors: I_1		
2. Conflicting tests: $PC_1.PR_1$		
3. Combined tests: $I_1.PC_1.PR_1$		

Note. For concrete names on symptom- and disease names, see the General Method in Section 3.1 of this thesis. The dot between the symptoms indicates co-occurrence.

A number of studies (Fagot et al., 1998; Kruschke, 1996; in press; Medin & Bettger, 1991; Medin & Edelson, 1988; Shanks, 1992) have shown that these novel symptom combinations lead to an inconsistent pattern of base-rate effects: When tested with the imperfect probe, *I*, the majority of participants choose the common disease, in line with the underlying base-rates. When tested with the combined probe consisting of three symptoms, *I.PC.PR*, participants again tend to choose the common disease. However, when tested with two perfect predictors, *PC.PR*—the conflicting probe—the majority of participants choose the rare disease *in contrast, or inverse, to the base-rates.* It is this counter-intuitive finding that has been labeled the inverse base-rate effect.

These results have been replicated and extended in a number of ways and are not restricted to particular base-rates or to particular procedures. For example, Medin and Bettger (1991) obtained the inverse base-rate effect using a 3:1 ratio during the first half of training, and a 2:2 ratio during the second half, but not when the ratios were reversed (i.e., when the 2:2 ratio preceded the 3:1 ratio; see also Kruschke, Experiment 2, in press). They concluded that the early stage of learning is important for the inverse base-rate effect to emerge. Shanks (1992) performed a study where he used a between-subjects design featuring two different base-rate ratios, 3:1 and 7:1. He could only observe the effect with the 7:1 base-rate ratio and not with the 3:1 base-rate ratio, and discussed the possibility that perhaps the participants in the original study by Medin and Edelson attended more to the 3:1 base-rate ratio because it was the only information available to them. Fagot et al. (1998) performed a comparative study with human participants and baboons. A perceptual analogue of the medical problem-solving task was given to the participants, but the inverse base-rate effect was only found in humans and not in baboons.

1.2. Why is Research on the Inverse Base-Rate Effect Important?

Two fields of psychological inquiry are particularly noteworthy in regard to human reasoning and the inverse base-rate effect: The rationality of human thinking and the nature of human knowledge representation. Each of these fields will be discussed in the following Section.

1.2.1. The normative issue. Research into the role of base-rates in human decision-making uses normative, Bayesian prescriptions as a standard against which to evaluate the rationality (i.e., correctness) of people's judgments and decisions, and observed deviations are interpreted as evidence of irrational decisions. A common assumption is that the verbally presented base-rates and/or base-rates provided through direct experience defines the prior probability (see e.g., Koehler, 1996). Bayes's Theorem is a mathematical formula that specifies how prior probabilities should change in light of new evidence. The amount of change depends on the diagnosticity of the actual evidence. To illustrate, consider a participant in the transfer phase of the inverse base-rate task who is confronted with the imperfect probe, *I*, which has occurred in the presence of two diseases, one common and one rare disease, *C* and *R*, respectively (cf. Table 1). Under these circumstances, Bayes' Theorem predicts that the participant should choose the common disease, *C*, because the imperfect predictor, *I*, is equally diagnostic of both diseases.

$$\frac{p(C|I)}{p(R|I)} = \frac{p(I|C) * p(C)}{p(I|R) * p(R)} = 3 \qquad \text{(Equation 1)}$$

where $p(C|I)$ and $p(R|I)$ are the posterior probabilities of the common and rare diseases given the imperfect symptom, *I*, respectively, $p(I|C)$ and $p(I|R)$ are the probabilities of the imperfect symptom, *I*, given the common and rare diseases, respectively (i.e., the likelihood ratio), and $p(C)$ and $p(R)$ are the prior probabilities (i.e., the base-rates) for the common and rare diseases, respectively. If translated into actual probabilities $p(I|C)$ = $p(I|R)$ = 1, $p(C)$= 0.75 and $p(R)$ = 0.25, which equals 3 (cf. Equation 1). This prediction corresponds fairly well with human data for this probe. For example, in Experiment 1 by Medin and Edelson (1998), participants chose the common disease, *C*, as their diagnosis with probability .78.

The intriguing result in the inverse base-rate design, however, concerns the conflicting transfer probe, *PC.PR*, that gives rise to the inverse base-rate effect. For this probe there is no straight-forward normative principle by which the participants' responses can be ascertained to be "correct" or not. The reason for the unclear normative status is that there is no way of determining participants' estimates of $p(PC.PR|C)$ and $p(PC.PR|R)$. Despite this difficulty of applying a normative analysis, several researchers have interpreted the inverse base-rate effect as an irrational decision (but see Anderson, 1990), thereby implicitly assuming that the likelihood of the unknown probabilities $p(PC.PR|C)$ and $p(PC.PR|R)$ are equal (e.g., 1). This conclusion is not altogether insensible, although it would perhaps be more appropriate to consider the inverse base-rate effect as a counter-normative decision, since it is hard to imagine how a normative account could license choice of a rare disease. In contrast, according to the eliminative inference idea (detailed in Section 2 below), the preference of the rare category, *R*, when confronted with the conflicting probe, *PC.PR*, is fully rational. Symptom *PR* in the symptom configuration *PC.PR* has never occurred with the common disease during learning, *I.PC*$\rightarrow$*C*, and this well-known disease is therefore eliminated. As was shown in Table 1, the training phase is deterministic, and the transfer phase presents entirely novel symptom combinations; therefore, it is questionable whether Bayes' Theorem, which makes predictions for probabilities, can be applied to the inverse base-rate task.

In summary, most researchers have interpreted the inverse base-rate effect as an irrational or counter-normative decision, but the account presented in this thesis points to a different conclusion, that the inverse base-rate effect is in fact a rational response. At any rate, the normative issue is of considerable interest if we wish to come closer to an understanding of the experiential impact of base-rates on human judgment and categorization.

1.2.2. Limitations of exemplar-based models of categorization. Exemplar-based models of category learning assume that experience with members of a category leads to concrete representations (memory traces) of each encountered exemplar and predict a consistent use of base-rates (Estes, 1994; Kruschke, 1992; Lamberts, 2000; Medin & Schaffer, 1978; Nosofsky, 1986, 1987; Nosofsky & Johansen, 2000; Nosofsky, Kruschke, & McKinley, 1992; Nosofsky & Palmeri, 1997; Smith & Minda, 2000). For example, in the context model by Medin and Schaffer (1978) it is assumed that experiencing particular exemplars of a category leads to a memory trace of each exemplar being stored. If applied to the inverse base-rate task, this model implies that one set of *I.PC*→*C* traces would be stored along with a smaller set of *I.PR*→*R* traces. On presentation of a new transfer probe, such as a patient with symptom *I*, the decision to classify the exemplar (i.e., the patient) as belonging to category *C* or category *R* would be governed by the summed similarity between symptom *I* and the *I.PC*→*C* traces versus the summed similarity between symptom *I* and the *I.PR*→*R* traces. Because there are more traces of symptoms *I.PC* in memory, the transfer probe *I* will have greater summed similarity to *I.PC* than to *I.PR*, and therefore category *C* will be chosen in preference to category *R*. However, on presentation of the conflicting probe, *PC.PR*, participants prefer category *R* (Fagot et al., 1998; Kruschke, 1996, in press; Medin & Bettger, 1991; Medin & Edelson, 1988; Shanks, 1992). Thus, a categorization model, such as that of Medin and Schaffer (1978), is unable to account for Medin and Edelson's data. In fact, together with apparent base-rate neglect (e.g., Gluck & Bower, 1988; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989), the inverse base-rate effect is one of the most provocative challenges to the otherwise extremely successful exemplar-based models of categorization.

1.2.3. Human knowledge representation: Rule-based or similarity-based? The I inverse base-rate effect is of considerable interest from the viewpoint of learning research where theorists debate about the relative importance of rule-based versus similarity-based representations (see e.g., Brooks, 1978, 1987; Logan, 1988; Medin & Ross, 1989; Medin & Smith, 1981; Palmeri, 1997; Rips, 1989; Ross, 1987; Shanks & St. John, 1994; Sloman, 1996; Smith & Sloman, 1994; Smith, Langston, & Nisbett, 1992). The rule-based system is symbolic and describes the world by capturing different kinds of structure, structure that is logical, hierarchical and causal-mechanical. The similarity-based system comprises exemplar-based and associative retrieval and its computations reflect similarity, temporal structure, and featural overlap. More generally, rule-following occurs because there is a one-to-one correspondence between the symbols of the rule and the components of the mental event (for more precise criteria of rule-following, see Smith et al., 1992). The mental event can be coded as an instantiation of an abstraction, and this abstraction can then be applied to a novel situation. Most important, rules have a logical structure and a set of variables.

To exemplify the distinction between rules and associations, consider the *blocking effect*, a learning phenomenon reported by Kamin (1969). Animals first learn a predictive relation between a conditioned stimulus (*A*) and an unconditioned outcome (outcome *O*). Once the *A*→*O* relation has been well-learned, a second conditioned stimulus (*B*) is presented in conjunction with the first stimulus (*A*) and the pair is reinforced (i.e., *A.B*→*O*). When tested later only the first stimulus evokes responding despite the fact that the conditional probability of observing *O* has been 1.0 both after observing *A* and *B*, that is *O* has always followed presentation of *A*, but also presentation of *B*. The explanation for this effect according to the rule-based view is that a rational participant would more or less explicitly disregard *B* because that cue is perfectly confounded with the highly valid cue *A* in the participant's experience, i.e., there is little ground for forming a strong belief about *B*, given that it covaries with *A*. In other words, the first cue-outcome pair (i.e., *A*→*O*) is coded and processed as an instantiation of an abstraction. Information that is coded abstractly can be assimilated to an abstract rule; If *A* then *O*; *A* therefore *O*. Thus, according to the rule-based approach people extract contingency information by applying rules and then use these learned relationships to test causal hypotheses in a manner equivalent to a scientist (e.g., Cheng, 1997).

Proponents of the associative view would object to the assumption that the blocking effect is caused by anything as abstract as the acquisition and subsequent access of a rule. Rather, they would propose that the blocking effect is the result of conditioned associations between two contiguous events. Indeed, most contemporary theories of animal and human learning are of this kind (e.g., Kruschke, 1992; Mackintosh, 1975; McClelland & Rumelhart, 1985; Pearce, 1987; Pearce & Hall, 1980; Rescorla & Wagner, 1972; van Hamme & Wasserman, 1994). For example, Rescorla and Wagner (1972) suggested that cues consist of sets of elements, each of which forms an independent association with the outcome event. Novel stimuli excite the representation of the outcome to the extent that they are composed of elements previously associated with the outcome. The blocking effect reported by Kamin can be explained by Rescorla and Wagner's analysis of cue-competition: Once an $A \rightarrow O$ association is established, any new cue (*B*) that is presented together with the previous cue (*A*) without improving the outcome contingency (i.e., $A.B \rightarrow O$) will fail to be associated with the outcome. Thus, due to competition between the cues, the connection between $A \rightarrow O$ automatically ”blocks out” the association between *B* and the outcome *O* (see also Mackintosh, 1975).

Thus, one difference between similarity-based and rule-based approaches to human knowledge representation comes down to how abstractly instantiations are encoded and retrieved. Another distinction focuses on the levels of awareness (see Table 2).

Table 2

General Cognitive Psychological Distinctions between Rules, Exemplars, and Associations[2]

Cognitive Psychological Distinctions					
Mental representation	Knowledge system	Level of stimulus abstraction	Level of awareness	Memory system/ cognition	Processing: Controlled vs. automatic
Rule	Rule-based	High	High	Semantic /explicit	Controlled
Exemplar	Similarity-based	Low	Medium	Episodic /explicit	Automatic
Association	Similarity-based	Medium	Low	Procedural /implicit	Automatic

[2] Note that Table 2 only is an efficient and simple way of dividing different cognitive psychological distinctions. It does not imply that, for example, a distinction in one column necessarily maps absolutely symmetrically to a distinction in another column. For example, some researchers on memory would probably disagree with the view that episodic memory is automatic since it is assumed to be phylogenetically new and “fragile”.

Reber (1993) has proposed that human knowledge representation depends on two distinct cognitive systems: Explicit and implicit cognition. He argues that implicit cognition evolved long before high-level consciousness and has developed a number of criteria which can be used to distinguish the two systems from one another: (1). Implicit systems should be more robust than explicit systems, operating despite injuries, diseases, and other disorders. (2). Implicit processes should be more age independent, revealing fewer differences than explicit processes in both infancy and old age. (3). Implicit processes should be IQ independent. (4). Implicit processes should show lower population variance than explicit processes. (5). Implicit processes should show across-species commonalities. Despite these criteria, research has shown that inferential rules in reasoning can be more or less conscious. Some inferential rules may be applied only unconsciously (Nisbett & Wilson, 1977), whereas others may be applied some of the time with a recognition that the rule is being used[3] (e.g., Shanks & Darby, 1998).

Many investigators maintain that the rule-based and similarity-based systems are indeed distinct (e.g., Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Palmeri, 1997), whereas others argue that the mind is composed of a single system (e.g., Nosofsky & Johansen, 2000). However, recent neuroscientific work on category learning has revealed a neuroimaging contrast between rule-based and similarity-based processing, which supports the modular view (Smith, Patalano, & Jonides, 1998). Interestingly, these results fit nicely with research on memory, which suggests that human memory is composed of several systems (see e.g., Schacter, 1990; Schacter & Tulving, 1994; but see also Roediger & McDermott, 1993). With this distinction in mind, the rule-based system would belong to the semantic system, which is concerned with memory for facts. The similarity-based system is more complex because it encompasses both exemplars (episodic memory) and associations (procedural memory), but, of course, none of these distinctions are clear-cut (see Table 2).

More recently, Shanks and Darby (1998) have reported results from discrimination learning demonstrating rule-based and similarity-based generalization as a function of learning performance in the same experimental procedure. These data point to individual differences in human knowledge representation.

1.2.3.1. Individual differences in human knowledge representation. Shanks and Darby (1998) presented participants with a *patterning task* commonly used in research on human causal learning (see also Lachnit & Kimmel, 1993). They were required to learn associations between various foods (i.e., cues) that hypothetical people ate and various allergic reactions that ensued (i.e., outcomes). In the training stage, participants were presented with a negative patterning structure consisting of elements that individually predicted the outcome (e.g. $A \rightarrow O$ & $B \rightarrow O$), but failed to predict the outcome when presented together as a compound ($A.B \rightarrow$ *no O*). The same participants were also presented with a positive patterning structure for which the roles of the elements and compounds were reversed (e.g., *A & B → no O but A.B → O*). After each response, feedback was provided about the correct outcome.

[3] Part of the reason for these inconsistent results may stem from how one chooses to define a rule (for a discussion, see Johansen & Palmeri, 2001).

In a subsequent transfer phase, participants were tested with parts of other incomplete patterning structures with each problem being instantiated with different foods. The critical issue was how people would respond to a number of novel compounds. For example, participants were presented with the novel compound *I.J* after having been trained on the elements *I* → *O* and *J* → *O*. If they generalized on the basis of rule-based reasoning, they should predict the absence of the allergy outcome on *I.J* trials, having learned the rule "*a compound and its elements predict opposite outcomes*". However, if they generalized on the basis of similarity-based reasoning, such as exemplar-based or associative retrieval, they should predict the occurrence of the allergy on *I.J* trials (see Table 3).

Table 3
Trial Types in the Learning- and Transfer Phases in Shanks and Darby (Experiment 2, 1998)

<table>
<tr><th colspan="3">Learning Phase</th></tr>
<tr><td>A → O</td><td>B → O</td><td>A.B → no O</td></tr>
<tr><td>C → no O</td><td>D → no O</td><td>C.D → O</td></tr>
<tr><td>E → O</td><td>F → O</td><td>E.F → no O</td></tr>
<tr><td>G → no O</td><td>H → no O</td><td>G.H → O</td></tr>
<tr><td>I → O</td><td>J → O</td><td></td></tr>
<tr><td></td><td></td><td>K.L → no O</td></tr>
<tr><td>M → no O</td><td>N → no O</td><td></td></tr>
<tr><td></td><td></td><td>O.P → O</td></tr>
<tr><th colspan="3">Transfer Phase: Tests for generalization mode</th></tr>
<tr><td>A?</td><td>B?</td><td>A.B?</td></tr>
<tr><td>C?</td><td>D?</td><td>C.D?</td></tr>
<tr><td>E?</td><td>F?</td><td>E.F?</td></tr>
<tr><td>G?</td><td>H?</td><td>G.H?</td></tr>
<tr><td>I?</td><td>J?</td><td>I.J?</td></tr>
<tr><td>K?</td><td>L?</td><td>K.L?</td></tr>
<tr><td>M?</td><td>N?</td><td>M.N?</td></tr>
<tr><td>O?</td><td>P?</td><td>O.P?</td></tr>
</table>

Note. *A* - *P* are foods, *O* is the allergy, and *no O* is no allergy. See Experiment 5 in Section 3.3.3 in this thesis for concrete examples of foods. The dot between the foods indicates co-occurrence.

The results demonstrated that the different modes of generalization were linked to people's learning performance. When participants were divided into efficient and inefficient learners on the basis of their learning accuracy, the results demonstrated a compelling cross-over effect: The responses for the inefficient learners were consistent with similarity-based responding, whereas the responses for the efficient learners were consistent with rule-based responding. Shanks and Darby (1998) discussed the possibility that with enough training all participants would become rule learners[4]. They also

[4] Interestingly, this conclusion contradicts the proposal by Logan (1988) who argued that automaticity in cognition is a function of the shift from strategic and algorithmic processes, such as the use of explicit rules, to the retrieval of specific stored exemplars of previous solutions to cognitive problems.

noted that other variables might affect the ease of rule learning, such as psychometric intelligence (McGeorge, Crawford, & Kelly, 1997).

Despite the evidence for individual differences in human knowledge representation, most researchers have applied similarity-based principles to human reasoning. In attempting to describe the inverse base-rate effect, associative concepts such as cue-competition have dominated. These explanations of the inverse base-rate effect support the hypothesis that humans, like animals, are governed by similarity-based rather than rule-based processes. If this claim is correct, humans should be the victims of a number of systematic misunderstandings about the relationships between cues and outcomes in the environment (see e.g., Kruschke & Johansen, 1999). The critical question for present purposes is whether this proposal is correct or not.

1.2.4. Modeling the human mind: Symbols and/or connections? As addressed in the previous section, it is still unclear what knowledge representations best describe the human mind, and, of course, this problem transfers to the domain in which theorists attempt to model such processes and representations. There are two contemporary approaches to computational modeling: The traditional symbolic kind and connectionism. Each of these can be regarded as a category of models with large family resemblances. The focus here will be on the assumptions that are common to most of them. In general, however, they range from uniform theories which hold that only symbolic models can account for cognitive processes, through hybrid theories which resort to both connectionist and symbolic accounts for differing aspects of cognition, to uniform theories which seek purely connectionist accounts of the fundamental aspects of cognition.

Traditional symbolic computationalism views the human mind as a symbol-manipulating device, analogous to the computer (e.g., Newell & Simon, 1972). A symbol is a locally available code that can provide access to distal information that is relevant to a particular task (Smith et al., 1992). The symbols can be patterns of any kind, such as external pictorial objects or conceptual configurations. The prerequisites for thinking are that the symbols can be stored and manipulated, and behavior is explained by rule-like operations on these specific symbol structures (Simon, 1996). This approach to computational modeling has been fruitful in many areas of cognition, but not all. Part of the reason for the failures is that simulations of symbolic processes only seem to work for well-specified problem domains in a conscious sequential fashion (such as problem-solving), whereas humans can do many things in parallel without explicit awareness (such as talking while driving a car). Some theorists argue that these problems are better captured by connectionist systems.

Since the 1980s, connectionism, or parallel distributed processing (PDP) has challenged computationalism in its classical form (Smith et al., 1992). Unlike traditional computational models, there is no need for a symbolic level of representation in such systems, but cognitive activities are represented “directly” over distributed representations. In this way representations are stored by a pattern of activation throughout the connectionist network. These traces cannot be measured explicitly and therefore connectionist models are often referred to as intuitive or subsymbolic processors, in contrast to symbolic systems which are often referred to as rule-interpreters (see e.g., Ellis & Humphreys, 1999; Fodor & Pylyshyn, 1988).

Connectionist systems usually consist of multiple units (or nodes) that are neuron-like in character and organized in layers. The units are small processing elements, which compute in parallel. For each connection there is a negative (inhibitory) or positive (excitatory) numerical value that determines the influence of the sending unit on the receiving unit. Each unit typically takes the weighted sum of all of its input links, and produces a single output to another unit if the weighted sum exceeds a certain threshold value. The network learns the association between the different inputs and outputs by modifying the weights on the links between units in the net (e.g., back-propagation; Rumelhart, Hinton, & Williams, 1986). This process repeats itself until the net produces a required output pattern given a certain input pattern. Thus, the model can be made to learn what is explicitly programmed in traditional symbolic models. This kind of learning by incremental weight adjustment over distributed representations is reminiscent of associationist conceptions of learning (Estes, 1991).

