Journal of Experimental Psychology:
Learning, Memory, and Cognition
2003, Vol. 29, No. 6, 1396–1400

0278-7393/03/$12.00 DOI: 10.1037/0278-7393.29.6.1396

Attentional Theory Is a Viable Explanation of the Inverse Base Rate Effect: A Reply to Winman, Wennerholm, and Juslin (2003)

John K. Kruschke
Indiana University

A. Winman, P. Wennerholm, and P. Juslin (2003) have admitted that J. K. Kruschke (2001a) cogently demonstrated the shortcomings of eliminative inference as an explanation of the inverse base rate effect, but they raise criticisms of Kruschke's attentionally based explanation. First, Winman et al. pointed out that attentional shifting does not improve learning performance in Kruschke's (1996) ADIT model, contrary to the claims that attentional shifting accelerates learning. This reply demonstrates that the deceleration of learning is a natural consequence when attentional shifts are not learned, as is the case in ADIT; however, when attentional shifts are learned, as was assumed by the underlying theory and as is the case in the EXIT model (Kruschke, 2001a, 2001b), then performance is indeed accelerated by attentional shifts. Second, Winman et al. pointed out that, whereas EXIT captures essentially all of the notable effects in the transfer data, it fails to capture a small effect [viz., $p(C|PC) > p(R|PR)$]. This reply demonstrates that when this trend in the data is merely weighted more heavily in the model fitting, then the EXIT model accommodates it. EXIT accomplishes this by emphasizing base rate learning more strongly. Thus, the EXIT model, and attentional theory more generally, remains a viable explanation of the inverse base rate effect.

The inverse base rate effect refers to a perplexing behavior, first reported by Medin and Edelson (1988), in which people make choices contrary to the base rates when generalizing from previous training examples. In a typical procedure, participants are trained to diagnose symptom lists when prompted with a list of possible (fictitious) diseases. On some trials, the learner is shown a list that includes a symptom, abstractly denoted here as I, and a second symptom, here denoted as PC. Whenever these two symptoms co-occur, the correct diagnosis is the disease denoted abstractly here as C. The actual symptoms seen by the learner could be words such as "dizziness" and "headache." Cases of this type are denoted I.PC $\rightarrow$ C, in which the period between symptoms denotes co-occurrence. On relatively fewer trials, there are cases of I.PR $\rightarrow$ R. Thus, symptom I is an imperfect predictor of the correct diagnoses, symptom PC is a perfect predictor of the common diagnosis C, and symptom PR is a perfect predictor of the rare diagnosis R.

After learning those cases, people are presented with new combinations of symptoms and asked to provide their best diagnosis (without being provided corrective feedback). In particular, when presented with symptom I by itself, people tend to diagnose it as C, the common disease. But when presented with symptom pair PC.PR, people tend to diagnose it as the rare disease R, contrary to the base rates. A variety of other symptom combinations are presented in the transfer phase, and it has proven to be a difficult challenge for theories of learning and performance to account accurately for the full pattern of results.

This research was supported by National Science Foundation Grant BCS 99-10720.

Correspondence concerning this article should be addressed to John K. Kruschke, Department of Psychology, Indiana University, 1101 E. 10th Street, Bloomington, Indiana 47401. E-mail: kruschke@indiana.edu

Attentional Shifting and Learning Accelerates Accuracy

Attentional theory (e.g., Kruschke, 1996, 2001a, 2001b) suggests that people learn first about I.PC $\rightarrow$ C because it occurs so frequently. When learning I.PC $\rightarrow$ C, people attend (on average) to both symptoms and build moderate strength associations from both symptoms to C. When subsequently learning I.PR $\rightarrow$ R, however, people shift attention away from symptom I to symptom PR. This shift of attention reduces error: Because symptom I is already associated with disease C, and C is not the correct response, attention is directed away from symptom I and toward symptom PR. This shift of attention protects and preserves the previously learned (and still useful) knowledge of I.PC $\rightarrow$ C.

The shift of attention should also accelerate learning on cases of I.PR $\rightarrow$ R, because there is reduced interference from symptom I. Indeed, by definition, the shift of attention is executed specifically to reduce error. Winman, Wennerholm, and Juslin (2003) pointed out, however, that in the ADIT model (Kruschke, 1996), which implements the idea of attention shifting, performance on the training cases of I.PR actually decelerates, rather than accelerates, when attention shifting increases. Winman et al. (2003) call for clarification of this apparent contradiction.

