An Activation-Based Account of Belief Bias in Relational Reasoning: The Effect of Concurrent Working Memory Load

Adrian P. Banks (a.banks@surrey.ac.uk)
Department of Psychology, University of Surrey
Guildford, Surrey, GU2 7XH, UK

Abstract

Working memory plays an important role in explaining reasoning performance. A model of belief bias in relational reasoning is presented here which explains the typical belief bias effects through the activation level of conclusions in memory. This account is formalized in an ACT-R model. The model predicts that a concurrent working memory load will increase the effect of belief on responses to reasoning problems, specifically on indeterminately invalid problems. This prediction is confirmed experimentally.

Keywords: Relational reasoning; Belief bias; ACT-R

Introduction

Prior knowledge has frequently been shown to influence deductive reasoning, an effect known as belief bias. Current explanations of this phenomenon are typically dual process theories, suggesting that belief bias arises from an interaction of rapid, automatic, heuristic processes with slow, controlled, analytic processes (e.g. Evans, 2006; Stanovich, 2004). The analytic processes rely on limited working memory resources and so explaining the role of working memory in reasoning is critical to understanding belief bias. However, descriptions of working memory are varied. The concept has been used to describe the inhibition of heuristic responses (De Neys, 2006) and as a limited capacity store for the representations used in reasoning (Quayle & Ball, 2000) and serving both of these functions (Evans, 2008). The purpose of this paper is to further understanding of the role of working memory by developing and testing a well-specified account of its role in reasoning. This account is derived from ACT-R theory (Anderson, 2007). Specifically it suggests that the activation level of the representations involved in reasoning affects their chance of retrieval from working memory and therefore the influence they have on judgments of validity. This activation-based account explains belief bias effects and makes a novel prediction about the influence of concurrent working memory load on belief bias which is tested here.

Belief Bias in Relational Reasoning

Relational reasoning problems can be constructed such that the conclusions vary independently in terms of belief and logic. Here is an example from Roberts and Sykes (2003):

The Rock 'n' Roll era was before the Punk music era
Grunge music was popular before the Punk music era
'The Silence of the Lambs' was released during the Punk music era
'Jailhouse Rock' was released during the Rock 'n' Roll era
Therefore, 'The Silence of the Lambs' was released after 'Jailhouse Rock'

Participants are asked to accept only conclusions that necessarily follow from the premises, i.e. logically valid conclusions. The conclusion here is both believable and logical and is commonly accepted. Problems where the conclusion necessarily follows, that is it is consistent with all possible interpretations of the premises, are referred to as *determinately valid*. If the conclusion is consistent with no possible interpretation of the premises it is *determinately invalid*. If the conclusion is consistent with some but not all interpretations, it is possible but not necessary and so is *indeterminately invalid*. Typically it is found that (a) valid conclusions are accepted more than invalid conclusions; (b) believable conclusions are accepted more than unbelievable conclusions; and (c) the effects of belief are stronger on indeterminately invalid conclusions than determinately invalid or valid conclusions (e.g. Roberts & Sykes, 2003). A model of belief bias in relational reasoning must account for these three phenomena.

ACT-R Model of Belief Bias

There are two stages to the model of belief bias presented here. Firstly, in the construction stage of the process, mental models are constructed based on the premises and conclusions are drawn from these mental models, referred to here as 'initial conclusions'. Secondly, in the retrieval stage of the process, these initial conclusions are retrieved and compared with the conclusion presented in the problem. If they all match then the problem is evaluated as valid. If one or more do not match then it is evaluated as invalid. If no conclusions are retrieved (i.e. they have all been forgotten) then the response is a guess. It is in this retrieval stage that the activation level of the conclusions determines the likelihood of their retrieval and it is through this process that belief exerts its influence. The construction and retrieval stages are described in detail below.