Altogether then, the reason why research on the inverse base-rate effect is important, can be summarized in the following three statements: It is incompatible with the normative view that human judgments are rational, it provides an anomaly for the otherwise extremely successful exemplar-based models of categorization, and it suggests that human knowledge representations are distorted due to associative processes such as cue-competition. This latter process (cue-competition) has been emphasized in most previous accounts of the inverse base-rate effect, and many theorists have implemented it in models with connectionist architectures. Next, we will turn to these accounts as well as a number of others.

1.3. Previous Accounts of the Inverse Base-Rate Effect

Several explanations of the inverse base-rate effect have been proposed. For example, Medin and Edelson (1988) presented a revised version of the original context model (Medin & Schaffer, 1978) that could explain their intriguing base-rate results. This model involves two processes. The first process, the competition principle, addresses the idea that symptoms compete to predict a disease. To explain the inverse base-rate effect, Medin and Edelson (1988) maintained that when feedback about the incorrectness of a response is provided, the participant may pay attention to in what way this situation deviates from the preceding one. Because *I.PC* is a more common pattern than *I.PR*—and therefore generally encountered first—this implies that symptom *PR* achieves more associative strength with the rare disease, *R*, than symptom *PC* does for the common disease, *C*. The second process, the principle of context change, embodies retrieval failure induced by changes in context from that of learning to that of transfer. For example, if participants have learned that the pair *I.PC* is associated with the common disease, *C*, and that symptom *PR* is associated with the rare disease, *R*, then symptom *PR* will suffer less from the change in context associated with the conflicting transfer probe, *PR.PC*, than will symptom *PC*.

Although Medin and Edelson (1988) argued that these two processes provided a good account of their data, they did not present a quantitative model. They did, however, propose that the competition principle is qualitatively similar to competitive learning models, such as the Rescorla-Wagner model (R-W, e.g., Rescorla & Wagner, 1972), and suggested that the inverse base-rate effect could emerge from the R-W model as a pre-asymptotic effect (Medin & Edelson, 1988, p.75).

Gluck and Bower (1988) proposed a similar account by incorporating the R-W model into their adaptive network model. They argued that the inverse base-rate effect emerges as a result of the differential associative strengths of the perfect predictors, *PC* and *PR*. However, Markman (1989) showed that modeling the inverse base-rate effect by application of this learning rule is inadequate, and also demonstrated that even the addition of a layer of hidden units to the network, and the use of the back-propagation algorithm (Rumelhart et al., 1986) still fails to reproduce the inverse base-rate effect.

Gluck (1992) noted that the R-W model predicts that with extended training the perfect predictors, *PC* and *PR*, would gain all the predictive strength, leaving the imperfect predictor, *I*, with none. Instead, he proposed that a variant of the adaptive network model (Gluck & Bower, 1988), incorporating the distributed cue representation of stimulus sampling theory (Atkinson & Estes, 1963) could explain the inverse base-rate effect. However, this modified model only exhibited a modest preference for the rare disease, *R*, on the conflicting probe *PC.PR*, and a similarly modest preference for the common disease, *C*, on the combined probe, *I.PC.PR*. In addition, no quantitative fits were presented.

Shanks (1992) proposed that a variant of the component cue model (ACM, for the Attentional Connectionist Model) inspired by Wagner (1978) could generate the inverse base-rate effect. Although this algorithm was successful when applied to the conflicting test case, *PC.PR*, it was never applied to the equally critical imperfect, *I*, and combined probes, *I.PC.PR*. Furthermore, ACM is unable to account for recent data reported by Kruschke (1996, Experiment 1).

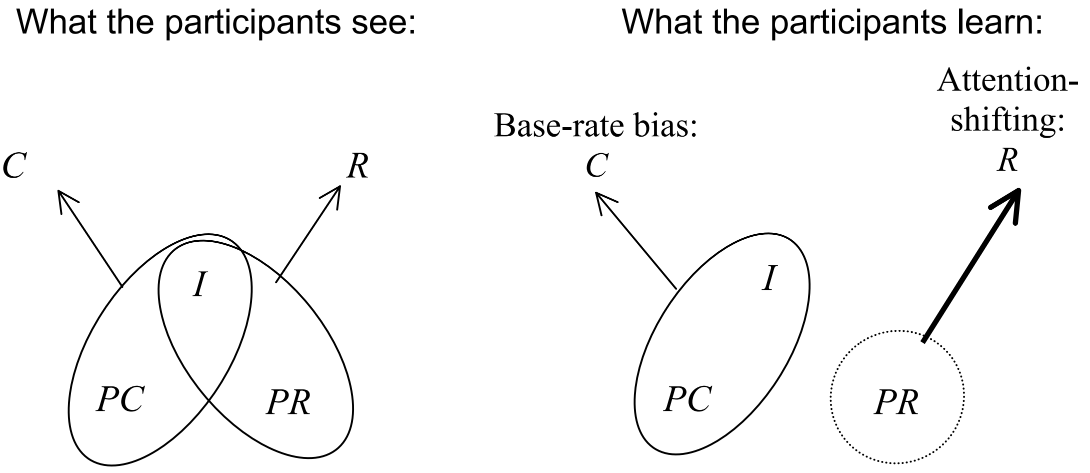
According to the Rational Model (Anderson, 1990) that focuses on the diagnostic value of mismatching features, the reversal of base-rates is an outcome of mismatches detected when the conflicting test case, *PC.PR,* is matched against the category for a disease, where mismatches are weighted more seriously for common diseases[5]. Although the Rational Model provides a satisfactory account of two out of three probes in the Medin and Edelson (1988) design, it fails to explain why people choose the common disease, *C*, when presented with an imperfect symptom paired with a perfect predictor of a common-other disease, *I.PCo* (For details of such an experimental design, see Kruschke, 1996).

Kruschke (1996, in press) reported strong inverse base-rate effects, and presented attentional theory, implemented in ADIT and EXIT (two connectionist models), that could explain both this phenomenon, and apparent base-rate neglect (Gluck & Bower, 1988). Because much of the empirical work presented in this thesis (detailed in Section 3 below) is concerned with contrasts between the predictions by attentional theory and the eliminative inference approach, this theory will given relatively more attention compared to the previously described accounts of the inverse base-rate effect.

1.3.1. Attentional theory. Attentional theory postulates mental representations in the form of cue-category associations and explains the inverse base-rate effect by rapid

[5] In fact, this model's explanation of the inverse base-rate effect largely resembles the logic of the eliminative inference idea (detailed in section 2 of this thesis). The main difference concerns the level of explanation. With Marr's (1980) distinction in mind, the Rational Model focuses on the computational level whereas the eliminative inference approach focuses on the algorithmic level.

attention shifts in the learning of those associations (Kruschke, 1996, in press). Specifically, it is proposed that during training participants apply their base-rate knowledge consistently on all trials (i.e., a base-rate bias). The inverse base-rate effect is caused by rapid shifts of attention away from symptoms that conflict with previous knowledge and toward distinctive features (e.g., symptom *PR*) to protect previously learned associations (i.e., attention-shifting). Because the common disease, *C*, occurs more often than the rare one, *R*, participants first learn to associate both *I* and *PC* with *C*. When later learning which symptoms are associated with the rare disease, *R*, they focus on the perfectly predictive symptom of this disease, namely *PR*, and thereby encode it by this single distinctive symptom. As a result, symptom *PR* is more strongly connected to disease *R* than symptom *PC* is to disease *C*, and this attentional assymetry explains why participants choose the rare disease on the conflicting transfer probe, *PC.PR*, that gives rise to the inverse base-rate effect (see Figure 1). When confronted with the remaining two critical transfer probes, *I* and *I.PC.PR*, people apply both their base-rate knowledge and their associative knowledge, where the base-rate knowledge dominates the response.


What the participants see:

What the participants learn:

Figure 1. What the participants see and learn in the training phase of the Medin and Edelson (1988) design as described by attentional theory (Kruschke, 1996, in press).

In the following, I will examine the psychological meaning of the attentional shifts in regard to developmental effects in children, and learning efficiency. This inquiry is motivated by a number of theory-critical predictions (detailed in Section 2.5) that are tested in the Empirical Section of this thesis.

1.3.1.1. What are rapid attention shifts? To derive à priori qualitative predictions from attentional theory in regard to the inverse base-rate effect, it is essential to define the key explanatory process according to this theory—rapid attention-shifting. The theoretical weight given to this concept is highlighted by the fact that it has come to be posited as a critical theoretical principle in accounting for human associative learning (Kruschke, 2001; Kruschke & Johansen, 1999). Despite this emphasis, psychological definitions of the term in the published articles which refer to it are meager (see Fagot et al., 1998; Kruschke, 1996, in press; Kruschke & Johansen, 1999). For example, in Fagot et al. (1998) rapid attention-shifting is described as “the ability of attention to rapidly shift away from one stimulus dimension to another, contingent upon prior learning” (p. 123). This is the only definition found in the texts.

The relatedness between attentional theory and several well-known learning models developed in relation to animal studies of classical and operant conditioning, is frequently discussed in the articles which refer to rapid attention shifts (Fagot et al., 1998; Kruschke, 1996, in press; Kruschke & Johansen, 1999). For example, Fagot et al. (1998), establish that ADIT (Kruschke, 1996), is an extension of the Rescorla-Wagner conditioning model of animal learning with the addition of rapid attention shifts. Similarly, in Kruschke and Johansen (1999), extensive cross-species comparisons are made, and the authors conclude that: “…many different species have been exposed to similar environmental demands for rapid learning, and many different species have therefore evolved functionally similar adaptations: rapid attention shifts that produce ’irrational behavior’” (p. 1084). In addition, the authors maintain that Mackintosh’s (1975) model of attentional learning in animals is close to a special case of ADIT.

The central idea in the model by Mackintosh is that an attention parameter, α, for a particular conditioned stimulus, can vary as a function of the degree to which the organism comes to learn that the stimulus signals probability of reinforcement. For example, from trial-to-trial observation, pigeons can learn to attend to a red response key that predicts reinforcement, and to ignore a black response key that is uncorrelated with the same reinforcement. To conclude, the psychological meaning of attentional shifts is underspecified, and one has to rely on interpretations of this key-explanatory mechanism to derive qualitative predictions from attentional theory.

1.3.1.2. Rapid attention-shifting in children. The repeated focus on similarities between human and animal associative learning in work on attentional theory indicates that the concept of rapid attention shifts reflects a rather rudimentary low-level ability, not only present in humans but possibly in non-humans too (see e.g., Fagot et al., 1998; cf. Reber’s criteria in Section 1.2.3). According to this *strong interpretation*, there is no *à priori* reason to assume that the mechanism should be operating at different magnitudes in adult humans than in children or lower animals. Nonetheless, other psychological interpretations of rapid attention shifts are possible, too. It could, for example, be argued that there is a gradual increase with developmental age in the ability to rapidly shift attention. In the following, such an interpretation of the concept will be referred to as the *weak interpretation* of rapid attention shifts.

In general psychological terms, ”to attend to something” refers to perceiving in relation to a task or a goal, either internally or externally motivated. Research in cognitive development has confirmed the presence of such an ability in very young children. *Latent inhibition*, defined as slower learning to a previously irrelevant preexposed stimulus than to a novel one (Lubow, 1989), is a typical example of a phenomenon that involves shifting attention away from an irrelevant stimulus to a relevant one. Lubow and Josman (1993) found that hyperactive 59-90 month old children demonstrated a loss of latent inhibition whereas nonhyperactive 58-85 month old children did not. Hence, their results suggest a well-functioning attention-shifting ability in healthy children, but not in children suffering from attentional deficits. Another example of young children's ability to shift attention is the ”sucking into focus” task (Kalnins & Bruner, 1973) in which a motion picture becomes available to the infant contingent upon appropriate changes in sucking rate.

These results suggest that according to a strong interpretation of rapid attentional shifts very young children have the necessary abilities to rapidly shift attention. Therefore, there should be no doubt as to whether children of 8-9 years of age should exhibit the inverse base-rate effect. On the other hand, if the weak interpretation of attentional shifts is correct, such children may not yet have developed the ability to rapidly shift attention, and should therefore not exhibit the inverse base-rate effect.

1.3.1.3. Rapid attention-shifting and learning efficiency. The more general argument for shifting attention away from cues that fail to predict an observed outcome is that it will speed up the learning process. For example, in Kruschke (in press) it is argued: “In general, attentional shifts can be highly adaptive, despite the fact that it can generate side effects such as the inverse base-rate effect. Learners want to make correct choices in as few training trials as possible. This need for speed can be accomodated with attention shifting”. Similarly, in Kruschke and Johansen (1999) it is proposed that: “The shift of attention is both rash and rational: rash because it is rapid, and rational because it quickly reduces error and mitigates interference between previous and novel learning. Thus the rash shift helps to achieve the rational goal of learning quickly” (p. 1084).

These assumptions of attentional theory seem to suggest that the stronger the attentional shifts, the more efficient is learning. Because the rapid attention-shifting mechanism is the key explanatory mechanism of the inverse base-rate effect, this suggests that efficient learners should be more prone to exhibit the inverse base-rate effect than inefficient learners. This prediction will be of interest in Study II and III of this thesis (detailed in Sections 3.3.2 and 3.4.1, respectively).

The two quantitative implementations of attentional theory, ADIT (Kruschke, 1996) and EXIT (Kruschke, in press) provide exceptional fits to human data. Nonetheless, there are some noteworthy discrepancies between the predictions by the models and human data. For example, in Kruschke (Experiment 1, 1996) it was reported that already from the outset participants responded better-than-chance for the rare categories. This result seems to violate the assumption that the participants form associations for the common diseases before the rare ones. This better-than-chance performance was attributed to some “*non-random guessing strategy*” (Kruschke, 1996, p. 6; see also Kruschke & Bradley, 1995, on guessing strategies). Thus, to account for the full complexity of the inconsistent base-rate effects in the Medin and Edelson (1988) design, Kruschke had to postulate three separate mechanisms: (1) a base-rate bias, (2) a rapid attention-shifting mechanism, and (3) a guessing strategy (Kruschke & Bradley, 1995). More recently, a fourth mechanism has been added, an (4) exemplar-based module that learns the trial-to-trial changes in the attentional shifts (Kruschke, 2001, in press).

In the following, a high-level reasoning process similar to the non-random guessing strategy discussed by Kruschke (1996) and Kruschke and Bradley (1995) will be presented that has the potential by itself to account for the entire pattern of data in the inverse base-rate design.

2. The Eliminative Inference Approach

The eliminative inference approach is based on the conception that participants in the standard inverse base-rate design (Medin & Edelson, Experiment 1, 1988) perceive the task as one concerning the learning of a set of *inference rules* (for similar approaches, see e.g., Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994). The task is deterministic throughout the training phase, and the consistent use of only six inference rules allows for 100 percent correct classifications (see the right-most column of Table 4).

To this broad conception, two inferential mechanisms are added: (a) *inductive inference*, and (b) *eliminative inference*, according to which participants sometimes eliminate category options that are inconsistent with well-supported inference rules.

Table 4
The Basic Design of the Training Phase in the Medin and Edelson Experiment 1 (1988). The Right-most Column Presents the Six Inference Rules of the Eliminative Inference Approach That Yield 100 Percent Correct Classifications.

Training Frequency	Symptoms	Disease Category	Inference Rules
3	$I_1.PC_1$	C_1	$I_1.PC_1 \rightarrow C_1$
1	$I_1.PR_1$	R_1	$I_1.PR_1 \rightarrow R_1$
3	$I_2.PC_2$	C_2	$I_2.PC_2 \rightarrow C_2$
1	$I_2.PR_2$	R_2	$I_2.PR_2 \rightarrow R_2$
3	$I_3.PC_3$	C_3	$I_3.PC_3 \rightarrow C_3$
1	$I_3.PR_3$	R_3	$I_3.PR_3 \rightarrow R_3$

Note. For concrete names on symptom- and disease names, see the General Method in Section 3.1 of this thesis. The dot between the symptoms indicates co-occurrence. Copyright © 2001 by the American Psychological Association. Reprinted with permission.

2.1. Induction and Elimination

To illustrate the processes of induction and elimination, consider the following example (see Figure 2). You are told that a friend of yours has bought a new pet animal called George. You are also informed that George either is a goldfish or a Brotogeris tirica. Not being a zoologist, you have a pretty good idea of what a goldfish is, but you have no notion whatsoever of what a Brotogeris tirica is. Your task is to guess what kind of pet animal George is.

First, consider a situation where you receive the cue ”*George lives in water*”. George is thus similar to a goldfish in the sense that he lives in water. In the absence of knowledge about what a Brotogeris tirica is, you might be tempted to guess that George is a goldfish. This illustrates one—admittedly weak—form of inductive inference, and you would probably not be very confident in this guess. Now consider the situation where you instead are given the cue ”*George can fly*”. You would probably guess with high confidence that George is a Brotogeris tirica—he is certainly not a goldfish! (Brotogeris tirica is Latin for the plain parakeet). You use your knowledge about the category ”goldfish” to eliminate the possibility that George is a goldfish. The idea of an eliminative inference mechanism embodies this sort of inference. The bottom line is that the mind need not only use inference rules for induction and ”verification”, as addressed in previous models (e.g., the context model by Medin & Schaffer, 1978), but also for “falsification” of possibilities that lead to absurdity or contradiction.

It is important to note the relationship between category base-rates and elimination. The more common goldfish are (i.e., the higher the base-rate of goldfish), the more information about goldfish you will have, and the more certain you will be that George *is not* a goldfish. Therefore, when people use eliminative inference they will, if anything, *respond inversely to the base-rates*.

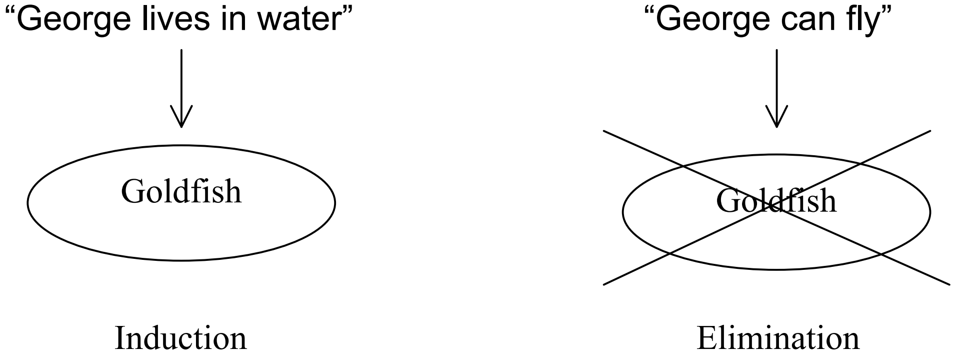


Figure 2. The two inferential mechanisms in the eliminative inference approach: Induction and elimination.

In the following, I will first describe a simple computational model of the eliminative inference idea, ELMO (for ELimination MOdel) that produces “an existence-proof” that this process can give rise to the inverse base-rate effect. Then I will present a number of qualitative predictions that derive from the eliminative inference idea and that contrast with the predictions by attentional theory (Kruschke, 1996, in press).

2.2. Representational Assumptions

In the Medin and Edelson (1988) design, there are six diseases. Assume that on a given transfer trial, each participant has a certain subset of inference rules that are active and available for inference—*the active set*, whereas another subset of potential rules are inactive—*the guessing set*. The active set consists of the categories (response options) for which the inference rules are active in working memory. The guessing set contains the remaining categories for which no inference rules are active, but which remains possible response options. For example, on one particular trial the rules for C_1, C_2,and C_3 may be active and part of the active set, and the diseases R_1, R_2,and R_3 may form the participant’s guessing set (cf. Table 4). The inference rules that are active for a particular individual are expected to vary from moment to moment and over different individuals as described by a probability distribution. The probability that a category enters the active set on a given trial is a positive function of the number of training trials. Given a fixed number of training trials, this implies that the probability is a positive function of category base-rate[6]. The probability that an inference rule for a common disease enters the active set at a given trial is denoted *c*, and the corresponding probability for a rare disease is denoted *r* (with $c>r$). If these probabilities are mutually independent, the probabilities of the different active subsets can be estimated (see the Appendix of Study I).

Throughout the learning phase of the inverse base-rate task, every probe has symptoms that precisely match one of the inference rules (cf. Table 4). In the transfer phase, however, entirely new symptom combinations are presented which do not allow for a perfect match with any of the six inference rules. In this phase, the participants' inferences, or perhaps better guesses, will have to be based on the similarity between the new symptom combinations and the conditions of the inference rules.

2.3. Inferential Processes

According to the eliminative inference approach, inference rules can be applied in two distinct ways. *Direct induction* applies when a new probe has exactly the symptoms present in the condition part of the inference rule, or when the symptoms of the probe are sufficiently similar to the conditions of the inference rule. In this case, inference rules that share features with the probe within the active set are activated (see "Activation" in Figure 3) and entered into an induction procedure (see "Induction" in Figure 3), where they are executed with a probability proportional to their similarity (for detailed computations of the similarity structures, see Study I).

In contrast to direct induction, inference rules can also be used for *eliminative inference*. The process of eliminative inference is activated when the new probe is dissimilar to the conditions of an inference rule (see "Elimination" in Figure 3). Under these conditions the probe is randomly assigned to one of the diseases within the guessing set. Specifically, because the probe does not agree with any of the inference rules in the active set, the participants will have to guess randomly among the diseases in the guessing set.