The key to resolving this issue is that the ADIT model does not learn its shifts of attention. In other words, although ADIT shifts attention to reduce error on individual trials, it does not retain those shifts on subsequent trials. Consider a trial of I.PR $\rightarrow$ PR. At the beginning of the trial, the model distributes attention equally to both I and PR. The previously learned association from I to C produces an error (because the correct response on this trial is not C). The model reduces this error by shifting attention away from I and toward PR. The model then increases its associative strength from the attended-to symptom, PR, to C. The association from the unattended-to symptom, I, is relatively unchanged. The shift of attention away from I toward PR is not learned, however. There-

fore, at the beginning of the next trial of I.PR → R, symptom I is again attended to, which produces an error again because the association from I to C has remained relatively intact. The revival of attention to I resurrects the error that was previously destroyed by the reduced attention to I. The resurrected error produces the decelerated learning in ADIT.

The lack of attention learning in ADIT was merely a parsimonious simplification for its particular applications (Kruschke, 1996). The general theoretical approach, however, has always assumed that attention shifts are learned (as emphasized, e.g., in Kruschke, 2001b, p. 815). The description of the ADIT model in Kruschke (1996) emphasized the lack of learning in ADIT merely to reduce possible confusion with the previous ALCOVE (attentional learning covering map) model by the same author (Kruschke, 1992); there was never any claim that lack of attentional learning was important for generating the behavioral effects. Indeed, the EXIT model (Kruschke, 2001a, 2001b) provides an extension of ADIT that implements learning of attention shifts.

EXIT has separate parameters for the attentional shifting rate and the attentional learning rate.[1] The shifting rate governs the extent to which attention can be shifted across cues when a stimulus is presented and corrective feedback is provided. The learning rate governs the extent to which the shifted distribution of attention is remembered across trials, such that when the same stimulus appears again, the shifted distribution is re-evoked, even without corrective feedback. For example, consider the first case of I.PR → R after cases of I.PC → C have been trained. The model will partially activate outcome C because of the previously learned association from I to C. To reduce this erroneous response, attention shifts away from I toward PR. The extent of this shift is governed by the shifting rate parameter. Note that this shift of attention is generated by the predictive error on the current trial, and the shifted distribution is not retained at the onset of the next trial. It would be useful, however, to regenerate this shifted distribution of attention the next time this input pattern re-occurs. This is accomplished in EXIT by learning associations between input exemplars and attentional gains. The extent of learning is governed by the attentional learning rate parameter.

When EXIT's attentional learning rate is fixed at zero, its processing is essentially the same as in ADIT, and in this case it also behaves quantitatively much like ADIT; that is, attentional shifting can decelerate learning performance. When EXIT's attentional learning rate is increased to moderate values, however, it does show the accelerated performance claimed by the high-level attentional theory.

Figure 1 displays the behavior of EXIT when it is applied to Experiment 1 of Kruschke (1996). This experiment was summarized in Kruschke (2001a), where both the eliminative inference model (ELMO; Juslin, Wennerholm, & Winman, 2001) and the EXIT model were fit to the data. The top row of Figure 1 shows EXIT's behavior when there is no attentional shifting (and no learning of attentional shifts). The top left panel shows accuracy as a function of training block on the I.PC → C trials and the I.PR → R trials. These dotted and dashed curves form the baseline learning curves relative to which acceleration or deceleration of performance must be judged. The top right panel shows predicted choice percentages on selected items in the transfer phase, where it can be seen that the model produces no inverse base rate effect when there is no attention shifting.

The middle row of Figure 1 shows EXIT's behavior when there is attentional shifting but no attentional learning. All other parameter values remained fixed at the same values as used for the top row. The middle left graph shows that accuracy on I.PR → R rises more slowly than the dashed reference curve. This is analogous to Winman et al.'s (2003) report regarding ADIT's deceleration of learning, because EXIT without attention learning is much like ADIT. The middle right graph shows that EXIT robustly produces an inverse base rate effect that closely matches human performance, even without attentional learning. That is, the inverse base rate effect can be largely captured by the asymmetries produced by the attentional shift alone, without the additional influence of attentional learning. Attentional learning is important for capturing other effects, such as retarded learning about a blocked cue (see Kruschke, 2001b).