Construction Stage

Most theories of relational reasoning suggest that people construct a representation which integrates the information in the premises into mental models (e.g. Schaeken, Johnson-

Laird, & d'Ydewalle, 1996). Each mental model represents a layout consistent with the premises. There is some debate about the exact form of these mental models however (e.g. Vandierendonck, Dierckx & De Vooght, 2004). The model presented here assumes that reasoners initially attempt to construct a single mental model which integrates all of the information in the premises. This model is similar to the 'isomeric' mental models described by Schaeken, Van der Henst and Schroyens (2006). These mental models retain the ambiguous relationship between some of the elements in the premises. For example:

A happens before B
C happens before B
D happens at the same time as B
E happens at the same time as C
Therefore, E happens before D

In this case it is not certain whether A occurs before C and E or afterwards, both are possible. A mental model is therefore constructed in which A, C, and E are represented as occurring before B and D. For example:

A	B
C	D
E	
Prior to $Time_2$	$Time_2$

Time is represented spatially, running from left to right. An initial conclusion can be drawn from this model by focusing on the two elements mentioned in the conclusion (E & D) and retaining their relative position in a new representation. This leads to the initial conclusion that E happens before D.

Sometimes, the problem can refer to elements that are ambiguously represented in the model. For example:

A happens before B
C happens before B
D happens at the same time as A
E happens at the same time as C
Therefore, E happens before D

This leads to the following model:

A	
D	B
C	
E	
Prior to $Time_2$	$Time_2$

Focusing on the two elements in the conclusion, D and E, does not lead to an unambiguous new representation. This ambiguity triggers further mental models to be constructed in which the ambiguity is resolved by explicitly considering all the possible layouts:

C	A	B	A	C	B
E	D		D	E	
$Time_1$	$Time_2$	$Time_3$	$Time_1$	$Time_2$	$Time_3$

A conclusion can now be drawn from both of these, leading to the initial conclusions 'E before D' and 'D before E'.

Retrieval Stage

ACT-R does not have separate modules for working memory and long term memory. Instead, working memory is 'equated with the portion of declarative memory above a threshold of activation.' (Anderson, Reder & Lebiere, 1996, pp. 221-222). In this model the most influential factor in determining activation is Base level activation. A chunk (symbolic representation) in declarative memory begins with a given level of activation which decays over time but is raised again whenever the chunk is retrieved or encountered again. Hence rehearsing a chunk will maintain its higher activation level through repeated retrievals, and if the chunk is experienced again in the course of reasoning the new chunk will merge with the existing chunk to create a merged chunk of higher activation.

The believability of a problem is modelled by placing a chunk in declarative memory before reasoning begins. This chunk takes the same form as a conclusion and represents the relationship between the two elements in the conclusion. It is the prior belief about the relationship between those two elements. If the conclusion presented in the problem describes the same relationship as this memory then it is a believable conclusion, else it is unbelievable.

The influence of this prior belief is twofold. Firstly, it can be unintentionally retrieved. In the retrieval stage of the model the intended goal is to retrieve the initial conclusions that have been drawn from the mental models to evaluate the problem. However the prior belief takes a similar form and may be recalled and used in this evaluation process instead or as well as the initial conclusions. In other words, the conclusion to the problem is familiar and is judged as valid but the source of this memory has been wrongly attributed to an initial conclusion that has just been drawn. In fact it stems from the matching of the conclusion in the problem to the prior memory chunk.

Secondly, the initial conclusions drawn from the mental models can merge with the prior belief chunk if they describe the same relationship. The initial conclusions formed for each problem type as a result of this process are presented in Table 1. Prior belief initially has the lowest activation level, less than a newly created conclusion, but if the conclusion merges with prior belief the activation of this chunk will be the highest.

The believable valid and unbelievable determinately invalid conditions lead to initial conclusions which match prior belief, and so form highly active 'merged conclusions'. These are very likely to be retrieved. In the believable valid condition this chunk matches the

Table 1: Merging of conclusions and prior belief chunks in each condition

	Valid		Determinately Invalid		Indeterminately Invalid	
	Matching	Non-matching	Matching	Non-matching	Matching	Non-matching
Believable	PB + C	-	PB	C	PB + C	C
Unbelievable	C	PB	-	PB + C	C	PB + C

Matching = chunk matches the conclusion in the problem, Non-matching = chunk does not match the conclusion in the problem, PB = prior belief chunk, C = conclusion chunk, PB + C = prior belief merged with conclusion, '-' = no chunks. Relative activation levels: PB+C > C > PB.

conclusion in the problem and so it will be judged as valid whereas in the unbelievable determinately invalid condition it does not match and will be judged invalid.