The gain in inferential power is illustrated by the fact that with inductive inference alone the participant will need to know all six diseases to have 100 percent correct classifications. By adding the eliminative inference mechanism, knowing only five diseases is sufficient for perfect performance (cf. Table 4).

The number of elements in the guessing set is a random variable determined by the probabilities *c* and *r*, which are assumed to be independent. For example, in the Medin and Edelson (1988) design with six diseases there is a probability of $(1\text{-}c)^3(1\text{-}r)^3$ that no inference rules are active, and that the guessing set contains all six diseases. Thus, the probability of a response in one particular category is 1/6. If the active set contains all inference rules for the common diseases and the guessing set contains all the rare diseases, the probability of a guessing response in one particular rare category is 1/3 (and so on for other possible guessing sets).

To obtain the predicted response proportions, the expected values across all guessing sets have to be computed, as determined by the probabilities *c* and *r*. Com-

[6] In retrospect, this assumption may be questioned since the base-rate that occurs early in training has proved to be more important than the base-rate that occurs late in training for the inverse base-rate effect to emerge (e.g., Kruschke, 2001b; Medin & Bettger, 1991; see also section 3.4.4 of this thesis).

putation of these expected values is straightforward but tedious (for detailed computations, see Tables A1-A3 of the Appendix in Study I). Here it is sufficient to note that, by virtue of the base-rate manipulation, the guessing set is more likely to contain rare diseases rather than common ones.

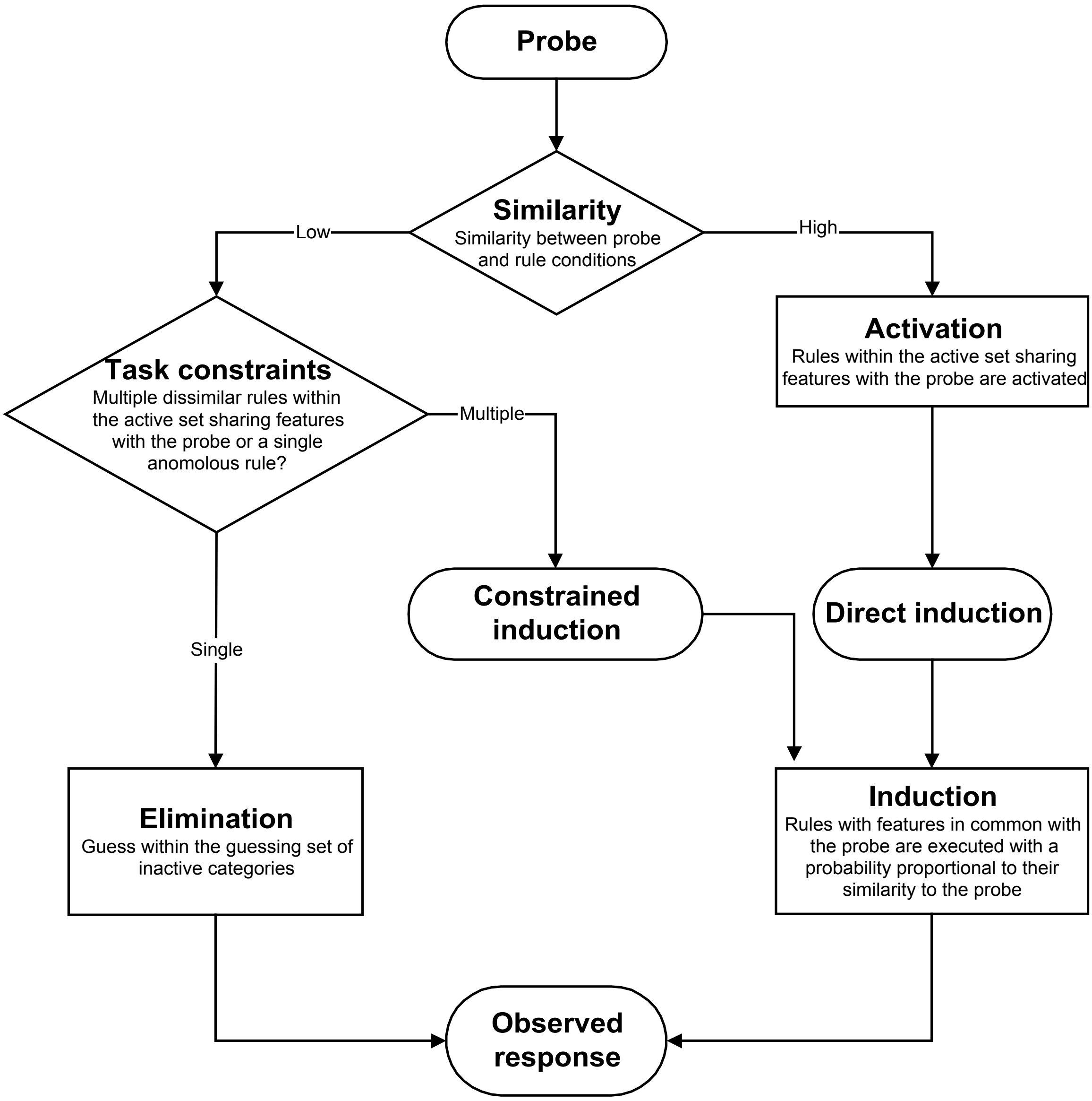

Probe
Similarity
Similarity between probe and rule conditions
Low
High
Task constraints
Multiple dissimilar rules within the active set sharing features with the probe or a single anomolous rule?
Multiple
Single
Activation
Rules within the active set sharing features with the probe are activated
Constrained induction
Direct induction
Elimination
Guess within the guessing set of inactive categories
Induction
Rules with features in common with the probe are executed with a probability proportional to their similarity to the probe
Observed response

Figure 3. A schematic presentation of the main processing stages of the eliminative inference approach. Copyright © 2001 by the American Psychological Association. Reprinted with permission.

It is assumed that a probe with all features in common with a rule condition, or at most one deviating feature (i.e., either one missing feature or one feature too much), elicits induction, whereas a probe with two or more deviating features elicits elimination. One crucial implication is that the conflicting probe, *PC.PR*, which deviates from all inference rules with two or more features, will elicit elimination. More specifically, with regard to the two rules with the most similar rule conditions, $I.PC \rightarrow C$ and $I.PR \rightarrow R$, the conflicting probe has both one lacking feature, *I*, and one contradicting feature, *PC* or *PR*. The imperfect probe, *I*, and the combined probe, *I.PC.PR*, on the other hand, only deviate with one feature from two of the rules, $I.PC \rightarrow C$ and $I.PR \rightarrow R$. The imperfect probe lacks one perfect predictor and the combined probe contains one perfect predictor too much. Therefore, these probes elicit direct induction. This crucial assumption, that *PC.PR* elicits more elimination than the other two critical probes, *I*, and, *I.PC.PR*, is tested in Experiment 3.[7]

At the most general level, the six diseases available by the task design provide a constraint on the admissible response categories. More specifically, it is assumed that at the end of training the participants will have detected that each symptom is associated with at most two diseases. Therefore, when they find that more than one inference rule within the active set has features in common with the transfer probe, they will resort to induction even if the probe is dissimilar to the conditions of the inference rule. That is, convinced that each symptom commonly is associated with at most two diseases, when the participant finds that the transfer probe shares a feature with two inference rules within the active set, he or she will conclude that one of these two rules must be the correct choice. Induction imposed by knowledge of the task structure is referred to as *constrained induction*, in contrast to the direct induction procedure discussed above (see "Constrained Induction" in Figure 3).

In summary, the eliminative inference approach postulates mental representations in the form of inference rules and explains the inconsistent pattern of base-rate effects in the inverse base-rate design with two cognitive mechanisms: Induction and elimination, as summarized in Figure 3. At the end of the training phase, the participants operate in a direct-induction mode in a routine-like manner with well-established inference rules. In the transfer phase, however, they are presented with novel and ambiguous symptom patterns. Therefore, the participants will have to enter a mode of controlled and flexible problem-solving that generates a more complex pattern of inductive- and eliminative inferences for these probes.

2.4. Elimination and the Inverse Base-Rate Effect

To understand why the eliminative inference approach predicts an inverse base-rate effect, it is useful to focus on one common-rare disease pair, C_1 and R_1, referred to as the *focal disease pair.* (The computations provided in the Appendix of Study I are also organized around the heuristic concept of a focal disease pair.) In regard to the focal disease pair, at a given trial of the transfer phase a participant may be in one of four mutually exclusive knowledge states: *State 1:* Neither the inference rule for the common nor the rare disease are elements of the active set. *State 2:* Only the inference rule for the common disease is an element of the active set. *State 3:* Only the inference rule for the rare disease is an element of the active set. *State 4:* Both the inference rules for the common and the rare diseases are elements of the active set.

Panel A of Figure 4 illustrates the processing of the imperfect (i.e., I_1) and combined probes (i.e., $I_1.PC_1.PR_1$). In State 1, in which no rules are active, there is no pos-

[7] This key assumption in the present quantitative implementation of the eliminative inference approach (ELMO) is very simple. In a future version of the model it would probably be meaningful to assume that one deviating feature elicits *more induction* and *less elimination*, whereas two or more deviating features elicit *more elimination* and *less induction*. However, these ideas are still under development (see also sections 3.4.4. and 4.2.4. of this thesis).

sibility for induction, and the participant decides randomly among the diseases within the guessing set. In State 2, only the inference rule for the focal common disease is active. The probes are sufficiently similar to the rule conditions of the common disease and the eliminative inference approach therefore predicts that the participant executes direct induction and assigns the transfer probe to that (common) disease. In State 3, only the inference rule for the focal rare disease is active. The imperfect and combined probes evoke direct induction favoring the focal rare disease. In State 4, finally, where both inference rules are active, there is direct induction for both transfer probes favoring either of the two active diseases.

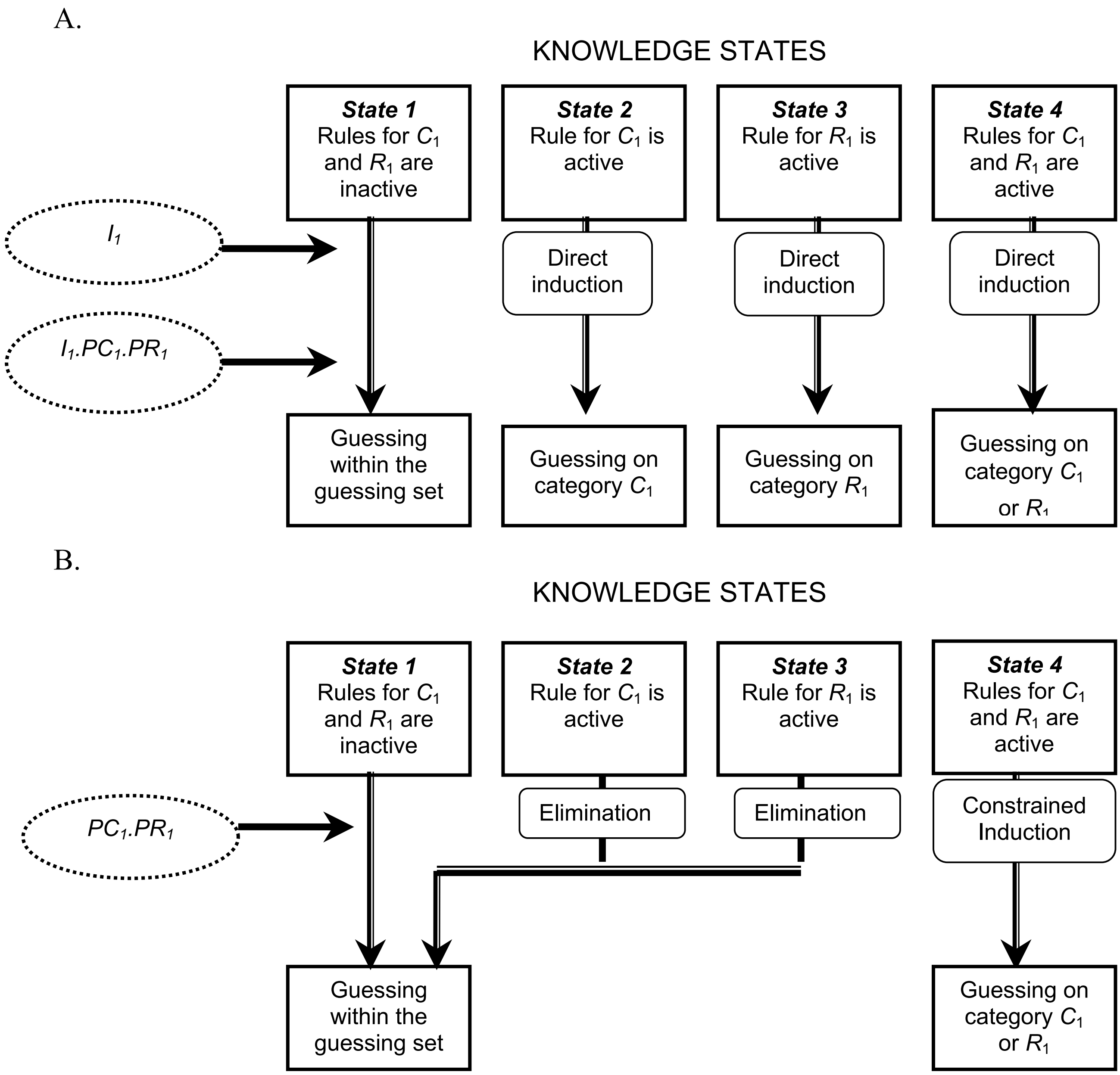

A.
KNOWLEDGE STATES
State 1 Rules for C_1 and R_1 are inactive
State 2 Rule for C_1 is active
State 3 Rule for R_1 is active
State 4 Rules for C_1 and R_1 are active
I_1
$I_1.PC_1.PR_1$
Direct induction
Direct induction
Direct induction
Guessing within the guessing set
Guessing on category C_1
Guessing on category R_1
Guessing on category C_1 or R_1
B.
KNOWLEDGE STATES
State 1 Rules for C_1 and R_1 are inactive
State 2 Rule for C_1 is active
State 3 Rule for R_1 is active
State 4 Rules for C_1 and R_1 are active
$PC_1.PR_1$
Elimination
Elimination
Constrained Induction
Guessing within the guessing set
Guessing on category C_1 or R_1

Figure 4. The four knowledge states of the eliminative inference approach applied to the imperfect (*I*) and combined probes (*I.PC.PR*) (Panel A), and the conflicting probe (*PC.PR*) (Panel B). Copyright © 2001 by the American Psychological Association. Reprinted with permission.

In Panel B of Figure 4 the processing of the conflicting transfer probe (e.g., $PC_1.PR_1$) is illustrated. In State 1, in which no rules are active, there is no possibility for induction, and the participant decides randomly among the diseases within the guessing set. In State 2, only the rule for the focal common disease is active, but the transfer probe is too dissimilar to the rule conditions of that disease and therefore the participant will eliminate the focal common disease and decide randomly among the diseases in the guessing set. In State 3, only the inference rule for the focal rare disease is active. Again the conflicting probe is too dissimilar to the rule conditions of that disease and the participant will therefore eliminate that disease with a random guess among the diseases within the guessing set. Thus, in both States 2 and 3, dissimilar probes elicit elimination: In State 2 the common disease is eliminated, in State 3 the rare disease is eliminated. However, because of the base-rate difference (with $c>r$), overall State 2 will be the more frequent knowledge-state and therefore a common disease will be eliminated more often. The guessing set will predominantly contain rare diseases. This situation will contribute to an increased rate of rare disease responses that conform to the inverse base-rate effect. In State 4, finally, where both inference rules are active, there is constrained induction for probes that share features with multiple dissimilar inference rules.

To obtain the overall predicted response proportions, one has to multiply the probabilities of the knowledge states with the expected response proportion for each disease, where the response proportions depend on whether induction or elimination is elicited in the particular knowledge state. These rather cumbersome computations are summarized for each of the transfer probes in Table A1 of the Appendix of Study I.

In order to illustrate the predictions, a quantitative implementation of the eliminative inference approach, ELMO, was fitted to data from the transfer design of Experiment 1 in Kruschke (1996). In this experiment data for a large number of transfer probes are presented and therefore they provide a suitable first gambit for the model. A partial replication of Experiment 1 by Medin and Edelson (1988) was conducted in Kruschke, although his design was delimited to four diseases. Each participant was trained for 120 trials.

The left-side panels of Figure 5 illustrate the quantitative predictions by ELMO (see also Kruschke, in press). As is evident, ELMO is very successful in reproducing the puzzling pattern of base-rate effects observed in the data. The residual deviations between predictions and data nevertheless suggest some cue-competition between the perfect predictors for common and rare diseases, perhaps arising from the kind of rapid attention shifts modeled by attentional theory. For example, ELMO with no cue-competition predicts the same base-rate use for the probes *I* and *I.PC.PR*. The data indicate less base-rate use for the combined probe, *I.PC.PR*, presumably because *PR* pulls stronger toward category *R* than *PC* pulls toward category *C*. Similar effects can account for the deviations for *I.PCo* and *I.PRo* (see also Sections 3.4.4 and 4.2.4 of this thesis).

The argument is not that the particular implementation of the eliminative inference approach, ELMO, provided here is viable as a general model of categorization. Nor is it denied that there are cue-competition effects in the data. However, the results in Figure 5 demonstrate that the high-level reasoning processes modeled by ELMO are sufficient by themselves to produce the basic pattern of base-rate effects in the inverse base-rate design (see also Kruschke, in press). In Study I, ELMO is also fitted to the data sets presented in that study. However, the main objective of this thesis is not any particular implementations, but novel qualitative predictions derived from the eliminative inference mechanism that contrast with those by attentional theory. To understand these theory-critical predictions, it is necessary to describe the psychological meaning of the eliminative inference mechanism in more detail.

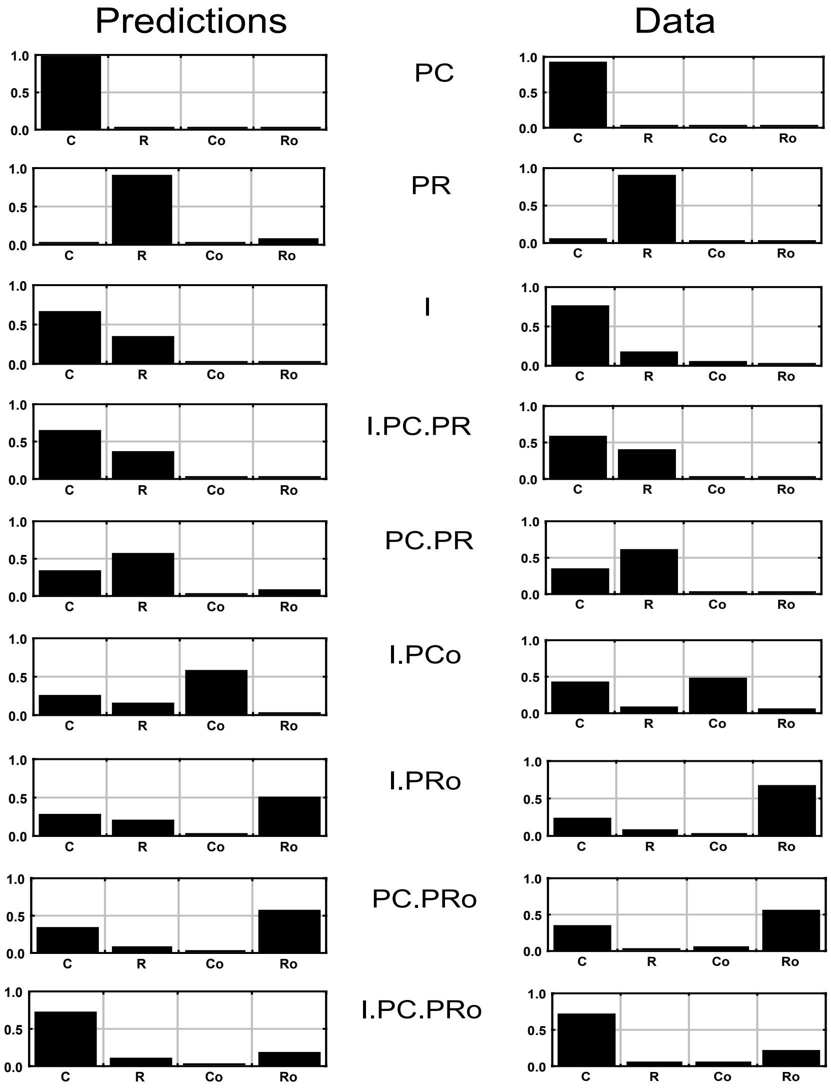


Figure 5. The predicted (left-side panels) and observed (right-side panels) mean response proportions for data in Kruschke (Experiment 1, 1996). The ordinate presents the response proportions and the abscissa presents the response categories. Also note that, for example, *PCo* in the probe *I.PCo* refers to the perfect common transfer probe of the category-pair that is not associated with the presented *I*, that is *o* = other. Copyright © 2001 by the American Psychological Association. Reprinted with permission.