The bottom row of Figure 1 shows EXIT's behavior when there is learning of the attention shifts. The bottom right graph shows that the inverse base rate effect is only slightly influenced by attentional learning. The bottom left graph shows that attentional learning does indeed accelerate performance on the learning trials: The learning curve for I.PR → R is now much higher than the dashed reference curve. This result confirms the claims in previous articles that attentional shifting accelerates learning, but the result also clarifies the claims, which presume that the shifts of attention have been learned.

The behavior of EXIT illustrated in Figure 1 is not closely tied to particular choices of parameter values. Attentional shifting robustly produces an inverse base rate effect in the model, regardless of whether the shifts are learned. Learning of attentional shifts robustly accelerates performance over baseline, regardless of the exact parameter values.

In previous studies, I have suggested that seemingly irrational behaviors such as the inverse base rate effect can be explained as unfortunate side effects of attentional shifting but that attention shifting accelerates learning and is therefore a rational solution to the need for speed in learning. The claim that attention shifting accelerates performance was made in the context of models that learn the attention shifts (e.g., the RASHNL model of Kruschke & Johansen, 1999, and the EXIT model of Kruschke, 2001b). What has been clarified in the present exchange is that attentional learning, not just shifting, is essential for that claim to be true.

EXIT Can Accommodate $p(C|PC) > p(R|PR)$

A second criticism raised by Winman et al. (2003) is that EXIT, in its best *overall* fits to data, has $p(R|PR)$ somewhat greater than $p(C|PC)$, whereas this is sometimes slightly reversed in human performance. Winman et al. argue that this is a significant shortcoming of the attentional approach because it predicts that the associative strength from PR to R should be greater than the associative strength from PC to C, but this is apparently contradicted when $p(C|PC) > p(R|PR)$.

Winman et al.'s (2003) argument does not hold, however, because there are influences on overt response proportions beyond

[1] The attention shifting rate was denoted λ_g and referred to as the "shift rate for attention" (Kruschke, 2001a, 2001b). The attentional learning rate was denoted λ_x and referred to as the "learning rate for associative weights from the exemplar nodes to the gain nodes" (Kruschke, 2001a, 2001b).