The unbelievable valid and believable determinately invalid conditions lead to initial conclusions which do not merge with prior belief. They are still active and likely to be retrieved, but not as likely as the previous two conditions. Therefore these conclusions are less likely to be judged as valid because the initial conclusion is more likely to be forgotten. Also, the weaker prior belief chunk may occasionally be retrieved. This does not match the conclusion in the unbelievable valid problems and does match the conclusion in the believable determinately invalid problems and so the logically incorrect response will sometimes be made. (This is the situation in which the prior belief is retrieved in unintentionally).

Finally, the indeterminate problems differ from the determinate problems because two initial conclusions will be drawn, as described above. One of these will always match the conclusion in the problem and one will not, and so there will be one merged conclusion and one unmerged conclusion. In the believable indeterminately invalid condition the merged chunk will match the conclusion in the problem whereas in the unbelievable indeterminately invalid condition the merged conclusion will not match the conclusion. As the merged chunk is most likely to be retrieved, the conclusion presented in the problem is more likely to be supported in the believable condition and rejected in the unbelievable condition. Hence merging explains the interaction of belief and logic that exists with indeterminately invalid problems but not in the other conditions.

In summary, this model outlines a reasoning process in which mental models are created, conclusions drawn and then retrieved from working memory to evaluate the problem. This explains the three main effects found in belief bias experiments. The main effect of logic arises from the mostly successful retrieval of initial conclusions drawn from mental models. The main effect of belief arises from the increased likelihood of retrieving believable conclusions because of their higher activation. The interaction of belief and logic arises only in indeterminate problems because here both matching and non-matching initial conclusions are retained in working memory and the merging process means that the matching conclusion will be more active in the believable conclusions and the non-matching conclusion more active in the unbelievable conclusion.

The Experiment

An experiment was conducted to test how well this model fits the pattern of responses found with relational reasoning problems and the effects of belief on them. Valid, determinately invalid, and indeterminately invalid problems were tested, each with both believable and unbelievable conclusions.

A novel prediction about a direct manipulation of working memory was also tested. This model proposes that conclusions are retrieved from working memory to evaluate reasoning problems. What effect would a concurrent working memory load have on this process? De Neys (2006) has shown a reduction in the accuracy of reasoning in syllogisms when belief and logic conflict, but did not break down problems any further than conflict and no-conflict, for example he does not distinguish between believable valid and unbelievable invalid problems which both conflict. Nor does De Neys test belief-logic conflict in determinately invalid problems, only indeterminately invalid problems were used. This model predicts different effects for each of these problem types, and so extends these earlier findings.

Participants were presented with a random five digit number before they attempted to evaluate each problem. They were asked to retain it and recall the number after they had evaluated the problem. This means that a representation of the number must be maintained in working memory at the same time as the conclusions to the reasoning problem. Rehearsal of the number takes time and slows the processing of the reasoning problem which in turn means that the activation levels of the initial conclusions in memory are not maintained as effectively. Therefore the concurrent working memory load will tend to reduce the activation levels of the initial conclusions.

As a result, initial conclusions are more likely to be forgotten. But not all conclusions will be influenced equally. Those with higher activation are less likely to be forgotten. This suggests that those conclusions that have merged with prior beliefs are less likely to be forgotten than the conclusions which have not merged. There will be a greater influence of merged conclusions when there is a concurrent working memory than without.

This will have a particularly noticeable effect in indeterminate problems. Without the working memory load it is more likely that both conclusions will be recalled leading to the correct invalid response. When only the merged conclusion is recalled the response will depend on the believability of the conclusion. In the believable condition the merged conclusion matches so a valid response is given, in the unbelievable conclusion the merged conclusion does not match and so an invalid response is given. In other words, the interaction of belief and logic when comparing valid and indeterminately invalid problems will be accentuated with a concurrent working memory load. The effect of working memory load on determinately invalid problems will not be so great.

Method

Participants Thirty-two students from the University of Surrey participated in return for course credit.

Design A within subjects, three factor design was used. Problems varied in the believability of their conclusion (*believable* or *unbelievable*), their validity (*valid, determinately invalid* or *indeterminately invalid*) and whether there was a concurrent working memory load or not (*load* or *no load*). The dependent variable was the number of conclusions accepted in each condition. Problems were presented in a different random order to each participant. The order of the working memory load condition was counterbalanced. Half of the participants completed the working memory load condition before the no load condition and the other half completed the no load condition first.