2.4.1. Eliminative inference and developmental age. According to the eliminative inference approach, participants activate simple inference rules during the learning phase of the inverse base-rate task, for example *I.PC*→*C* (cf. the right-most column of

Table 4). These rules are then used in a flexible and controlled manner to both induce and eliminate category membership in the transfer phase. It has been shown that young children can activate simple inference rules if the task is perceptual rather than conceptual in nature. For example, research on simple rule induction has shown that even 30 month old children are successful in matching a toy frog with a box with a picture of a toy frog on top of it (i.e., the "sameness rule" see Smith, Deloache, & Schreiber, 1995). It is questionable, however, whether such tasks evoke inference rules of the more complex conceptual kind proposed by the eliminative inference approach. Inducing the membership of a concrete stimulus such as *PC* by comparing its similarity with the conditions of an inference rule such as *I.PC* appears to be less complicated than ruling out its applicability and looking beyond the concrete information given, such as is proposed for *PC.PR*. In this case, an additional mental operation is required since the reasoner must abstract and access other hypothetical alternatives (i.e., the categories in the guessing set). Thus, the child needs to understand that the concrete features in front of him/her are inconsistent with ("rule out") a category, thereby increasing the likelihood that one of the other categories is the correct one.

In a recent review on cognitive development beyond childhood, Moshman (1998) maintains that: "Although young children routinely make inferences in accord with the rules of logic, only later in development do individuals increasingly think about such rules and understand their epistemic role in justifying connections among propositions (p. 957). Moshman (1998) continues: "There is surprisingly strong support for Piaget's 1924 proposal that formal or hypothetico-deductive reasoning—deliberate deduction from propositions consciously recognized as hypothetical—plays an important role in the thinking of adolescents and adults but is rarely seen much before the age of 11 or 12" (1998, p. 972).

Thus, recent research on cognitive development seems to suggest that eight and nine year-old children (the age-group chosen in Experiment 4 in Section 3.3.1 below) are not yet able to reason formally, and therefore they cannot be expected to invoke the inference rule of elimination as frequently as adults that gives rise to the inverse base-rate effect according to the eliminative inference approach.

2.4.2. Rule-based reasoning and learning efficiency. As mentioned previously, recent research has demonstrated individual differences in rule-based and similarity-based generalization as a function of learning performance (Shanks & Darby, 1998). More specifically, rule-based responding was evidenced in efficient learners, whereas similarity-based responding was evidenced in inefficient learners. Interestingly, the discrimination task developed by Shanks and Darby (1998) has several similarities with the task developed by Medin and Edelson (1988). In both designs, the task is to envisage a potential medical state on the basis of various features. In addition to the conceptual relatedness of the designs, both tasks consist of similar numbers of symptoms, similar modes of stimuli presentation, similar participant populations, and a deterministic training stage. Furthermore, in both tasks, the theories about the underlying learning processes are evaluated by requiring participants to make decisions on transfer probes consisting of novel constellations of previously encountered stimuli. Because of the many similar features of the tasks, it seems reasonable to expect individual differences as a function of learning performance in participants responding in the inverse base-rate task, too. In the next section, this prediction as well as a number of other predictions implied by the idea of an eliminative inference mechanism will be presented in detail. Importantly, all of these predictions contrast with those implied by attentional theory (Kruschke, 1996, in press).

2.5. Theory-Critical Predictions

2.5.1. Beginner's luck. The first theory-critical prediction is *beginner's luck*, which implies that in the learning phase participants will perform better than chance already at the outset for the rare category exemplars (i.e., *R* given *I.PR*). The explanation is that the common categories, *C*, are learned before the rare ones, *R*, and the common categories are therefore eliminated on every first-time encounter with a rare category exemplar. This process of elimination leads to a guess among the diseases in the guessing set, and because of the base-rate difference, these are the rare ones. This process also speeds up the learning process (cf. Section 2.3).

This prediction has not been systematically explored in previous research, but it provides an explanation for the anomalous finding referred to as a "non-random guessing strategy" reported in Kruschke (1996) and in Kruschke and Bradley (1995). The predictions derived from attentional theory were premised on the assumption that predictors of common diseases, *PC*, become associated with common diseases, *C*, before rare predictors, *PR*, become associated with rare diseases, *R* (cf. Figure 1). In Experiment 1 in Kruschke (1996), two common-rare disease pairs are involved. With random decisions among the four categories, 25% correct classifications are expected. Surprisingly, however, it was found that participants selected the correct rare disease, *R*, in 49% of the initial trials. This phenomenon, which violates one of the key-assumptions of attentional theory, is explained and directly predicted by the eliminative inference mechanism. In fact, a simulation of ADIT in Study III below shows that in contrast to the rationale of attentional theory rapid attention shifts *decelerate* learning (see Section 3.4.1).

2.5.2. A novel symptom effect. The second theory-critical prediction is the *novel symptom effect* which implies that presentation of an entirely novel symptom in the transfer phase will lead to a preponderance of rare category responses, *R*, mirroring the inverse base-rate effect for conflicting probes, *PC.PR* (Juslin, Wennerholm, & Winman, 1999). The participants will notice that the novel symptom is dissimilar to the symptoms of the active inference rules, and therefore guess on some disease within the guessing set. Because of the base-rate manipulation, the elements of the guessing set are likely to be rare categories, *R*, rather than common ones, *C*.

This prediction is important for two reasons: First, it is a direct test of the existence of an eliminative inference mechanism. To the extent that this type of inferences underlie the responses for both the conflicting and novel transfer probes, data for these probes should be similar. Second, this prediction is critical in the sense that it points to a response pattern contrary to the one by attentional theory (Kruschke, 1996, in press). The novel probe has not been affected by any shifts of attention during learning, so the only mechanism at work is the base-rate bias pulling toward common categories, *C*.

Therefore, attentional theory predicts common responses, *C*, whereas the eliminative inference approach predicts rare responses, *R*.

2.5.3. A repetition effect. The third theory-critical prediction is the *repetition effect*, which implies that if the eliminative inference mechanism is an important factor contributing to both the inverse base-rate effect and the novel symptom effect, a repetition of the transfer phase should produce a smaller inverse base-rate effect and a smaller novel symptom effect. This prediction follows from the eliminative inference mechanism in conjunction with the following auxiliary assumption: When encountering a novel and puzzling symptom pattern for the first time, it seems sensible to eliminate the categories associated with well-established inference rules (i.e., the elements of the active set), and to guess on some of the diseases in the guessing set. However, when presented with an additional learning phase, the chance increases that the participant will notice that the ambiguous probes actually do not occur in learning at all. Therefore, when the second transfer phase takes place, participants are more likely to note that the ambiguous probes have no straightforward relation to the categories presented in learning. In this case, they may resort to a true random decision between all categories, or even respond according to the base-rate.

In contrast, exposure to the conflicting probe, *PC.PR*, in a previous transfer phase with no feedback should have no pervasive effect on the associations formed between predictors and diseases that underlie the inverse base-rate effect according to attentional theory. In sum, the eliminative inference approach predicts that the inverse base-rate effect should decrease after the second learning phase, whereas attentional theory predicts an unchanged inverse base-rate effect.

2.5.4. A novel disease effect. The fourth theory-critical prediction is the *novel disease effect*. According to the eliminative inference approach, introduction of a novel disease in the transfer phase will lead to a qualitative change in the response pattern (in the experimental setting the novel category is simply labeled “other disease”). A novel disease is necessarily an element of the guessing set, and will therefore accumulate many responses from eliminative inference. The conflicting probe, *PC.PR*, is especially prone to elicit elimination, and will therefore produce particularly many “other disease” responses. For the conflicting probe the modal response should no longer be the rare category, *R*, but the novel category “other disease”. Note, however, that the probability that a rare category, *R*, is an element of the guessing set is still higher than the probability of a common category, *C*. As a result, there will be more rare than common responses for the conflicting probe, even if the modal response is the “other disease”. The remaining transfer probes will primarily evoke induction, and the modal response will continue to be the common category, *C*. In one sense this means that the inverse base-rate effect is predicted to disappear: For none of the probes, the rare categories will constitute the modal response.

According to attentional theory, the inverse base-rate effect arises from rapid attention shifts away from the common disease, *C*, toward the rare disease, *R*, in the learning phase. These processes are not affected by the addition of a novel disease in the transfer phase. A strict interpretation of attentional theory therefore predicts that there will be no responses assigned to the novel disease, because all the predictors will be associated with the diseases presented in the learning phase (i.e., *C* and *R*). A less strict interpretation might allow for some noise in the learning- and/or response process that assigns some portion of responses also to the novel disease. However, to the extent that the associations formed in training directly determine the transfer responses, the rare categories, *R*, should constitute the modal response for the conflicting probes, *PC.PR*, whereas the common categories, *C*, should be the modal response for the remaining critical transfer probes (i.e., *I* and *I.PC.PR*).

2.5.5. No inverse base-rate effect in children. The fifth theory-critical prediction concerns *no inverse base-rate effect in children*. As previously argued, if the strong interpretation of attentional theory is correct, that the inverse base-rate effect is the outcome of a rudimentary low-level encoding mechanism, the phenomenon can be expected to be present early in human development. In contrast, if the inverse base-rate effect is a result of eliminative inferences, a meta-cognitive ability that is rarely seen before the age of 11 or 12 (Moshman, 1998), we can expect that children should fail to exhibit the phenomenon. The finding of an inverse base-rate effect in children would provide support for the strong interpretation of rapid attention shifts, but leave the eliminative inference approach and the weak interpretation of rapid attention shifts in doubt.

2.5.6. A stronger inverse base-rate effect in rule-based learners. The sixth theory-critical prediction is concerned with the representations that underlie the inverse base-rate effect. It has recently been demonstrated that efficient learners are more likely to rely on rule-based principles, whereas inefficient learners rely on similarity-based principles (Shanks & Darby, 1998). If the inverse base-rate effect depends on rule-based processes, such as those described by the eliminative inference approach, we can expect *a stronger inverse base-rate effect with rule-based learners* than with similarity-based learners. In contrast, according to attentional theory, the inverse base-rate effect depends on cue-category associations affected by rapid attention shifts. On this account, there is no reason to expect a reversal of responding as a function of learning mode because the theory does not incorporate any rule-based architecture. In addition, in contrast to the critical assumption that rapid attention shifts accelerate learning, a simulation of ADIT actually demonstrates that rapid attention shifts decelerate learning. Thus, if anything, according to attentional theory a stronger inverse base-rate effect can be expected in inefficient (similarity-based) rather than efficient (rule-based) learners. Because rule-based learning is highly correlated with learning efficiency (see Reanalysis of Experiment 1 from Study I below), however, learning efficiency is statistically controlled for in Experiment 5 below (see Section 3.3.3 of this thesis).

2.5.7. Whenever there is an inverse base-rate effect, *PC* should elicit more category *C* responses than *PR* should elicit category *R* responses. The seventh theory-critical prediction concerns the primary cause of the inverse base-rate effect; rapid attention shifts or eliminative inference. According to attentional theory, people's knowledge is assymetrical: "Symptom *I* more strongly indicates *C* than disease *R*, and symptom *PR* more strongly indicates disease *R* than symptom *PC* indicates *C*" (Kruschke, in press) (cf. Figure 1). Indirect support for this encoding asymmetry comes from participants' choice of the rare disease, *R*, when presented with the conflicting transfer probe, *PC.PR* (and also to some extent, by the larger proportion of *C* responses for *I* and *I.PC.PR*). Because attentional theory predicts stronger weights from *PR* to *R* than from *PC* to *C*, comparing the response proportions for these two transfer probes should provide a stronger and more direct test of the presence of rapid attention shifts. Thus, whenever there is an inverse base-rate effect, *PR* should elicit more category *R* responses than *PC* should elicit category *C* responses (see Equation 2 below). In contrast, induction and elimination of simple inference rules suggests that whenever there is an inverse base-rate effect, *PC* should elicit more *C* category responses than *PR* should elicit category *R* responses. After training the participants may have formed two inference rules that involve *PC* and *PR*: $I.PC \rightarrow C$ and $I.PR \rightarrow R$. Because of the base-rate manipulation, the probability that $I.PC \rightarrow C$ is accessible and retrieved on a specific transfer trial is higher than the probability that $I.PR \rightarrow R$ is accessed and retrieved (see Equation 3 below). Thus, the predictions by the eliminative inference approach and attentional theory pull in different directions:

$$p(C|PC) < p(R|PR) \text{ AND } p(C|PC.PR) < p(R|PC.PR) \qquad \text{(Equation 2)}$$

$$p(C|PC) > p(R|PR) \text{ AND } p(C|PC.PR) < p(R|PC.PR) \qquad \text{(Equation 3)}$$

Next, the results from the empirical studies, aimed at testing these predictions, will be presented. Study I tests the predictions "beginner's luck", "a novel symptom effect", "a repetition effect", and "a novel disease effect". Study II tests the predictions "no inverse base-rate effect in children", and "a stronger inverse base-rate effect in rule-based learners". Study III, finally, tests the prediction "whenever there is an inverse base-rate effect, *PC* should elicit more category *C* responses than *PR* should elicit category *R* responses".

3. Empirical Studies

3.1. The General Method

The general method used in the experiments reported in Study I, II, and III was developed by Medin and Edelson (Experiment 1, 1988) and indexes participants' use of base-rate information from experience with classifying examples of a category; the inverse base-rate task (or the *Medin and Edelson task*). In each of the experiments in each study the participants were given a brief instruction asking them to imagine that they were physicians at a hospital, and that they were to learn how to make medical diagnoses from experience with hypothetical patients exhibiting various symptoms (but see Experiment 4, detailed in Section 3.3.1). They were informed that the names of the diseases were fictitious, and that no real-life experience could be used to infer the correct responses. Participants were told that they would first be allowed to practice on a number of patients with feedback informing them if they had made the correct or incorrect diagnosis. Then they were instructed to apply the knowledge they had acquired during the learning phase on a number of novel patients in a transfer phase without feedback.

Each participant made the responses individually on a computer. The stimuli consisted of symptoms randomly drawn from a list of twelve symptoms: *Bad knees, anemia, stuffy nose, epidermophytosis, impaired hearing, impaired short-term memory, stomach pain, hair loss, loosening of the teeth, visual defect, back pain,* and *swollen arms*. Participants were asked to select their responses from six fictitious diseases: *Coralgia, Buragamo, Terrigitis, Midosis, Althrax,* and *Namitis*. Symptoms and diseases were assigned randomly to each participant, as was the ordering of the diseases in the list for each trial.

In the learning phase every particular instance of a common disease, *C*, occurred in the presence of two symptoms: One imperfect predictor, *I*, and one perfect, *PC*. Similarly, every particular instance of a rare disease, *R*, was paired with two symptoms: One imperfect, *I*, and one perfect, *PR*. Thus, each *imperfect predictor* was associated with both a common and a rare disease, and each *perfect predictor* was uniquely associated with only one disease (see the learning phase in Table 1).

In the transfer phase, participants were required to diagnose novel patients exhibiting one up to three symptoms. On these trials the imperfect predictors, *I*, and the perfect predictors—three *PC* and three *PR*—were tested separately to see whether participants would choose the common or rare diseases. Other trials tested participants' base-rate knowledge when presented with the combined test, *I.PC.PR*. However, the most important trials in the transfer phase were the conflicting transfer trials, divided into conflicting trials combining the perfect predictors that belonged to the same category-pair against one another (e.g., $PC_1.PR_1$), and conflicting trials combining perfect predictors from different category-pairs against one another (e.g., $PC_1.PR_2$). On the remaining trials, the imperfect predictors were paired with perfect predictors. These were mainly used to disguise the underlying purpose of the transfer phase (see the transfer phase in Table 1).

3.2. Study I (Juslin, Wennerholm, & Winman, 2001)

The purpose of Study 1 was twofold: To present a novel explanation of the inverse base-rate effect and to test the psychological plausibility of this account. The first part presented the eliminative inference approach and a quantitative implementation of this idea, ELMO, which was fitted to data reported by Kruschke (Experiment 1, 1996) and to the data sets reported in Study I, to demonstrate that ELMO is able to capture the basic pattern of base-rate effects observed in the Medin and Edelson design (see Figure 5). The second part reported three experiments that contrasted four predictions derived from the eliminative inference mechanism with the predictions by attentional theory. Experiment 1 addressed *beginner's luck*, *the novel symptom effect*, and *the repetition effect*. Experiment 2 investigated whether the inverse base-rate effect in Experiment 1 was caused by *the repetition effect* or *prolonged learning*. Experiment 3 addressed the *novel disease effect*.

3.2.1. Experiment 1: Beginner's luck, a novel symptom effect, and a repetition effect. Participants were presented with the standard Medin and Edelson task, but different from the original study participants were presented with a second learning- and transfer phase to test the repetition effect. The base-rate ratio was varied with 55 participants in a 3:1 ratio group, and 54 participants in a 7:1 ratio group. This manipulation was included to be certain of replicating the inverse base-rate effect (cf. Shanks, 1992). Only participants who reached asymptotic learning were used in the subsequent analysis. A learning criterion was set at 96% (23/24) correct responses in the last block of 24 trials in Training Session 1. Participants who failed to meet this criterion were excluded. In the 3:1 group, 40 participants out of 55 (73%) met the criterion. In the 7:1 group, 41 out of 54 participants (76%) met the criterion.

Similar to the pattern observed in Kruschke (Experiment 1, 1996), the common diseases were learned more rapidly than the rare diseases and the participants attained a higher proportion correct for the rare diseases than expected by chance alone already at the outset. In the 3:1 group, the proportion "rare" was .45 ± .08; which is 2.6 times higher than expected by chance (.17 in the present design)[8]. In the 7:1 group, the proportion "rare" was .41 ± .12; which is 2.5 times higher than expected by chance. Thus, in both base-rate conditions, the participants performed better than chance already at the outset for rare-category exemplars; beginner's luck.

In the 3:1 group, the results for the novel transfer probes after Training Session 1 mirrored the responses for the conflicting probes, *PC.PR*, after Training Session 1. Although slightly less pronounced, the participants favored the rare diseases, .55 ± .08 (proportion "common" = .45 ± .08). After Training Session 2, the participants had altered into a slight but non-significant use of the base-rate (proportion "common" = .58 ± .09, and proportion "rare" = .42 ± .09). In the 7:1 group there was an inverse base-rate effect for the conflicting probe after Training Session 1 (proportion "common" = .38 ± .08, and proportion "rare" = .55 ± .08), and Training Session 2 (proportion "common" = .40 ± .10, and proportion "rare" = .57 ± .10). The corresponding proportions of common-category and rare-category responses for the novel probe were .42 ± .10 and .58 ± .10, respectively, after Training Session 1, and .52 ± .11 and .48 ± .11, respectively, after Training Session 2. Figure 6 presents a comparison between the responses for the novel and conflicting transfer probes. Although the results are not identical, the participants in both base-rate groups 3:1 (panel A) and 7:1 (panel B) clearly favor the rare diseases on both the conflicting and novel transfer probes; the novel symptom effect.

Although not statistically significant, the inverse base-rate effect was smaller for the 3:1 group than the 7:1 group. This base-rate paradox, also reported by Shanks (1992), implies that the less frequent a rare disease is, the more participants will tend to choose it when presented with a conflicting transfer probe.

To test the prediction that the inverse base-rate effect for conflicting and novel transfer probes should become smaller after Training Session 2, the data for these probes from both conditions were entered into a 2 (Transfer Session 1 vs. 2, within-subjects) x 2 (novel or conflicting probe, within-subjects) x 2 (base-rate ratio 3:1 or 7:1, between-subjects) split-plot analysis of variance. The only significant effect was the main effect of learning, $F\ (1_{,}\ 79) = 9.07$, MSE=.17, p<.005; all other ps were larger than .1. Overall, the mean proportion of rare-category responses for the novel and conflicting probes was .42 after Training Session 1, and .50 after Training Session 2; the repetition effect.

[8] Throughout this thesis, all intervals denoted " ± " refer to 95% confidence intervals based on independent observations. For example, n = 40 in the 3:1 condition of Experiment 1 with 40 participants.