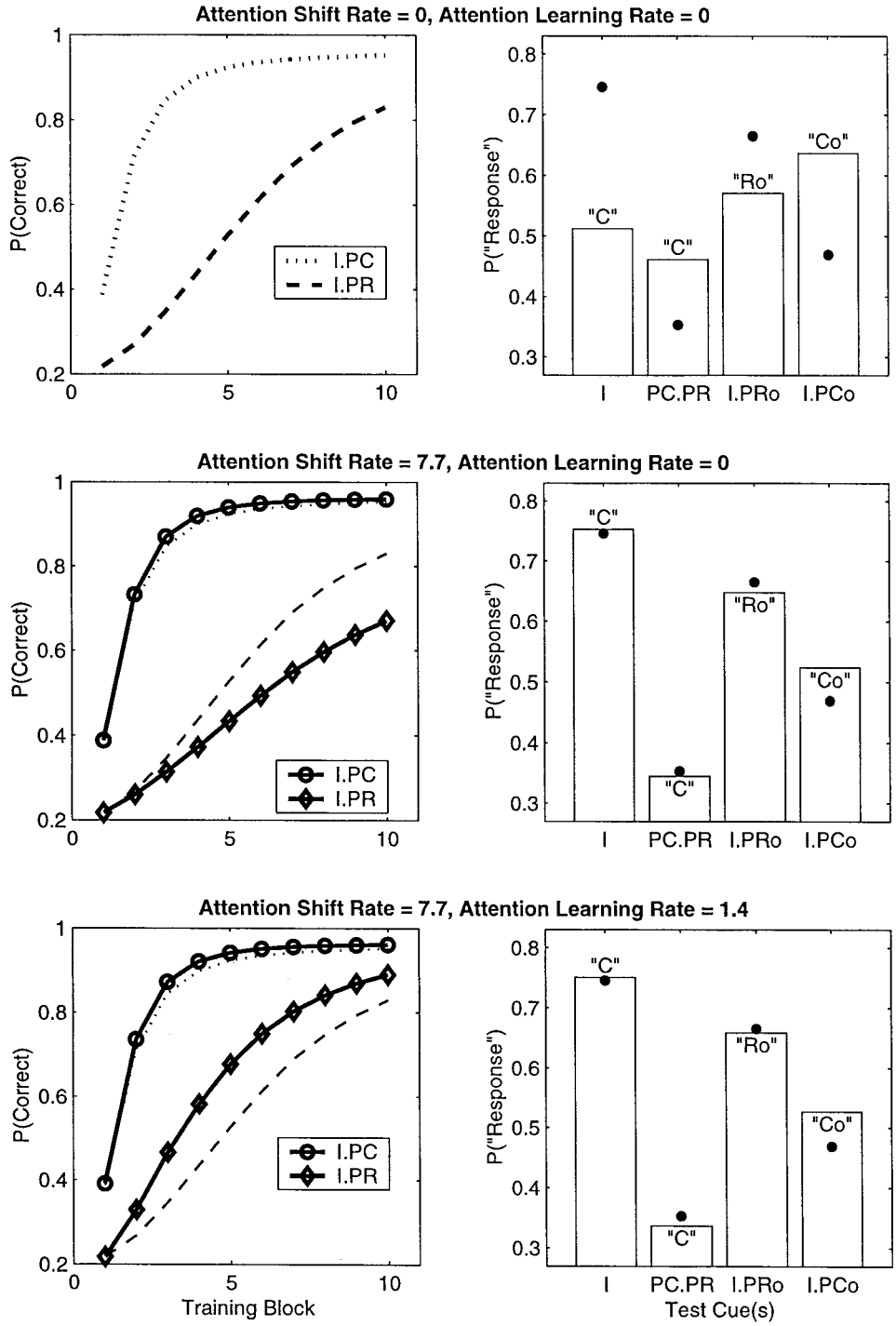

Figure 1. Behavior of the EXIT model (Kruschke, 2001a, 2001b) applied to Experiment 1 of Kruschke (1996). Top row has zero attention shifting, and consequently zero attention learning. Middle row has attention shifting but zero attention learning. Bottom row has attention shifting and attention learning. All other parameter values were held fixed across the rows, with exemplar specificity = 2.5, attention capacity = 2.6, choice decisiveness = 4.4, output association learning rate = 0.14, and bias salience = 0.010. These parameter values were chosen because they best fit (RMSD = 2.54 percentage points) the data from Experiment 1 of Kruschke (1996) using 40 simulated random subjects. The parameter values are somewhat different than those reported in Kruschke (2001a) because that fit used the particular training sequences observed by the participants. The behavior of the model is essentially the same. The left column of graphs shows accuracy as a function of training block, with separate curves for training cases I.PC and I.PR. The learning curves for zero attention shifting, indicated by the dotted and dashed curves in the top left graph, are copied into the middle-left and bottom-left graphs, for sake of easy comparison. The right column of graphs shows selected choice proportions for the subsequent test phase. Bars show model predictions; dots show human data. Symptom combination I.PCo denotes an imperfect symptom I from one pair of diseases combined with a perfect predictor, PCo, of the common disease in another pair of diseases. Notice in the human data that p(Ro|I.PRo) $>$ p(Co|I.PCo).

Table 1
Choice Percentages for Humans, ELMO, and EXIT, From Experiment 1 of Kruschke (1996)

	Human choice				ELMO choice				EXIT choice			
Symptom	C	R	Co	Ro	C	R	Co	Ro	C	R	Co	Ro
I	**74.6**	**17.4**	4.9	3.1	**64.2**	**35.6**	0.0	0.2	**77.0**	**11.3**	7.7	3.9
I.PC.PR	**58.0**	**40.2**	1.3	0.4	**64.2**	**35.6**	0.0	0.2	**58.8**	**39.4**	1.0	0.8
I.PCo	40.6	8.0	**46.9**	4.5	23.0	16.3	**60.3**	0.3	33.4	8.7	**56.1**	1.8
I.PRo	21.9	8.5	3.1	**66.5**	26.4	25.9	0.2	**47.6**	30.2	3.8	1.3	**64.6**
PC	*93.3*	3.1	3.1	0.4	*98.7*	0.6	0.1	0.6	*89.8*	1.8	5.4	2.9
PR	4.0	*91.1*	1.8	3.1	0.6	*93.5*	0.6	5.3	3.0	*87.2*	7.8	2.0
PC.PR	35.3	61.2	2.2	1.3	35.1	58.5	0.7	5.7	36.3	56.6	4.7	2.4
PC.PRo	35.3	2.7	5.8	56.3	35.1	5.7	0.7	58.5	41.1	0.7	1.1	57.0
I.PC.PRo	71.9	3.6	3.6	21.0	72.3	8.0	0.0	19.6	76.9	0.6	0.3	22.2