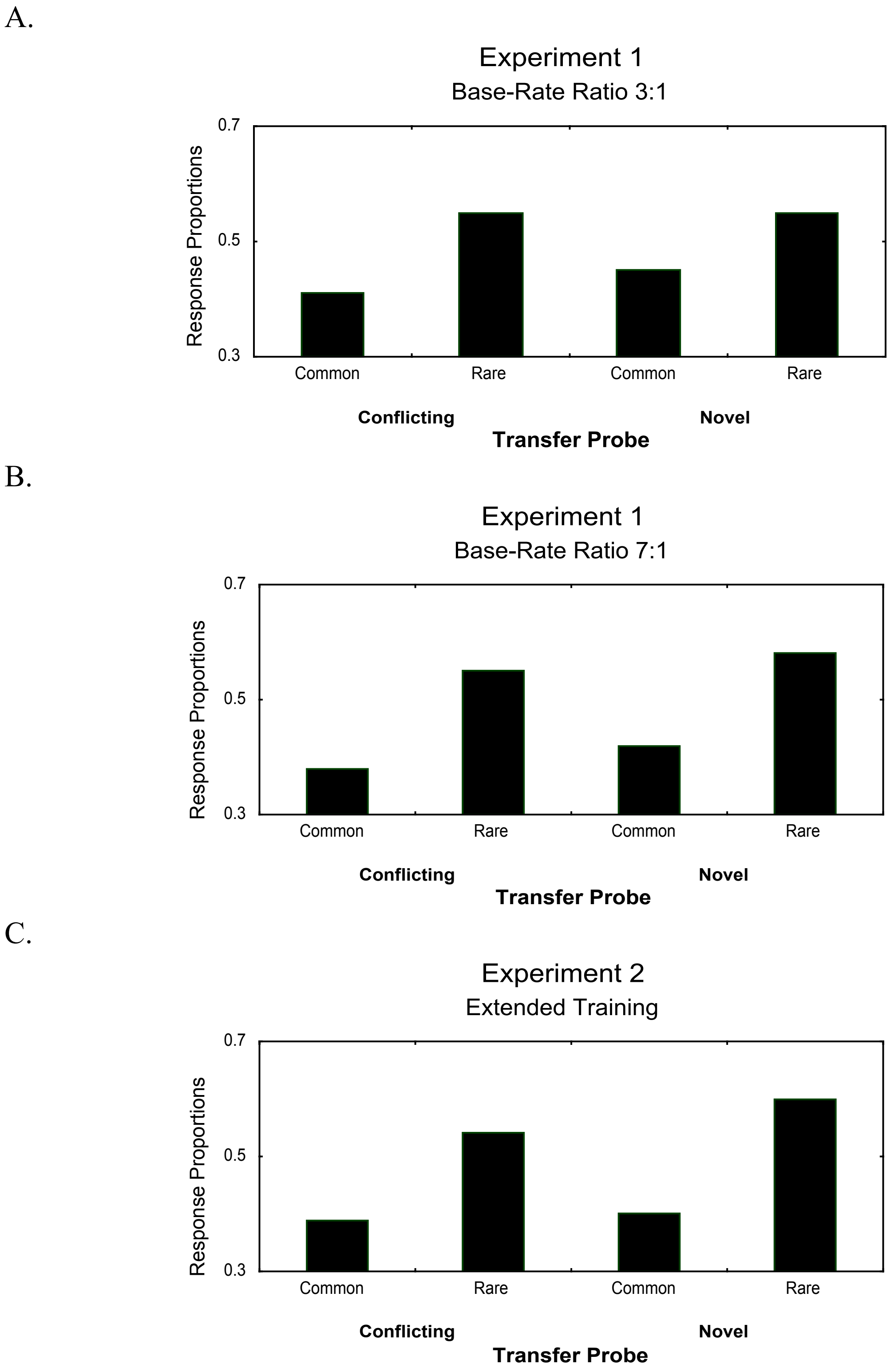


Figure 6. Mean response proportions for the conflicting and novel transfer probes for the 3:1 base-rate condition of Experiment 1 after Training Session 1 (Panel A), the 7:1 base-rate condition of Experiment 1 after Training Session 1 (Panel B), and for Experiment 2 with extended training (Panel C). Copyright © 2001 by the American Psychological Association. Reprinted with permission.

In conclusion, Experiment 1 was successful in replicating the inconsistent base-rate effects observed by Medin and Edelson (1988) with base-rate use for the imperfect and combined probes, but an inverse base-rate effect for the conflicting probe (for more details, see Table 2 in Study I). In addition, the results confirmed three of the predictions derived from the eliminative inference approach: beginner's luck, the novel symptom effect, and the repetition effect. None of these empirically verified predictions are predicted by attentional theory.

3.2.2. Experiment 2: A repetition effect or prolonged learning? In Experiment 1 the inverse base-rate effect was smaller after Training Session 2, and for the novel transfer probe it vanished altogether. The interpretation provided in Experiment 1 was that this finding is caused by the testing between Training Sessions 1 and 2. When the novel probe has been presented in the transfer phase after Training Session 1, it will not be ”novel” when the second transfer phase takes place; the repetition effect. A second potential explanation is that with extended training the inverse base-rate effect will disappear. For example, Medin and Bettger (1991) discussed the possibility that “with enough experience the inverse base-rate effect that we have attributed to competitive learning may be overcome altogether“ (p. 328).

Experiment 2 addressed whether the diminished inverse base-rate effect in Experiment 1 was due to repeated transfer exposures or to prolonged learning. The training phase involved 672 (4*168) training trials without transfer phases interspersed midways through the training trials. It was hypothesized that if the inverse base-rate effect would be eliminated after this amount of training prolonged learning would probably be the correct explanation of the inverse base-rate effect. In contrast, if the inverse base-rate effect would persist despite this prolonged training, it would suggest that the diminished inverse base-rate effect after Training Session 2 in Experiment 1 is a consequence of repeated transfer exposures; the repetition effect.

Twenty-five undergraduate students participated in Experiment 2. The base-rate ratio was 3:1 for all participants. At the end of training they were presented with the transfer phase, which was identical to the one in Experiment 1 and consisted of 24 trials. Because the content of the design was identical to the 3:1 condition of Experiment 1, and the participant population was the same, the two experiments were compared. Twenty-three out of 25 participants (92%) met the learning criterion.

Again, the data replicated the standard pattern in the Medin and Edelson design with base-rate use for the imperfect and combined probes, and an inverse base-rate effect for the conflicting and novel probes. Thus, these data demonstrated that the inverse base-rate effect persists even after ample learning both for the novel and conflicting probes (see panel C of Figure 6), which suggests that the diminished inverse base-rate effect after Training Session 2 in Experiment 1 was a consequence of repeated transfer exposures rather than prolonged learning; the repetition effect.

3.2.3. Experiment 3: A novel disease effect. Experiment 3 addressed the novel disease effect. During the transfer phase, the list of diseases was extended with a seventh category, labeled “*Other disease*”. Similar to the procedure in Experiment 2, the results from Experiment 3 were compared to the response proportions of Experiment 1, but this time to the 7:1 condition after Training Session 1 and primarily for conflicting probes. Experiment 3 also tested the critical assumption of ELMO that conflicting probes should elicit more eliminative inferences than the imperfect and combined probes. To reiterate, in ELMO it is assumed that a probe with all features in common with a rule condition, or at most one deviating feature (i.e., either one missing feature or one feature too much), elicits induction, whereas a probe with two or more deviating features elicits elimination. One crucial implication is that the conflicting probe, *PC.PR*, which deviates from all inference rules with two or more features, will elicit elimination.

Thirty-eight participants took part in the experiment. Twenty-seven out of 38 participants (71%) met the learning criterion. Similar to Experiments 1 and 2, the use of base-rate information was evident for the imperfect (proportion "common" = .51±.15) and combined (proportion "common" = .49±.17) probes. For the conflicting transfer probe the participants did show an inverse base-rate effect (proportion "common" = .16, and proportion "rare" = .31), but the modal response was not found in the rare category, but the novel one (proportion "other disease" = .48) (see Figure 7); the novel disease effect.

The data also showed a difference in eliminative responding between the conflicting vs. the imperfect and combined probes as assumed by the eliminative inference approach. The mean response proportions for these probes were subjected to a planned comparison test, the conflicting probe (proportion "other disease" = .48) vs. the imperfect and combined probes (proportion "other disease" = .32). There was significantly more elimination for the conflicting probe, $F\,(1, 26) = 4.99, p < .05$.

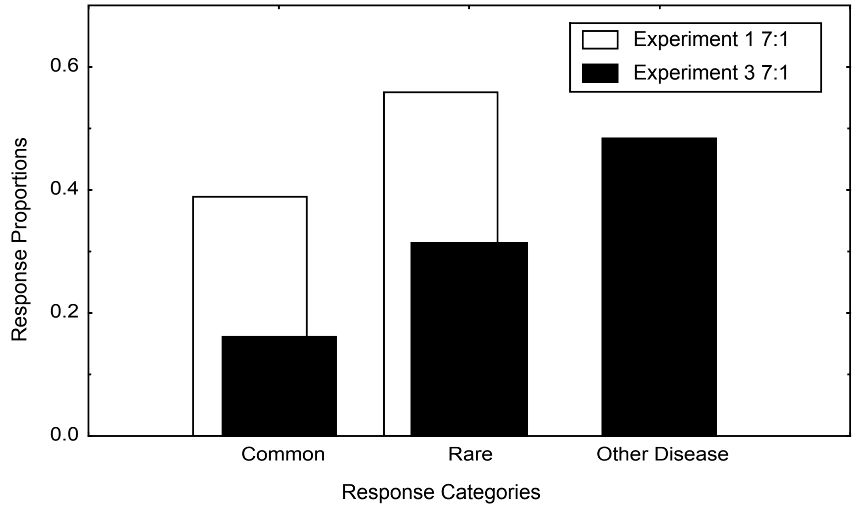


Figure 7. Mean response proportions of common, rare and "other disease" categories for the conflicting probes in Experiment 1 (base-rate ratio 7:1 after Training Session 1) and Experiment 3. Copyright © 2001 by the American Psychological Association. Reprinted with permission.

3.2.4. Summary of Study I. Study I successfully replicated the qualitative pattern of base-rate effects in the Medin and Edelson (1988) design, and provided a quantitative implementation of the eliminative inference idea that demonstrated that the high-level reasoning processes of ELMO can produce the basic pattern of base-rate effects in the inverse base-rate design (see also Kruschke, in press). A number of qualitative predictions derived from the eliminative inference idea were also confirmed. The most important of these were the predictions of “beginner’s luck”, which demonstrates that participants are better than chance already from the outset for rare-category exemplars, the “novel symptom effect”, which demonstrates that participants clearly favor the rare category for novel and conflicting probes, and the “novel disease effect”, which suggests that eliminative inferences of common diseases mediate the inverse base-rate effect rather than the strongest cue-category associations. In addition, Experiment 3 confirmed the critical assumption of ELMO that the conflicting probe, *PC.PR*, elicits more elimination than the imperfect, *I*, and combined probes, *I.PC.PR*. Finally, although ELMO correctly identified the modal response for each transfer probe, the quantitative fit was very poor when the task was changed by adding a novel category in the transfer phase. The problem was that the participants eliminated for the transfer probes where ELMO predicted that they should have induced their responses (i.e., *I* and *I.PC.PR*). It appears that once provided with the additional option “other disease”, the participants often preferred to respond that these probes, too, bear no straightforward relation to the inference rules in training (e.g., $I.PC \rightarrow C$). The introduction of a novel category seems to have raised the criterion for induction, leading to an increased overall rate of eliminative inferences. Thus, the criterion for inductive and eliminative inferences is likely to be a continouos function of similarity, and the function may differ depending on the set of response options. More generally, the basic qualitative patterns in Study I are consistent with the eliminative inference idea, but inconsistent with attentional theory.

3.3. Study II (Wennerholm, Winman, & Shanks, 2001)

The purpose of Study II was to trace the representations underlying the inverse base-rate effect. Participants were divided into groups which on different grounds could be expected to rely more or less on rule-based responding. The derivations in regard to the predictions by attentional theory were based on a literature search of key concepts and definitions as well as a computer simulation of ADIT (Kruschke, 1996; see also Study III). Two experiments and one reanalysis of Experiment 1 from Study I were included. Experiment 4 addressed *no inverse base-rate effect in children*. The reanalysis of Experiment 1 from Study I and Experiment 5 addressed *a stronger inverse base-rate effect in rule-based learners.*

In Study II it was explicitly assumed that the abstraction processes evoked in the patterning task in the Shanks and Darby task (Experiment 2, 1998; cf. Section 1.2.3.1 of this thesis) are similar to the processes implied by the eliminative inference mechanism. In both tasks, participants need to “look beyond” the information given and respond to more abstract interrelations between the stimuli.

3.3.1. Experiment 4: No inverse base-rate effect in children. Experiment 4 compared the qualitative predictions from rapid attention shifts and eliminative inferences in regard to 8-9 years old children’s behavior when confronted with a modified version of the inverse base-rate task. The cover story was changed to concern sick dogs instead of human patients. Participants were asked to pretend that they were veterinarians and their task was to cure each dog with a pill. To ensure that the participants encoded each dog with certain symptoms as a *unique* dog (rather than as the ”same” dog reappearing at the animal hospital later) each dog was given a unique name. Following Kruschke (1996), the number of features were decreased from nine to six, and the number of categories from six to four (cf. Table 1). In addition, the common diseases occurred five times more often than the rare ones (cf. Study I). In other respects, the method of Study II was identical to Study I. Fifty children and forty-nine adults participated.

The learning phase automatically stopped when the participants had reached a learning criterion of 22/24 (92%) correct responses on the last two blocks of learning (one block consisted of 12 trials)[9]. Each participant was required to solve at least 96 training trials and the maximum amount of trials was 180. In the adult group, 45 out of 49 participants (92%) met the criterion within the maximum number of trials. In the child group, 38 participants out of 51 (74.5%) met the criterion. On average, the adults reached the learning criterion after 100.8 trials, whereas the children reached it after 111.7 trials.

Similar to the previous experiments, the adult data replicated the standard pattern in the Medin and Edelson design with base-rate use for the imperfect and combined probes, and an inverse base-rate effect for the conflicting probes. Similarly, in the child group the use of base-rate information was evident for the imperfect and combined probes, and only a modest inverse base-rate effect was observed for the conflicting probe (for details, see Table 2 of Study II).

The participants’ responses on the conflicting transfer probes *PC.PR* were subjected to an additional analysis. Each participant responded to the conflicting probes four times. For these probes, participants were classified as exhibiting *base-rate use* if they showed a preference for the common category, *base-rate indifference* if they showed no preference for either the common or the rare category, and *base-rate inverse* if their responses predominantly were toward the rare categories. A majority of he adults exhibited the inverse base-rate effect, whereas most of the children were indifferent (see Figure 8).

In addition, for the children who outperformed the average adult in the learning phase of the inverse base-rate task, the lack of an inverse base-rate effect was actually more pronounced compared to the child group as a whole. Figure 9 shows the results of dividing both groups of participants (through a median split) into two subgroups on the basis of the number of trials it took them to reach the learning criterion. Whereas the most efficient adults show a larger inverse base-rate effect than the least efficient adults, the contrary pattern is found for the children. This interaction indicates a qualitative, rather than a quantitative difference between children and adults (*F* (1, 78) = 1.26; *N.S.*, *MSE* = .23).

[9] In the preceding experiments (1, 2, and 3) the learning criterion has been 23/24 (96%) correct responses in the last two blocks of learning. However, in Experiment 4 this criterion was somewhat lowered to 22/24 (92%) correct responses in order to receive a larger sample in the child group. It should be noted, however, that the response pattern with the smaller sample is very similar to the presented results.

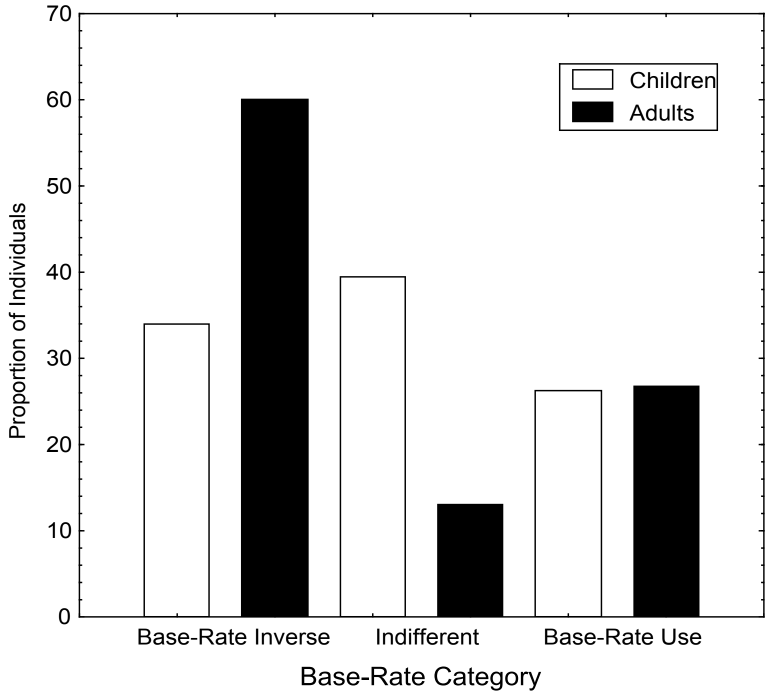


Figure 8. Proportion of children and adults who use the base-rates, are indifferent, or exhibit the inverse base-rate effect in Experiment 4.

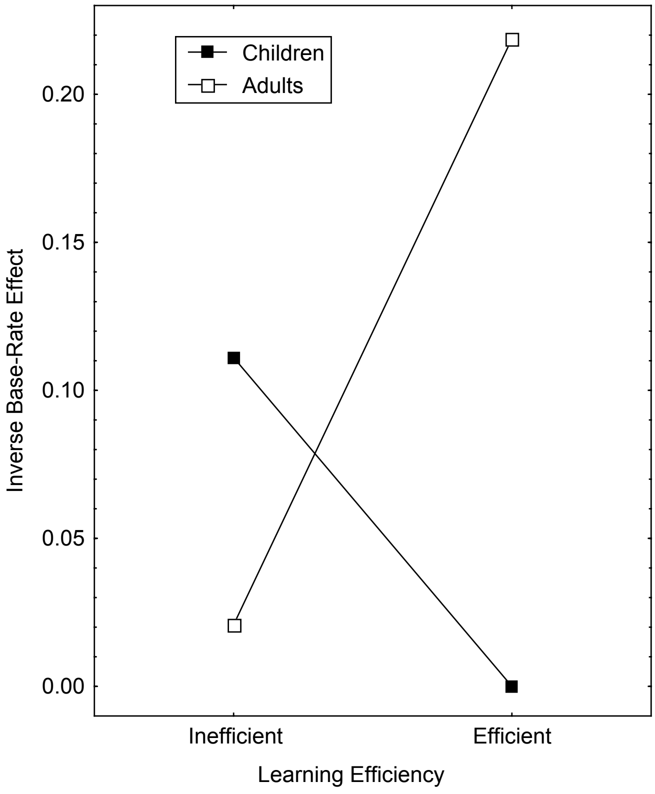


Figure 9. Size of the inverse base-rate effect (choice percentage difference favoring the rare disease) as a function of cognitive development (adults vs. children) and efficiency of learning.

3.3.2. Reanalysis of Experiment 1 from Study I: A stronger inverse base-rate effect in rule-based learners. The purpose of the reanalysis of Experiment 1 from Study I was to investigate the prediction of "a stronger inverse base-rate effect in rule-based learners" through a reanalysis of the data. All 109 participants who had completed the original task were divided into "Efficient" and "Inefficient learners", based on their overall proportion correct in the learning phase split at the 33rd and 66th percentile, respectively. This resulted in 75 participants. No learning criterion was applied in order to maximize the differences in learning performance and to obtain a higher statistical power. The division into efficient and inefficient learners was done separately for the two base-rate groups. Thus, there was approximately an equal number of participants from these two conditions in each learning group. This resulted in 37 participants classified as efficient learners and 38 participants classified as inefficient learners. The average proportions of correct responses in these two groups were .938 (*sd* = .023) and .729 (*sd* = .087), respectively.

The analysis was confined to the conflicting probe *PC.PR*. As can be seen in Figure 10 below, after Training Session 1, the inefficient learners demonstrated *base-rate use* in contrast to the efficient learners who showed an inverse base-rate effect. After Training Session 2, the inefficient learners still showed a modest preference for the common diseases, and the efficient learners again showed a clear inverse base-rate effect. This crossover effect in both learning stages results in a significant interaction between learning performance and the choice proportions for the rare and common diseases (F (1, 73) = 5.9, $p < .05$).

It could be the case that the inefficient learners failed to learn the task altogether, and that the significant group difference is an effect of more or less random responding on the part of those participants. However, a separate analysis excluding participants who failed to reach the learning criterion indicates a similar pattern to the data presented in Figure 10. Thus, the results cannot be a consequence of the inclusion of participants still at a pre-asymptotic level of learning. This should also be evident from the results illustrated in the right panel of Figure 10, where it can be seen that the inverse base-rate effect clearly remains after Training Session 2 when all participants have received twice as many learning trials as is normally the case.

A second objection to the results is that efficient learners are not necessarily more prone to learn rules than are inefficient learners in the inverse base-rate effect task. In fact, that efficient learners should exhibit strong inverse base-rate effects follows from the attentional theory too (J. K. Kruschke, Personal communication, July 17, 2001), provided that attentional shifts speed up the learning process (cf. Sections 1.3.1.3 and 3.4.1 of this thesis). But attentional theory does not predict that rule-based learners should exhibit the inverse base-rate effect, only that learners with the strongest cue-category associations should exhibit it. Therefore, in Experiment 5 a second task was used to independently assess participants' proneness to apply rules. Moreover, a statistical control for the learning performance was applied in that experiment to exclude the possibility that learning performance differences rather than differences in rule-learning proneness caused the results of the reanalysis of Experiment 1 from Study I.

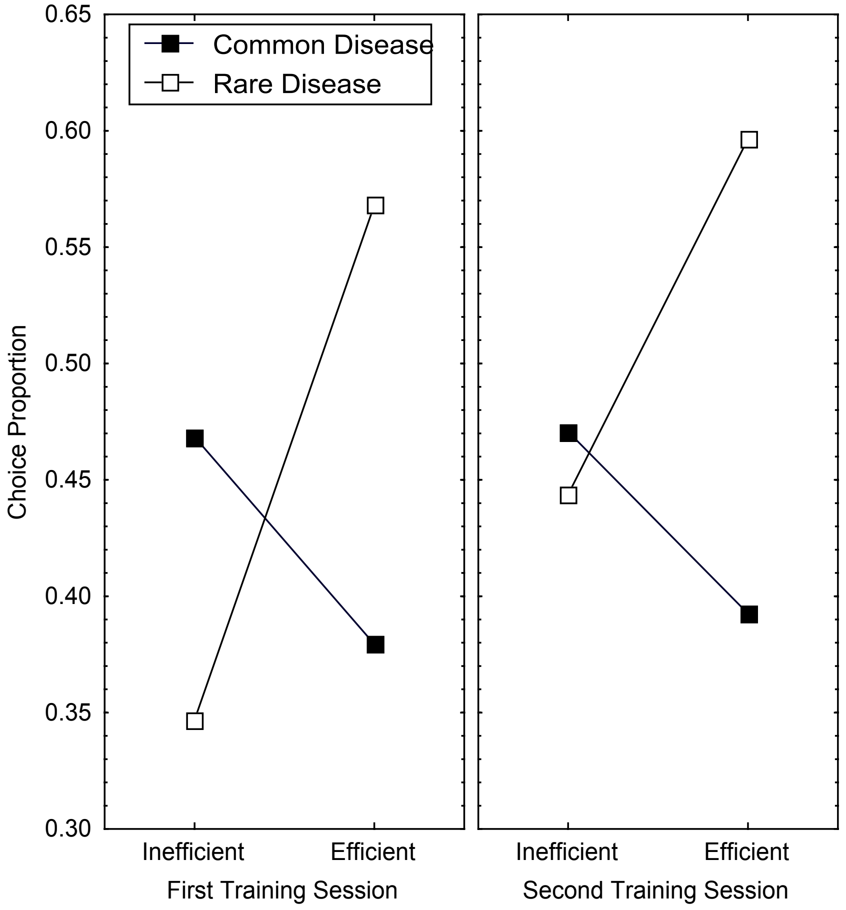


Figure 10. Response proportions for the *common* and *rare* diseases of the conflicting transfer probe *PC.PR* after the first (left panel) and second (right panel) learning phase for participants classified as *inefficient* and *efficient learners* in the reanalysis of Experiment 1 from Study I.

3.3.3. Experiment 5: A stronger inverse base-rate effect in rule-based learners. Ninety-seven participants took part in both the Medin and Edelson task, and the Shanks and Darby task discrimination task in counter-balanced order. After completion of the Shanks and Darby task, each participant answered an open questionnaire concerning the explicit knowledge they had about it. In addition, the time-limited (40 minutes) version of Raven's Advanced Progressive Matrices test (Raven, Raven, & Court, 1993) was administered to all participants to investigate if psychometric intelligence has any relationship with rule-based learning.

The Medin and Edelson task was identical to Experiment 1, but only one learning- and transfer phase was used and the base-rate ratio was 7:1 for all participants. Eighty-two out of 97 participants (84%) met the learning criterion 23/24 (96%) correct responses in the last two blocks of learning. For these participants, there was a clear inverse base-rate effect in the data (proportion "common" = .41; proportion "rare" = .53; t (81) = 2.16, p < .05).

In the Shanks and Darby task all participants were trained on the same number of items, 10×18 trials. After completion, participants were divided into three groups on the basis of their proportion correct in the final block of learning (cf. Shanks & Darby, Experiment 2, 1998). Participants achieving a score of 14 or lower were assigned to Group Low (*M* = 11.4, *SD* = 2.1, *n* = 40), those achieving a score between 15 and 17 were assigned to Group Medium (*M* = 16.2, *SD* = .79, *n* =38), and those scoring 18 were assigned to Group High (n = 19). In the following, this measure is referred to as the index of participants' *learning strategy*. The collapsed data of the trials in Figure 11 give a clear picture: as learning performance increases, the tendency to rely on rule-based generalization increases, and the proportion of similarity-based generalizations decreases. A one-way *ANOVA* with classification of participant as the independent variable, and proportion of rule-based generalizations as the dependent variable, demonstrates that this difference is statistically reliable (*F* (2, 94) = 8.0, *p* < .05).

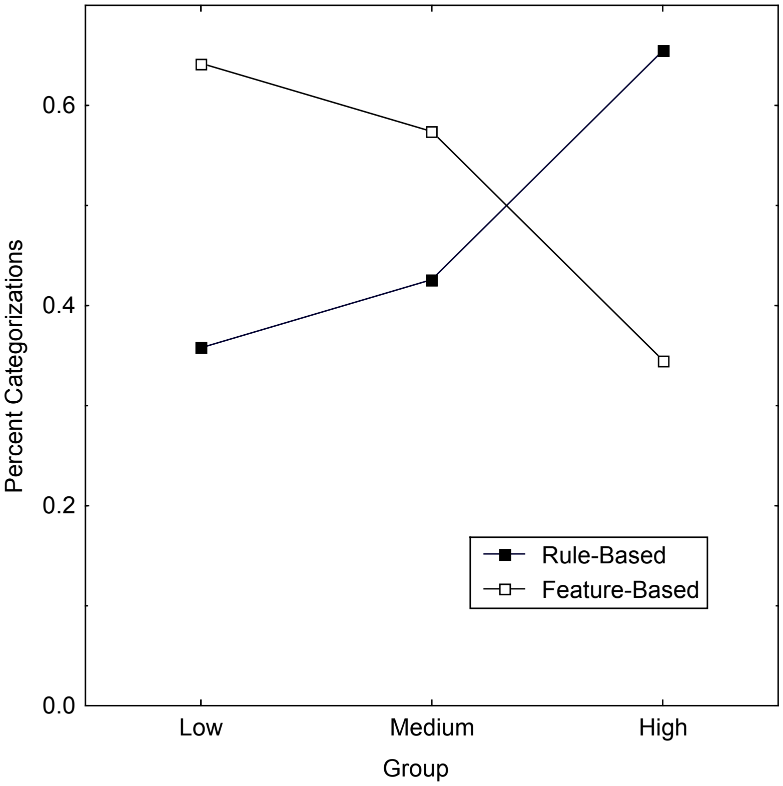


Figure 11. Proportions of *rule-based* and *feature-based* categorizations on novel transfer probes as a function of classification of participants in the last block of learning in the Shanks and Darby task.

The upper panel of Figure 12 shows the pattern of responding on the conflicting transfer probe *PC.PR* in the Medin and Edelson task as a function of learning strategy in the Shanks and Darby task. Group Low prefer the common over the rare diseases, whereas the opposite pattern is observed for the participants in Group High. Within-subjects t-tests reveal that the difference in preference between the common and rare diseases is only significant for Group High (*t* (18) = 2.2, *p* < .05). An independent t-test shows that the difference in the inverse base-rate effect between Group Low and Group High is statistically significant (*t* (57) = 2.0, *p* <.05).

The participants were rank-ordered by the proportion of rule-based responses to the critical transfer probes in the Shanks and Darby task (cf. Table 3), and divided into three groups by the 33rd and 66th percentile. The lower panel of the Figure 12 demonstrates the results of classifying participants into these groups: feature-based-, mixed-, and rule-based learners according to their mode of generalization to the novel critical transfer probes rather than on their learning strategies. This measure will be referred to as the index of participants' *mode of generalization*. The pattern is virtually the same as when the learning performance criteria by Shanks and Darby (1998) are used.

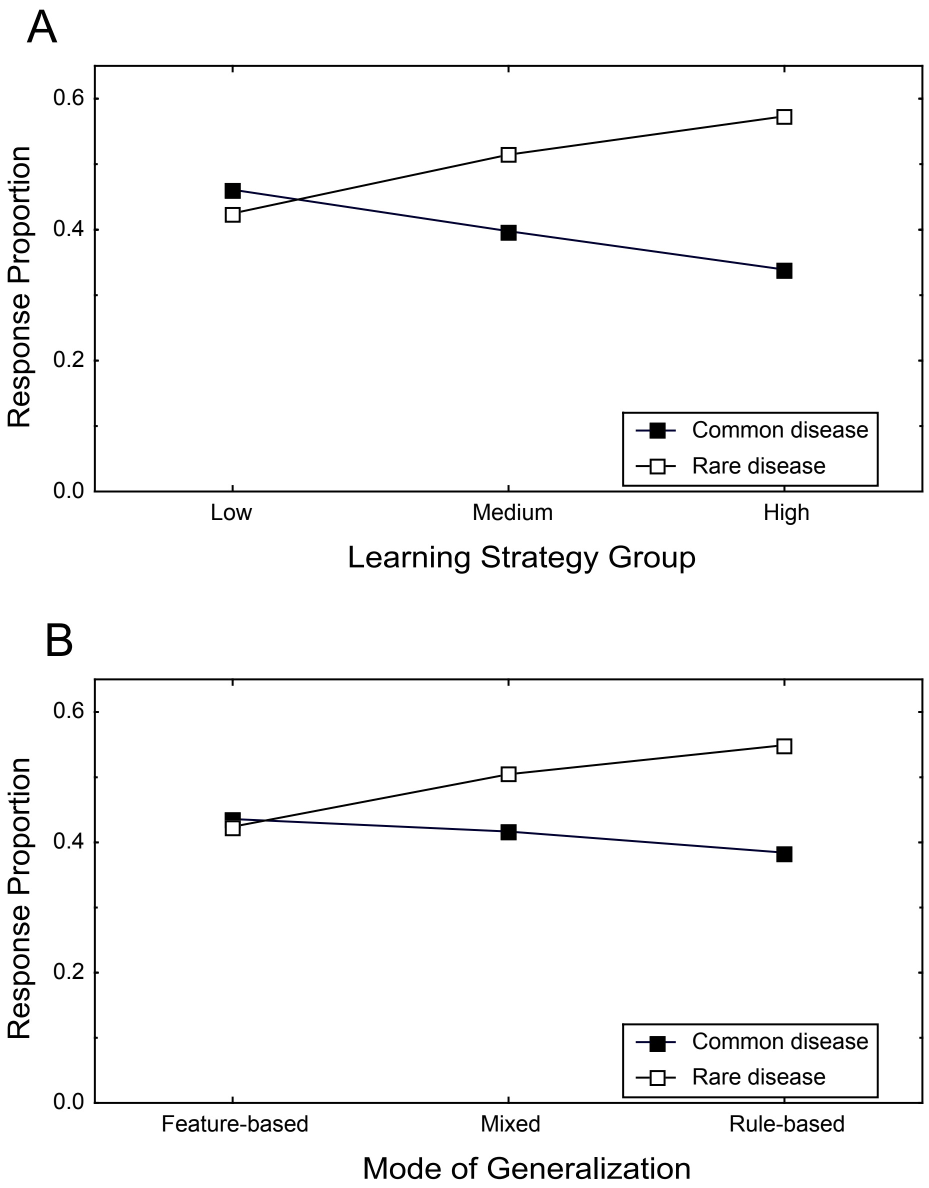


Figure 12. Response proportions for the *common* and *rare* diseases on the conflicting transfer probe, *PC.PR*, as a function of categorization of participants in the last learning block of the Shanks and Darby task (Panel A), and as a function of responses to critical test probes in the same task (Panel B).

Participants in Group High also performed better than the other participants in the learning phase of the Medin and Edelson task. They had an overall proportion correct of .91 in the learning phase of the Medin and Edelson task, compared to .84 for Group Low. This difference in learning performance between the groups approaches statistical significance (*t* (56) = 1.7, *p* < .10), which suggests a positive relation between learning efficiency and rule-learning propensity.

A matching procedure was undertaken to investigate whether the results were caused by differences in learning strategies in the Shanks and Darby task or by learning differences in the Medin and Edelson task. Participants in Group High and Low were ordered according to their overall proportion of correct diagnoses in the learning phase of the Medin and Edelson task. For each participant in Group High all participants with a comparable score in Group Low were entered into the analysis. This resulted in two groups of approximately the same overall learning performance in the Medin and Edelson task (*M* = 89.9 % and 90.6 % correct in the feature-based and rule–based groups, respectively), differing only in learning strategy. A comparison of these two groups with respect to the inverse base-rate effect revealed that the rule-based group showed a strong effect (average difference between the common and rare response proportions = -.23, *SD* = .47), whereas the feature-based group demonstrated base-rate use (average difference = .08, *SD* = .47). Thus, the matching procedure did not reduce the size of the inverse base-rate effect, and an independent t-test confirmed the difference between the groups (*t* (42) = 2.15, *p* <.05). The same conclusion was drawn by treating the overall performance in the learning phase of the Medin and Edelson task as a concomitant variable in an *ANCOVA*, with learning strategy classification (rule-based or feature-based) as the independent variable, overall learning performance in the Medin and Edelson task as the covariate and the degree of the inverse base-rate effect as the dependent variable (*F* (1, 55) = 4.33, *p* < .05).

Participants were also classified as more or less aware of the patterning rule (i.e., *a compound and its elements predict opposite outcomes*, cf. Section 1.2.3.1) on the basis of their responses to the open questionnaire. Failure to report the correct rule was scored as zero, a partial report of the rule was scored as 1, and a full report of the underlying rule was scored as 2. A comparison of this explicit *rule awareness score* between the different learning strategy categories shows that Group High is more aware (*M* = 1.6, *SD* = .68) than Group Medium (*M* = .97, *SD* = .91), which in turn is more aware than Group Low (*M* = .4, *SD* = .7) (*F* (2, 94) = 16.1, *p* <.001). Similarly, the correlation between the overall learning performance in the Shanks and Darby task and the Raven's Advanced Progressive Matrices test was positive and significant (*r* (97) = .24, *p* < .05), which suggests that the Shanks and Darby task to a certain degree taps a general cognitive functioning ability.

3.3.4. Summary of Study II. Study II also successfully replicated the qualitative pattern of base-rate effects in the Medin and Edelson (1988) design (Experiment 4 and 5). In addition, a number of predictions derived from the eliminative inference idea were verified. First, support was found for the prediction of "no inverse base-rate effect in children". In Experiment 4, it was demonstrated that most children did not exhibit the inverse base-rate effect, whereas most adult did. Arguably, unlike most adult humans, children of the age group 8-9 years lack the cognitive sophistication needed to appreciate a rule-based, meta-cognitive strategy like elimination. Second, support was found for the prediction of "no inverse base-rate effect in rule-based learners". In the

Reanalysis of Experiment 1 from Study I, and in Experiment 5, it was demonstrated that the rule-based assumptions of the eliminative inference approach better captures the representations that mediate the inverse base-rate effect, even when learning efficiency is controlled for (Experiment 5).

3.4. Study III (Winman, Wennerholm, & Juslin, 2001)

Study III is a response to the critique of Study I in Kruschke (in press) entitled "The inverse base-rate effect is not explained by eliminative inference". In Kruschke (in press), old and new data are presented that are problematic for the quantitative implementation of the eliminative inference approach, ELMO, as formulated in Study I. Instead, it is claimed that a novel connectionist implementation of attentional theory, EXIT, fits the data well.

The critique in Kruschke (in press) revolves around three issues. The first is that there are systematic deviations between the predictions by ELMO and human data for some transfer probes of the original inverse base-rate effect design, most notably that participants prefer the common disease, *C*, more strongly when presented with the imperfect predictor, *I*, than with the combined probe, *I.PC.PR*, whereas ELMO predicts equal response proportions for these probes (cf. Section 2.4). The second concern is that the inverse base-rate effect appears to depend on the presence of a shared imperfect predictor in the training phase (Experiment 1, Kruschke, in press). The inverse base-rate effect was not observed for a common-rare disease pair without such a shared symptom. The third criticism concerns data when the base-rates are reversed during training (Experiment 2, Kruschke, in press).

The purpose of Study III was twofold. First, to evaluate the proposal that attentional theory provides a better account of the inverse base-rate effect than the eliminative inference approach. Three central claims emanating from attentional theory were investigated: (a) that rapid attention shifts accelerate learning in the inverse base-rate design, (b) that the inverse base-rate effect is caused by a stronger association between *PR-R* than between *PC-C*, and (c) that it is participants relying on cue-category associations that contribute to the inverse base-rate effect. The second purpose was to propose a number of possible remedies to the shortcomings of ELMO, as discussed in Kruschke (in press; but see also Study I).

3.4.1. A simulation of ADIT: Rapid attention shifts decelerate learning. In Kruschke (1996, in press) it is argued that rapid attention shifts speed up the learning process. This proposal was investigated by a computer simulation of ADIT (Kruschke, 1996). Figure 13 presents the proportion of correct classifications predicted by ADIT for five different rates of attention-shifting. The predictions are appropriate for the simplified design in Experiment 1 from Kruschke (1996). The other parameters were kept at the values fitted to the data in Kruschke, Experiment 1, 1996, where $\phi = 4.16$, $\beta = .268$, and $\eta = 2.63$. Each data point contains 500 simulated participants confronted with 120 training trials each. The simulations reveal that in contrast to the functional rationale for the rapid attention-shifting mechanism, it does not seem to accelerate learning in the inverse base-rate effect task, but if anything to *decelerate* it. In fact, the most rapid learning is obtained with no attention-shifting at all. The reason why ADIT makes this prediction is that it does not learn its attentional shifts from one trial to the next. The failure to learn attentional shifts causes ADIT to give lower accuracy with higher attentional shifts. Thus, this simulation gives no support for the proposal that rapid attention shifts *per se* should speed up the learning process. It remains to be carefully specified under what conditions this alleged benefit in attentional theory occurs.

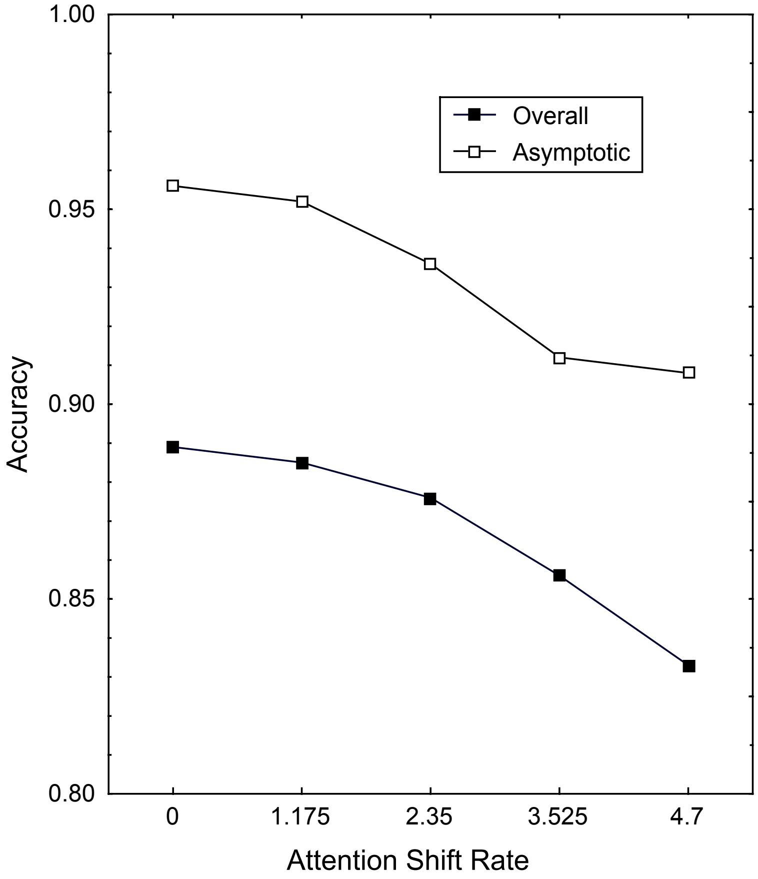


Figure 13. Implementation of ADIT (Kruschke, 1996), simulating overall- and asymptotic accuracy (two last blocks of learning) as a function of five different rates of attention-shifting.

3.4.2. An aggregated analysis: Whenever there is an inverse base-rate effect, *PC* should elicit more category *C* responses than *PR* should elicit category *R* responses. According to attentional theory, whenever there is an inverse base-rate effect in participants' data, *PR* should elicit more category *R* responses than *PC* should elicit category *C* responses. In contrast, according to the eliminative inference approach, *PC* is better learned than *PR* because it is part of the conditions of the inference rule *I.PC*→*C*, which has a higher base-rate of occurrence. These two contrasting explanations of the inverse base-rate effect were tested by collapsing previously collected data in the Uppsala lab. The data were taken from published and unpublished experiments with an inverse base-rate effect design using standard procedures with exclusion of participants not meeting a specified learning criterion. As evidenced in Figure 14, this analysis demonstrated a clear inverse base-rate effect (*t* (335) = 4.6, *p* <. 001), and a significantly higher response proportion for *C* on the *PC* probe than for *R* on the *PR* probe (*t* (335) = 4.2, *p* <. 001).