Note. Boldface numbers indicate data that were particularly challenging for ELMO. Italicized numbers indicate data that were focused on by Winman et al. (2003). C = common disease; R = rare disease. An "o" after a symptom label indicates that the symptom came from the other pair of diseases (e.g., I.PCo indicates I1.PC2 and I2.PC1, averaged). A dot between symptoms indicates co-occurrence. PC = perfectly predictive symptom of a common disease; PR = perfectly predictive symptom of a rare disease; I = imperfectly predictive symptom.

the associations from PR and PC. Among these influences are associations from a context (i.e., bias) cue that effectively represent base rates of the response items. These base rate associations can tip the scales slightly in favor of the more frequent outcome when they are made more salient. As is shown in detail below, when EXIT is refit to the data with the difference between p(C|PC) and p(R|PR) weighted heavily, the effect is captured and the best fitting parameter values give the context cue a notable salience. (Details of how EXIT and ADIT learn base rates can be found in Appendix 2 of Kruschke, 2001b.)

Table 1 shows the results of fitting EXIT to the data from Experiment 1 of Kruschke (1996), weighting the difference between p(C|PC) and p(R|PR) heavily in the fit. Previously the model was fit using a standard measure of discrepancy between data and prediction, namely the root mean squared deviation, (RMSD) = $\sqrt{1/36 \sum_{i=1}^{36} (p_i - \hat{p}_i)^2}$, where p_i is the observed human choice percentage and $\hat{p}_i$ is the model's predicted percentage and the sum spans all 36 cells in Table 1. In the new fit, the RMSD was added with another cost, $C = 3 \times sig[p(R|PR) - p(C|PC)]$, where $sig(x) = 1/[1 + \exp(-x)]$ is the sigmoid function. This was merely an arbitrary cost that approaches a value of 3 when p(R|PR) greatly exceeds p(C|PC), and approaches a value of zero when p(R|PR) is much less than p(C|PC). The fitting routine tried to minimize 0.33*RMSD + 0.67*C.

The best fit yielded RMSD = 3.76 (i.e., 3.76 percentage points discrepancy, on average) and a cost $C = 0.205$, compared with RMSD = 7.40 for ELMO.[2] Table 1 shows that EXIT yields p(C|PC) > p(R|PR) by 2.6 percentage points, slightly more than in the human data. The difference is robust in the model, and not merely the result of random sampling from simulated subjects. The model predictions come from an average of 40 random simulated subjects, and the advantage of p(C|PC) over p(R|PR) is highly reliable across other random number seeds.

The reweighted fit sets the salience of the context node to 0.938, whereas in the original overall fit the salience of the context node was only 0.010. The reweighting emphasizes base rate consistency in the data, which is achieved in the model by learning associations from the bias node. This detracts from the magnitude of the inverse base rate effect, however, and the fit of EXIT to the remaining data is definitely worse than before (the best RMSD is 2.54 percentage points when the cost C is given zero weight, with parameter values in Figure 1). Nevertheless, EXIT continues to show all the important effects in the data, unlike ELMO. In particular, EXIT robustly shows the base rate inversion, p(Co|I.PCo) < p(Ro|I.PRo), whereas ELMO predicts the opposite.

Performance Is Not Only Attention Shifting

The viability of attention shifting as an explanation of the inverse base rate effect does not in any way exclude the action of other mechanisms in human behavior. The mind is a busy place, and doubtlessly there are many learning and response processes involved in generating performance in these tasks. Eliminative inference is likely to play a strong role in situations when learners have gaps in their knowledge and therefore engage inference strategies during responding. The claim here is merely that among these numerous processes, attentional shifting plays a prominent role in the inverse base rate effect.

[2] Best fitting parameter values for EXIT with the modified cost function were as follows: exemplar specificity = 9.62, attention capacity = 17.0, choice decisiveness = 5.00, attention shifting rate = 2.65, output association learning rate = 0.135, attention learning rate = 0.0313, and bias salience = 0.938. See Kruschke (2001a, 2001b) for an explanation of the parameters in EXIT.

References

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, and Cognition, 27,* 849–871.

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. (2001a). The inverse base-rate effect is not explained by

eliminative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1385–1400.

Kruschke, J. K. (2001b). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology, 45,* 812–863.

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

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.

Winman, A., Wennerholm, P., & Juslin, P. (2003). Can attentional theory explain the inverse base rate effect? Comment on Kruschke (2001). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1390–1395.

Received March 25, 2002
Revision received April 14, 2003
Accepted April 16, 2003 ■