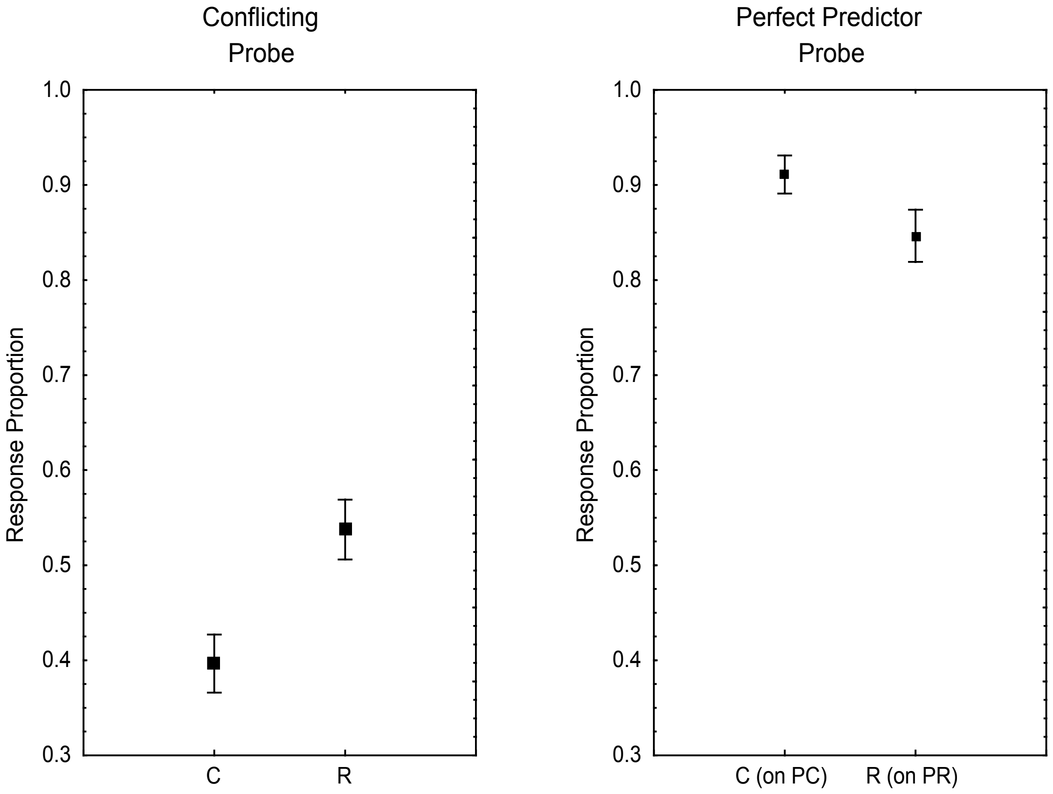


Figure 14. Left panel: Observed response proportions of the common (*C*) and rare (*R*) diseases for the conflicting probe (*PC.PR*) with 95% confidence intervals. Right panel: Observed response proportions of the common diseases (*C*) given *PC* and the rare diseases (*R*) given *PR*, respectively, with 95% confidence intervals.

Moreover, a quick search through the literature on the inverse base-rate effect [10], including previously unpublished data, demonstrated that out of 23 conditions there were 16 in which $p(C|PC) > p(R|PR)$ and 6 in which $p(C|PC) < p(R|PR)$. Thus, the data revealed a pattern inconsistent with attentional theory, but consistent with the eliminative inference approach.

3.4.3. A stronger inverse base-rate effect in rule-based learners. In Study III, the previously reported data from Experiment 5 in Study II were restated to demonstrate that the cue-category associations that mediate the inverse base-rate effect according to attentional theory are refuted by data showing that only the rule-based learners exhibit the inverse base-rate effect, whereas the feature-based learners *use* the base-rates (see Figure 12).

[10] Study I (Experiments 1-2), Kruschke (1996, Experiments 1-3), Medin and Bettger (1991, Experiment 1), Medin and Edelson (1988, Experiments 1-4), Shanks (1992, Experiment 1), Wennerholm (2001), and Study II (Experiments 4 and 5).

3.4.4. A theoretical discussion: Possible remedies to the shortcomings of ELMO. The second part of Study III was theoretical in nature and discussed possible remedies to the shortcomings of ELMO, as discussed in Kruschke (in press). The first concern is that there are systematic deviations between the predictions by ELMO and human data for some transfer probes of the original inverse base-rate effect design, most notably that participants prefer the common disease, *C*, more when presented with the imperfect predictor, *I,* than with the combined probe, *I.PC.PR*, whereas ELMO predicts equal response proportions for these probes. In Study III, it was suggested that this asymmetry occurs because larger importance is given to *PR* than to *PC* in the similarity computations, essentially in line with attentional theory. ELMO is easily modified to allow for more weight to *PR* than to *PC* by entering separate similarity parameters for the perfect predictors. However, Figure 14 clearly suggests that attentional asymmetry by itself cannot explain the data.

The second concern is that the inverse base-rate effect appears to depend on the presence of a shared imperfect predictor in the training phase. The inverse base-rate effect was not observed for a common-rare disease pair without such a shared symptom (Experiment 1, in Kruschke, in press). Unfortunately, there are some aspects of this experiment that render the interpretation of the results problematic. First, in contrast to previous experiments in which participants who fail to meet a predefined learning criterion are excluded, all participants were included in Kruschke's analysis. This is important, because as previously demonstrated (Study II), group data consists of a mix of two sub-populations with diametrically opposite behavior and different generalization patterns: Inefficient learners use the base rates whereas efficient learners exhibit the inverse base-rate effect. Therefore, it is questionable whether it is fruitful to model them quantitatively within a single representational scheme. Second, in Kruschke (in press) all results are subjected to Chi-square tests. An assumption underlying the use of this statistical test is that the observations in the sample are independent of one another. Because the tests are based on raw frequencies with each participant contributing with several responses for a single probe, this assumption is violated throughout the article (a fact that is acknowledged for only one of the ten tests), resulting in highly inflated chi-square values. In a previous article by Kruschke (1996), the violation of independence is controlled for by the conventional procedure of dividing the Chi-squares by the number of responses given by each participant. If the same procedure is applied to the tests of Experiment 1 the resulting Chi-squares are low and non-significant. Contrary to what is implied in Kruschke (in press), they are *not* in the region of approaching statistical significance. Nevertheless, if future studies were to confirm the observed base-rate pattern there are two reasons why the absence of an inverse base-rate effect for the common-rare disease pair without a shared imperfect predictor can be integrated into the eliminative inference approach.

First, the criterion for induction/elimination in Study I was merely a simple way of capturing the idea that people tend to induce for *I* and *I.PC.PR* and eliminate for *PC.PR*, an intuition that was verified in Experiment 3 of Study I. However, this specific criterion was found to be questionable by the results of the same experiment in that people eliminated more than implied by the criterion, and therefore, it was proposed that the criterion for inductive and eliminative inferences is likely to be a continuous function of similarity, and that that function may differ depending on the response options.

In Kruschke's Experiment 2, participants learn that two perfect predictors PaC_1 and PbC_1 lead to a common disease C_1 and that two perfect predictors PaR_1 and PbR_2 lead to disease R_1. If one considers that the imperfect predictors I are redundant for classification in the training phase in the original design (i.e., $I.PC \rightarrow C$, and $I.PR \rightarrow R$), it would seem reasonable for an inference mechanism to attach more weight to the perfect than to the imperfect predictors. In ELMO, this means that the similarity of a probe to a rule would decrease more due to a missing perfect than due to a missing imperfect predictor (formally expressed as $s_P<s_I$).

For the rules formed for the disease-pairs that involve no shared imperfect predictors ($PaC_1.PbC_1 \rightarrow C_1$ & $PaR_1.PbR_1 \rightarrow R_1$), the weight attached to the two features would be distributed equally across the two predictors. Given that attention is limited and shared across the two features, less weight would probably be attached to any of these two perfect predictors in isolation than to the single perfect predictor of the original design.

The second reason why the stronger inverse base-rate effect for $PC.PR$ than $PaC_1.PaR_1$ may be compatible with eliminative inference is more basic. Due to interference, it is reasonable to assume that inference rules that involve shared symptoms are more difficult to learn than inference rules that only contain perfect predictors. Because of the higher similarity (due to a common feature) of the former rules, the learning of $I.PC \rightarrow C$ will interfere more with the learning of $I.PR \rightarrow R$, than the learning of $PaC_1.PbC_1 \rightarrow C_1$ will interfere with the learning of $PaR_1.PbR_1 \rightarrow R_1$. If removal of a shared symptom enhances learning of these rules, this may explain the disappearance of the inverse base-rate effect effect. Although this explanation is admittedly ad hoc, enhanced learning of the pair with no shared symptom is a straightforward prediction from widely accepted knowledge in cognitive psychology (i.e., interference theory). A further hint that this may in fact be the explanation is found in Kruschke (in press): "It is clear from the test data that people learned the disease of the pair with no shared symptom much better than the rare disease of the pair with a shared symptom".

The third criticism concerns data when the base-rates are reversed in the training phase. The base-rate that occurs early in training has proved to be more important (Medin & Bettger, 1992). Thus, if there is an initial 3:1 base-rate ratio that is changed into a 1:3 base-rate ratio in the latter half of the training phase, the initial 3:1 ratio contributes to an overall inverse base-rate effect at the end of training, even though the overall base-rate is 1:1. This result fits nicely with attentional theory but also with a version of ELMO where the similarity parameters are primarily determined by the early phase of training. The first implementation of ELMO included such a "freezing" parameter to capture this phenomenon, but it was discarded for reasons of parsimony.

3.4.5. Summary of Study III. Study III demonstrated that rapid attentional shifts *per se* do not accelerate learning (cf. beginner's luck in Study I), but rather decelerate it, and it was proposed that the conditions under which circumstances this alleged benefit occurs need to be more carefully specified. Study III also presented an aggregated analysis that confirmed the prediction "whenever there is an inverse base-rate effect, *PC* should elicit more category *C* responses than *PR* should elicit category *R* responses”. In addition, the result from Experiment 5 from Study II demonstrating that only rule-learners exhibit an inverse base-rate effect, was restated. Finally, a theoretical discussion of possible remedies to the shortcomings of ELMO raised in Kruschke (in press), was presented. These findings suggest that a complete account of the inverse base-rate effect needs to integrate inductive- and eliminative inferences operating on rule-based representations with attentional shifts.

4. Discussion

4.1. Summary

This thesis has presented a novel explanation of the inverse base-rate effect (Medin & Edelson, 1988), the eliminative inference approach. The proposal is that participants eliminate category options that are inconsistent with well-supported inference rules. These assumptions contrast with those by attentional theory (Kruschke, 1996, in press), according to which the inverse base-rate effect is the outcome of rapid attention shifts operating on cue-category associations. Thus, the eliminative inference approach and attentional theory differ in two important ways. First, the mechanisms proposed to account for the inverse base-rate effect are theoretically different: Eliminative inference is a rule-mediated *response* strategy, whereas rapid attention-shifting is a feature-mediated *learning* strategy. Second, the representational assumptions underlying the two accounts are different: The eliminative inference approach postulates simple inference rules, whereas attentional theory postulates cue-category associations. The primary question addressed in this thesis was whether associative principles in general, and the notion of rapid attention shifts operating on cue-category associations in particular, provide the only or most important account of the inverse base-rate effect. This question was addressed in three studies.

In Study I, a quantitative implementation of the eliminative inference idea, ELMO, which was fitted to data in Kruschke (Experiment 1, 1996; but see also Study I), and demonstrated that this mechanism may be a significant contributor to the inconsistent pattern of base-rate effects in the inverse base-rate design, that is base-rate use for imperfect, *I*, and combined predictors, *I.PC.PR*, and an inverse base-rate effect for conflicting predictors, *PC.PR* (cf. Figure 5; but see also Kruschke, in press). The Empirical Section of Study I found additional support for the assumptions of the eliminative inference approach: Study I confirmed the predictions of “beginner’s luck”, “the novel symptom effect”, “the novel disease effect”, and “the repetition effect”. Study II confirmed the predictions of “no inverse base-rate effect in children”, and “a stronger inverse base-rate effect in rule-based learners”. Study III, finally, confirmed the prediction that “whenever there is an inverse base-rate effect, *PC* should elicit more category *C* responses than *PR* should elicit category *R* responses”. The confirmation of these predictions is important because they are inconsistent with the predictions by attentional theory (Kruschke, 1996, in press), as well as other models in the cue-competition tradition (e.g., Gluck, 1992; Gluck & Bower, 1988).

Specifically, in regard to the qualitative theory-critical predictions, the two most important results in Study I were the “novel symptom effect” and the “novel disease effect”. In Experiment 1 and 2, a majority of the participants favored the rare category when presented with a novel probe, a response pattern that mirrors the inverse base-rate effect for conflicting probes, *PC.PR* (cf. Figure 6). This finding suggests that the mechanism involved in classifications of these two probes may be similar. In Experiment 3, the participants favored the novel disease, never seen during learning, when presented with the conflicting probes, *PC.PR* (cf. Figure 7). This tendency was stronger for the conflicting probes, *PC.PR*, than for the imperfect, *I*, and combined probes, *I.PC.PR*, as presumed by the eliminative inference explanation. Experiment 1 further confirmed the prediction of "beginner's luck", according to which accuracy should be better than chance already from the outset for rare-disease trials (cf. "non-random guessing strategy", Kruschke, 1996; Kruschke & Bradley, 1995). Finally, Experiment 2 demonstrated that the diminished inverse base-rate effect after Training Session 2 in Experiment 1 was due to the repeated exposures with the transfer probes rather than to more extensive training; "the repetition effect".

In Study II the most important finding was that rule-based representations mediate the inverse base-rate effect. In the Reanalysis of Experiment 1 from Study I (cf. Figure 10) and Experiment 5 (cf. Figure 12) the data demonstrated that only efficient and rule-based learners, respectively, exhibit the inverse base-rate effect, as predicted by the representational scheme of the eliminative inference approach. In Experiment 4, it was further demonstrated that most adults exhibited a strong inverse base-rate effect whereas most children were indifferent in choosing between the common and rare categories when presented with the conflicting probe, *PC.PR* (cf. Figure 8). Arguably, unlike most adult humans, children of the age group 8-9 years lack the cognitive sophistication needed to appreciate a meta-cognitive strategy like elimination. This qualitative difference between children and adults would not be observed if the inverse base-rate effect was a result of either the strong or the weak interpretation of rapid attentional shifts (cf. Figure 9). In its present formalization the eliminative inference approach cannot explain this qualitative difference either. It would be interesting to see if this effect can be replicated and, if so, to investigate why the most efficient children were less prone to exhibit the inverse base-rate effect than the least efficient children were.

In Study III, finally, it was shown that in contrast to the proposed functional rationale for the rapid attention-shifting mechanism, it does not seem to accelerate learning in the inverse base-rate effect task, but if anything to decelerate it (cf. Figure 13). It was concluded that the conditions under which this alleged benefit occurs need to be more carefully specified. In the second section, it was shown that participants did not elicit more category *R* responses when presented with the transfer probe *PR* than category *C* responses when presented with the probe *PC*, although they still exhibited the inverse base-rate effect. Instead the results suggested that *PC* was better learned than *PR*, presumably because it is part of the conditions of the inference rule $I.PC \rightarrow C$, which has a higher base-rate of occurrence (cf. Figure 14), as predicted by the eliminative inference approach. This result is also consistent with the data in Kruschke (1996) as well as the overall pattern in the published studies on the inverse base-rate effect.

Taken together, as an account of the inverse base-rate effect the empirical evidence of this thesis suggest that rule-based elimination is a powerful component of the inverse base-rate effect. But previous studies have indicated that attentional shifts are important contributors to the inverse base-rate effect too (for a discussion, see Study I, III, and Kruschke, in press). Therefore, a complete account of the inverse base-rate effect needs to integrate inductive- and eliminative inferences operating on rule-based representations with attentional shifts. The theoretical discussion of Study III proposed a number of suggestions for such integrative work (but see also Section 4.2.4 below).

4.2. Implications and Suggestions for Future Research

4.2.1. Measurement error or individual differences in human knowledge representation? In all the experiments using the inverse base-rate design reported in this thesis, the variance in the data is immense, and difficulties in obtaining statistically reliable effects despite large samples and large deviations between means has frequently been encountered (see e.g., Study I). In retrospect, this is probably not a consequence of large measurement error *per se*, but an effect of large individual differences in the participants' responding: Some participants use a rule-based generalization strategy and exhibit strong inverse base-rate effects (such as efficient adult learners), whereas other participants use a feature-based generalization strategy and use the base-rates (such as baboons, Fagot et al., 1998, children and inefficient adult learners). These differences in human knowledge representation are interesting for several reasons.

First, they fit nicely with Reber's (1993) criteria for implicit and explicit cognition. Explicit rule-based processes are more age-dependent (Experiment 4), IQ-dependent (Experiment 5), and do not show cross-species-commonalities, presumably because the cortical structures involved in these high-level processes had a later evolutionary arrival. Second, they suggest that being a rational learner in one context (the Shanks and Darby task), is associated with a response that has been interpreted as being irrational in another (the inverse base-rate task). Why is it that good performance on one set of standards, being an efficient or rule-based learner, is associated with a bias that has been interpreted as being non-normative and irrational on a second? The first idea that comes to mind is that the inverse base-rate effect is in fact a rational response. This is at least in the spirit of the eliminative inference approach. It seems rational to eliminate the common disease when confronted with the conflicting transfer probe *PC.PR* because symptom *PR* has never occurred with the common disease during learning (i.e., the inference rule $I.PC \rightarrow C$). It would be interesting to investigate if the cause of other judgmental biases in the literature can be traced to similar representational differences, and perhaps even more interesting, to investigate what the reasons are for these individual differences in human knowledge representation. Evidently, psychometric intelligence is not the whole explanation. The explained variance in Experiment 5 between overall learning performance in the Shanks and Darby task and the Raven's Advanced Progressive Matrices test was only 5.8%.

One possibility that comes to mind has to do with people's memory-capacity. But before proceeding to this discussion, it is important to distinguish rule-formation from rule-application (see Section 4.2.2, see also Smith et al., 1992). The former process takes place during learning (or encoding), whereas the latter process takes place in the generalization (or retrieval) situation. It could be the case that "rule-learners" eventually become "exemplar-based learners", and in this sense move from semantic processing to episodic processing (cf. Logan, 1988, see also Table 2). For example, in the eliminative inference approach, it is proposed that people activate simple inference rules in the learning phase of the inverse base-rate task, for example, $I.PC \rightarrow C$. However, for this approach to the inverse base-rate effect it is not crucial that the encoding of elements in the learning phase is rule-based[11]. The main idea is that people eliminate in the *retrieval* situation, and this process is evidently mediated by rules (cf. Figure 12). As we have seen, rule-application involves an abstraction of configurations of features that are applied to a novel situation (cf. Experiment 5), one important criterion for rule-use. Moreover, in Experiment 5, it was shown that rule-based learners were successful in verbally reporting the patterning rules, another important criterion for rule-use (e.g., Ashby et al., 1998; see also Shanks & Darby, 1998). Provided that representational shifts from rules to exemplars take place during encoding, these learners could possibly make use of their episodic memory system as an additional buffer next to their working-memory in which the "on-line" elaborations of information take place. This idea fits well with the literature on expertise, which suggests that experts, such as chess-players, use their long-term memory when solving problems, such as retrieving exemplars of previously successful chess-moves. As a matter of curiosity, recent neuroscientific work using brain imaging methods supports this hypothesis too. A research team in France demonstrated that a mathematical genius, the German Rüdiger Gamm, activated his long-term memory when solving rule-based problems (in this case complex problems of multiplication) (Pesenti et al., 2001).

Second, and more generally, the demonstration of similarity-based and rule-based representations in the same cognitive task as a function of learning (or expertise) indicates that human knowledge representation cannot be faithfully captured by unified theories of either rules (symbols), associations (connections), or exemplars. Instead, hybrid models are required, and, indeed, this is the trend we now see in cognitive science (e.g., Anderson & Betz, in press; Erickson & Kruschke, 1998; Johansen & Palmeri, 2001; Palmeri, 1997; Vandierendonck, 1995; see also Ashby et al., 1998). This modular view seems to apply to the inverse base-rate effect too. For example, in the aggregated analysis in Study III[12], it was found that 55% (N=227) of the participants (N=413) exhibited the inverse base-rate effect (the rule-based learners), whereas the remaining 45% (N=186) used the base-rates (the feature-based learners). A plot of this analysis suggests a bimodal rather than a unimodal response distribution (see Figure 15).

[11] It may be true in the learning perspective investigated in this thesis, but in the long run it could well be the case that the inference rules eventually shift into exemplars, as previously suggested by Medin and Bettger (1991). Unfortunately, in the studies presented in this thesis it it not possible to determine whether such representational shifts have taken place. Perhaps future experiments could circumvent these problems by incorporating probes at different stages of learning that are diagnostic of different representational modes.

[12] In the aggregated analysis of Study III only the participants who has passed the learning criterion 23/24 correct responses on the last two blocks of learning were included (N=335). However, in the analysis presented here all participants are included (N=413).

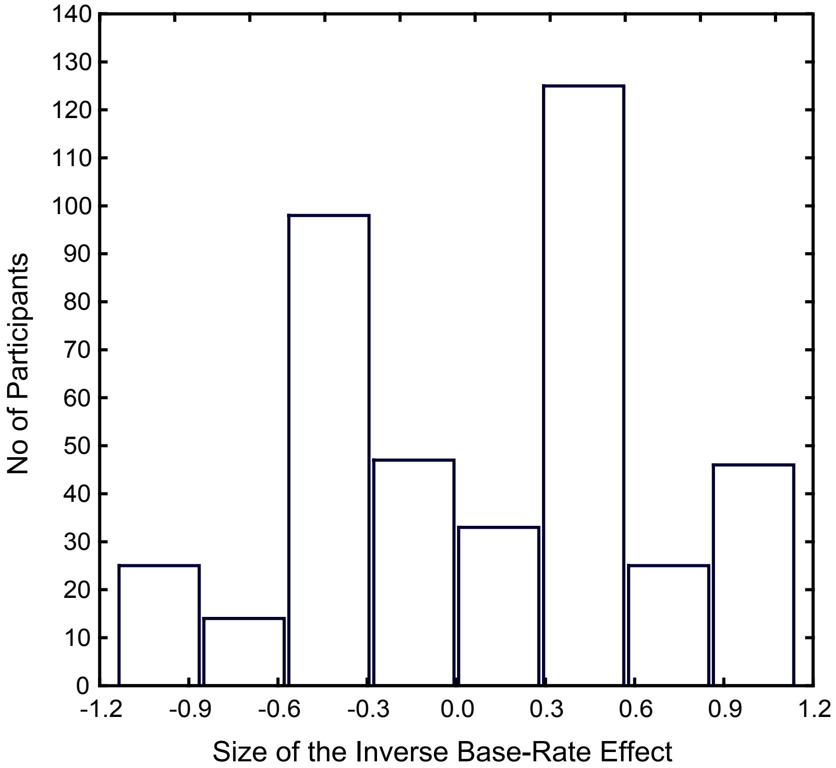


Figure 15. Number of participants in the aggregated analysis of Study III inversing or using the base-rate information when confronted with the conflicting transfer probe, *PC.PR* (N=413).

4.2.2. The ecological validity of the inverse base-rate effect. An obvious but important question that I have neglected thus far is whether the inverse base-rate effect occurs in real life, or whether it is an artificial phenomenon created in the laboratory. First, it is difficult to imagine a situation in real life in which two perfect predictors, one associated with a common outcome and another associated with a rare outcome, would reveal itself, and perhaps more importantly, in which you are “forced” to choose between them. Most of the time cues in the environment are probabilistically related to one another, such as “if the sky is grey it might rain”, but it might just as well “not rain”, and whether or not the base-rates for these events happen to be different, you are not ”forced” to choose between them. Thus, from an ecological standpoint, it is perhaps more reasonable to imagine a situation with one ambiguous predictor, such as “the sky is grey”, and that your belief in the event ”rain/no rain” is reflected on a continuous subjective probability scale rather than a forced-choice discrete one. Hence, for these reasons the inverse base-rate effect is probably a phenomenon of low ecological validity. On the other hand, if people were allowed to express their degree of belief in a given event, such as the probability of each disease in the inverse base-rate task given the symptom constellation *PC.PR*, would the inverse base-rate effect disappear then? In a related line of research, Wennerholm (2001) investigated this hypothesis in a between-subjects study. One group received the standard inverse base-rate task, and a second group received an identical learning phase, but a transfer phase in which they were allowed to express their subjective probability estimate for each disease given each critical transfer probe on an 11-point scale with eleven alternatives, ranging from 0% ”the patient’s symptoms are *absolutely not associated* with this disease” to 100% ”the patient’s symptoms are *absolutely associated* with this disease”. These probability estimates were then converted into response proportions to make them comparable to the original transfer condition. Interestingly, the participants in the ”probability group” still preferred the rare category when presented with the conflicting probe, although to a lesser degree than the original group. These results suggest that whatever processes produce the inverse base-rate effect, they do not only surface when participants make forced decisions in an ambiguous task that lacks a well-defined solution. Thus, there appears to be more to the inverse base-rate effect than can be dismissed by arguments about its low ecological validity. The transfer data in Experiment 2 suggest a similar conclusion since the participants still exhibited an inverse base-rate effect despite four times as many learning trials as is normally the case.

4.2.3. Toward an integrative account of the inverse base-rate effect. In general, the two predominant accounts of the inverse base-rate effect, the eliminative inference approach and attentional theory, have received some support by model fits and empirical data, but they have also encountered difficulties in explaining certain aspects of data in implementing data (see e.g., Study I, II and III of this thesis, and Kruschke 1996, in press). The present formalizations of attentional theory (ADIT, Kruschke, 1996, & EXIT, Kruschke, in press) cannot account for the data presented in this thesis, and the present formalization of the eliminative inference approach (ELMO, Study I) cannot account for data presented in Dennis and Kruschke (1998) and Kruschke (in press, but see Section 3.4.4 for a review of possible remedies to the shortcomings of ELMO). Nevertheless, in the Discussion of Study III of this thesis, it was proposed that: ”Altogether, the data presented in this article (i.e., Study III) and elsewhere (referring to Study I and II, and Kruschke, 1996, in press) suggest that a more promising and parsimonious approach is to integrate the attention shifts directly into a representational scheme that involves flexible, high-level inference rules that are used for both induction and elimination” (parentheses added). There are a number of reasons for this suggestion, which I will detail next.

First, as we have seen, attentional theory is unable to account both qualitatatively and quantitatively for the results of the three studies reported in this thesis, results that are predicted by the eliminative inference approach. Second, the ontology of attentional theory is already relatively rich. ADIT (Kruschke, 1996), for example, involves three fairly separate mechanisms: (1) A learning process that forms the cue-category associations, (2) a rapid attention-shifting mechanism, and (3) a base-rate bias. EXIT (Kruschke, in press) interprets base-rate bias with a bias node and adds (4) exemplar-based memory to learn the changes in the attentional shifts from one trial to the next, in addition to the mechanisms just mentioned. Also, to account for the data reported in this thesis and the “non-random guessing strategy” observed in the early training trials (Kruschke, 1996, p. 20; see also Kruschke & Bradley, 1995), some mechanisms for (5) eliminative inference would have to be added too.

Third, as noted in Kruschke (in press), the parameter estimates obtained when fitting the seven-parameter model EXIT to data exhibit an amazing variability. For example, the attention shift rate fitted to data from Experiment 1 in Kruschke (1996) is 4.42, but close to zero when fitted to Experiment 2 in the same article (no attention-shifting?). Similarly, the learning rate parameter for exemplar weights is 10.00 when fitted to the data from Experiment 2 in Kruschke (1996), but .0092 when fitted to the data from Experiment 1 in Kruschke (in press), a difference which numerically is three orders of magnitude. Unfortunately, a rationale and interpretation of these variations across what yet appears to be structurally rather similar experimental designs is lacking, which may indicate that the parameters themselves carry little or no information.

Fourth, although there are obvious shortcomings of the quantitative implementation of the eliminative inference mechanism, ELMO, provided in this thesis, the remedies to these problems discussed in the theoretical discussion of Study III seem reasonable. As was discussed, one explanation is that the responses may also be affected by the kind of attention-shifting processes modeled by attentional theory, and a sensible extension of ELMO would be to allow for such attention effects. The idea of attentional shifts has immense support in the literature on animal learning, and it seems reasonable that they contribute to the inverse base-rate effect to some extent. What is more debatable is whether these processes provide the most important accounts of the inverse base-rate effect. Similarly, Experiment 3 revealed that the quantitative fit of ELMO breaks down when the experimental task is changed: In this case by changing the response options. It should be noted, however, that attentional theory cannot even qualitatively account for these data. Nevertheless, the observation of too many eliminative responses suggests that a viable quantitative implementation requires a more complex formalization capturing how and when people turn from induction to elimination (e.g., as a function of the response options). Although this is a research task for the future, such an integrative model would hopefully approximate the underlying process that operates in the inverse base-rate effect task in a faithful manner and produce successful fits to human data.

More generally, a necessary first step toward an integrative account of the inverse base-rate effect would be to isolate the attentional shifts from the eliminative inference mechanism, or vice versa, to get an appreciation of how these processes individually affect the inverse base-rate effect (i.e., an independent estimate of each process). How could this be accomplished? First, as pointed out previously, we need to know what is meant by attentional shifts to allow for direct empirical observation. Arguably, the eliminative inference mechanism is easier to observe, as illustrated in, for example, Experiment 3 in Study I. For example, are attentional shifts to be interpreted as explicit or implicit mechanisms? The arguments of this thesis clearly suggest that attentional shifts are implicit processes. On the other hand, the results by Lubow and Josman (1993) who found that hyperactive 59-90 months old children demonstrated a loss of latent inhibition suggest that attentional shifts are explicit processes, at least with Reber’s criteria in mind: Implicit systems should be more robust than explicit systems, operating despite injuries, diseases and other disorders.

Whatever the exact nature of the attentional shifts, one possibility would be to divert participants’ attention while requiring them to solve the inverse base-rate task. Ideally, this would eliminate the influence of attentional shifts and leave the eliminative inference mechanism unaffected. The inverse base-rate task could be changed to allow for “conflict” between two attention-demanding tasks (and I take the inverse base-rate task to be one such task) while at the same time requiring them to learn the cue-outcome associations. On the other hand, it is reasonable that attention is a necessary element in the formation of rules, too, and the eliminative inference mechanism is mediated by well-formed inference rules, so we might end up with inconclusive evidence after all.

Another possibility would be to capture the attentional shifts in eye-movements (although I am far from certain that eye-movements reflect the attentional shifts implied by attentional theory). An eye-tracking device could be calibrated with a perceptual version of the inverse base-rate task[13], and one could follow participants' eye-movements between the cues and outcomes and thereby determine if attentional shifts have taken place. Rapid attention shifts operate in learning, and this suggests that it should be possible to see whether people's eyes focus more on, for example, *PR* than on *I* when learning which predictors are associated with the rare category *R*, since *I* is already associated with *C* according to attentional theory (cf. Figure 1). In its present formalization, the eliminative inference approach predicts no such difference. It only predicts that the probability of forming common inference rules is higher than the probability of forming rare ones because of the base-rate difference.

A third possibility would be to measure reaction times when participants are presented with the inverse base-rate task. Although neither attentional theory nor the eliminative inference approach are formalized to predict such effects, it seems reasonable to expect that implicit processes are more rapid than explicit meta-cognitive ones. Therefore (provided that attentional shifts are implicit), participants should spend less time when responding to the conflicting predictors *PC.PR* than the imperfect, *I*, and combined predictors, *I.PC.PR*, if the correct explanation lies in rapid attention shifts, whereas they should spend more time if the correct explanation lies in a meta-cognitive strategy such as elimination.

At present, it is in fact questionable whether behavioral data can give us much more information as to which processes are involved in the inverse base-rate effect than various researchers have already provided (Dennis & Kruschke, 1998; Fagot et al., 1998; Kruschke, 1996, in press; Medin & Bettger, 1991; Medin & Edelson, 1988; Shanks, 1992; and Study I, II and III of this thesis). A more fruitful alternative might be to investigate which neurobiological structures are invoked in the inverse base-rate task, preferably by both rule-based and similarity-based learners as indicated by the Shanks and Darby task (1998). As previously mentioned, Smith et al. (1998) found that different brain areas were involved in rule-based and exemplar-based processing. Perhaps these brain areas may be used as predicted regions of interest in such a study, an attempt that also could provide clues about which memory-systems are invoked.

4.3. Concluding Remarks

To conclude, in what ways has this thesis contributed to a deepened understanding of the processes and representations underlying the inverse base-rate effect? First, from a normative point of view, it suggests that the preference of a rare category, *R*, when confronted with the conflicting probe, *PC.PR*, is in fact a fully rational response. The perfect rare predictor, *PR*, in the conflicting transfer probe, *PC.PR*, has never oc-

[13] It would probably be necessary to eliminate the eye-movements arising from reading the symptom- and disease names as is the case in the original inverse base-rate task.

curred with the common disease during learning, *I.PC*$\rightarrow$*C*, and this well-known disease is therefore eliminated. Interestingly, as Anderson (1990) has previously demonstrated, this proposal is also consistent with Bayesian prescriptions of rationality (cf. Section 1.3).

Second, the inverse base-rate effect has been interpreted as incompatible with the exemplar-based models of categorization, which predict a consistent use of base-rates. This otherwise extremely successful area of theorizing would restore its position with the addition of the eliminative inference mechanism. Most of the time people's inferences work in accordance with their accumulated memory traces (i.e., induction), but sometimes when knowledge is scarce, and an inductive inference provides an unreasonable alternative, their inferences work in the opposite manner (i.e., elimination). This proposal also highlights the larger intention of this thesis, namely that the eliminative inference mechanism is more than a "response strategy" in an experimental task that lacks a well-defined (or "correct") solution. Rather, it is proposed that eliminative inference is a general cognitive mechanism that probably works to different degrees in a number of cognitive situations. Although it remains for future research to determine under which particular circumstances such eliminative processes actually take place, they point to the need for eventual integration with general theories of learning and categorization.

Third, the inverse base-rate effect has been explained by cue-competition, a concept derived from associative theories of learning (e.g., Rescorla & Wagner, 1972), and most successfully formulated in the rapid attentional shifts of attentional theory (Kruschke, 1996, in press). As we have seen, however, the empirical data of this thesis challenge that position. Rather than merely being an outcome of competition between feature-based symptom configurations, the inverse base-rate effect seems to be well captured by rule-based elimination.

5. References

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review.* (In press).

Ashby, F. G., Alfonso-Reese, L. A, Turken, A. U., & Waldron, E. M. (1998). A formal neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442-481.

Atkinson, R. C., & Estes, W. K. (1963). Stimulus sampling theory. In R. D Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology.* New York: Wiley.

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

Brooks, L. (1978). Nonanalytic concept formation and memory for instance. In E. Rosch & B. B. Lloyd (Eds.). *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Erlbaum.

Brooks, L. (1987). Decentralized control of categorization: The role of prior episodes. In U. Neisser (Ed.). *Concepts and conceptual development: The ecological and* *intellectual factors in categorization* (pp. 141-147). Cambridge, England: Cambridge University Press.

Butt, J. (1988). Frequency judgments in an auditing-related task. *Journal of Accounting Research, 26*, 315-330.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 537-545.

Christensen-Szalanski, J. J., & Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance, 29*, 270-278.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1-73.

Dennis, S., & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology*, *50***,** 131-138.

Ellis, R., & Humphreys, G. (1999). *Connectionist Psychology: A Text with Readings*. Hove, United Kingdom: Psychology Press.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107-140.

Estes, W. K. (1991). Cognitive architectures from the standpoint of an experimental psychologist. *Annual Review of Psychology, 42,* 1-28.

Estes, W. K. (1994). *Classification and cognition.* London: Oxford University Press.

Estes, W. K., Campbell, J. A., Hatsopoulus, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of a parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 556-571.

Fagot, J., Kruschke, J. K., Depy, D., & Vauclair, J. (1998). Associative learning in humans (Homo sapiens) and baboons (Papio Papio): Species differences in learned attention to visual features. *Animal Cognition*, *1*, 123-133.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3-71.

Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective Probability* (pp. 130-161). Great Britain: John Wiley & Sons Ltd.

Gigerenzer, G. & Hoffrage (1995). How to improve Bayesian reasoning without instructions: Frequency formats. *Psychological Review, 102,* 684-704.

Gluck, M. A. (1992). Stimulus sampling and distributed representations in adaptive network theories of learning. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes,* (Vol. 1, pp. 169-199). Hillsdale, NJ: Lawrence Erlbaum.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

Hasher, L. & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39,* 1372-1388.

Holyoak, K. J., & Spellman, B. A. (1993). Thinking. *Annual Review of Psychology, 44*, 265-315.

Johansen, M. K., & Palmeri, T. J. (2001). *Representational shifts in category learning.* (Submitted).

Juslin, P., Wennerholm, P., & Winman, A. (1999). Mirroring the inverse base-rate effect: The novel symptom phenomenon. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (pp. 252-257). Mahwah, NJ: Erlbaum.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgments under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

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

Kalnins, I. V., & Bruner, J. S. (1973). The coordination of visual observation and instrumental behavior in early infancy. *Perception, 2*, 307-314.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 276-296). New York: Appleton-Century-Crofts.

Koehler, J. (1996). The base-rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*, 1-53.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Kruschke, J. K. (1996). Base-rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 3-26.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*. (In press).

Kruschke, J. K. The inverse base-rate effect is not explained by eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. (In press).

Kruschke, J. K., & Bradley, A. L. (1995). *Extensions to the delta rule for human associative learning*. (Indiana University Cognitive Science Research Report #141. Avaliable via WWW at http://www.indiana.edu˜kruschke/kb95abstract.html).

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1083-1119.

Lachnit, H., & Kimmel, H. D. (1993). Positive and negative patterning in human classical skin conductance response conditioning. *Animal Learning & Behaviour, 21*, 314-326.

Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review, 107,* 227-260.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492-527.

Lubow, R. E. (1989). *Latent inhibition and conditioned attention theory.* New York: Cambridge University Press.

Lubow, R. E., & Josman, Z. E. (1993). Latent inhibition deficits in hyperactive children. *Journal of Child Psychiatry and Psychology, 34,* 959-973.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement . *Psychological Review, 82,* 276-298.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck, M. A., & Bower, G. H. (1988). *Journal of Experimental Psychology: General, 118,* 417-421.

McGeorge, P., Crawford, J. R., & Kelly, S. W. (1997). The relationships between psychometric intelligence and learning of an explicit and an implicit task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 239-245.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159-188.

Medin D. L., & Bettger, L.G. (1991). Sensitivity to changes in base-rate information. *American Journal of Psychology, 104,* 311-332.

Medin, D. L., & Edelson, S.M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General, 117*, 68-85.

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (vol. 5, pp. 189-223). San Diego, CA: Academic Press.

Medin, D. L., & Schaffer, M. M. (1978). Context model of classification learning. *Psychological Review, 85,* 207-238.

Medin, D. L. & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 241-253.

Moshman, B. (1998). Cognitive development beyond childhood. In W. Damon, D. Kuhn, & R., Siegler (Eds.). *Handbook in child psychology* (Vol. 2, pp. 947-978). New York: John Wiley & Sons, Inc.

Nelson, T. E., Biernat, M., & Manis, M. (1980). Everyday base-rates (sex stereotypes): Potent and resilient. *Journal of Personality and Social Psychology, 59,* 664-675.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental proceses. *Psychological Review, 84,* 231-259.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 87-108.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7,* 375-402.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 211-233.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 324-354.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94*, 61-73.

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

Pesenti, M., Zago,L., Crivello, F., Mellet, E., Samson, D., Duroux, B., Seron, X., Mazoyer, B., & Tzourio-Mazoyer, N. (2001). Mental calculation in a prodigy is sustained by right prefrontal and medial temporal areas. *Nature Neuroscience, 4,* 103-107.

Raven, J., Raven, J. C., & Court, J. H. (1993). *Advanced Progressive Matrices: Section 1.* Oxford Psychologists' Press.

Reber, A. (1993). *Implicit learning and tacit knowledge.* New York: Oxford University Press.

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

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21-59). Cambridge, England: Cambridge University Press.

Roediger, H. L., & McDermott, K. B. (1993). Implicit memory in normal subjects. In F. Boller and J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63-131). Amsterdam: Elsevier.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 629-639.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Vol. 1. Foundations* (pp. 318-362). Cambridge, MA: MIT Press.

Schacter, D. L. (1990). Perceptual representation systems and implicit memory: toward a resolutation of the multiple memory systems debate. In *Development and neural bases of higher cognition,* (Ed. A. Diamond), pp. 543-571. New York Academy of Sciences, New York.

Schacter, D. L., & Tulving, E. (1994). What are the memory systems of 1994? In D. L. Schacter and E. Tulving (Eds.), *Memory systems of 1994* (pp. 1-38). MIT Press, Cambridge, MA.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science, 4,* 3-18.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24*, 405-415.

Shanks, D. R., & St John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17,* 367-447.

Simon, H. A. (1996). Computational theories of cognition. In W. O'Donohue & R. E. Kitchener (Eds.), *The philosophy of psychology* (pp. 160-172). London: Sage Publications.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3-22.

Smith, L. B., & DeLoache, J. S., & Schreiber, J. C. (1995). *Induction of a sameness rule by young children.* Unpublished raw data.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 3-27.

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science, 16,* 1-40.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65,* 167-196.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition, 22,* 167-196.

Vandierendonck, A. (1995). A parallell rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review, 2,* 442-459.

van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*, 127-151.

Wagner, A. R. (1978). Expectancies and priming in STM. In S. H. Hulse, H. Fowler, & W. K. Honig (Eds.), *Cognitive processes in behavior* (pp. 177-209). Hillsdale, N. J: Erlbaum.

Wennerholm, P. (2001). *Probabilistic challenges to the inverse base-rate effect.* (Manuscript in preparation).