Materials Participants completed twenty-four problems, two of each type. The problems used were all two-model problems of the form described above. The content of the problems was based on those used by Roberts and Sykes (2003) and all described temporal relations between a range of historical events. An example has been presented above. Two premises were believable and two were unbelievable in each problem in order to control for premise believability. A second set of problems was created by reversing the relational term for each conclusion. Thus the believable valid problem in the first set becomes an unbelievable determinately invalid problem in the second set, and so on. This technique counterbalances the strength of belief in the believable and unbelievable conditions and avoids discrepancies arising if the conclusions chosen for one condition are stronger in belief overall than the other. Half of the participants used the first set of problems and half used the second set.

Procedure Participants completed the task individually. They were given the following written instructions: 'This is an experiment to test peoples' reasoning ability. You will be given 24 problems in total. For each problem you will be shown four statements and you are asked if a certain conclusion (given below the statements) may be logically deduced from them. You should answer this question on the assumption that the statements are, in fact, true. If, and only if, you judge the conclusion necessarily follows from the statements, you should press 'd' on the keyboard, otherwise press 'k'. Please answer all the questions as accurately and quickly as you can. Please press the spacebar when you are ready to move on to the next problem. Additionally, during twelve of these problems you will be asked to remember five random numbers between 1 and 9 whilst you are solving the problem. When you have solved it, recall these numbers out loud for the experimenter to note down. There will be a different set of numbers for each problem'. Participants were given a practice problem and their solution was discussed to ensure they had understood the task. Then they completed the experiment. The practice and experimental materials were presented using a computer.

Results

Responses to the forward and reversed problems were combined and the percentage of conclusions accepted calculated for each condition. These responses are presented in Figure 1. A 2x3x2 within subjects ANOVA showed a main effect of validity $F(2, 62) = 63.22$, $p<0.0001$, no main effect of belief $F(1, 31) = 1.00$, $p>0.05$ and an interaction of validity and belief $F(2, 62) = 5.64$, $p<0.01$. Thus the expected effects of validity and the interaction with belief were found, but unexpectedly there was no main effect of belief. Examining figure one it seems that the reason for this was that unbelievable valid conclusions were accepted more frequently than believable valid conclusions. This cannot be attributed to content effects of the problems because of the counterbalancing of the two sets of problems with forward and reversed conclusions. The explanation of this effect is not clear.

It was predicted that the effect of belief bias would be greater with the working memory load than without. Therefore the percentage of conclusions accepted for valid and indeterminately invalid problems were examined separately for the load and no load conditions. In the load condition there was a main effect of validity $F(1,31) = 19.36$, $p<0.0001$, a marginally significant effect of belief $F(1,31) = 3.53$, $p=0.07$ and a significant interaction of validity and belief $F(1,31) = 8.91$, $p<0.01$. It was also predicted that this interaction of valid and invalid problems would be present only with indeterminately invalid problems. Comparing valid and determinately invalid problems, a main effect of validity was found $F(1,31) = 38.02$, $p<0.0001$, but no main effect of belief was found $F(1,31) = 0.09$, $p>0.05$ and no interaction of validity and belief $F(1,31) = 0.90$, $p>0.05$. Thus the predicted effects of working memory load were found. Believable conclusions were accepted more often, specifically in the indeterminately invalid condition.

In the no load condition when comparing valid and indeterminately invalid problems there was a main effect of validity $F(1,31) = 40.63$ $p<0.0001$, but no main effect of

belief $F(1,31) = 0.03$, $p>0.05$ nor an interaction of validity and belief $F(1,31) = 2.05$, $p>0.05$. Comparing valid and determinately invalid problems a main effect of validity was found $F(1,31) = 181.19$, $p<0.0001$, no main effect of belief $F(1,31) = 1.34$, $p>0.05$ and no interaction of validity and belief $F(1,31) = 0.24$, $p>0.05$. Therefore the predictions of this experiment were supported. The effect of belief bias was greater when there is a working memory load than without and was especially strong in the indeterminately invalid problems.

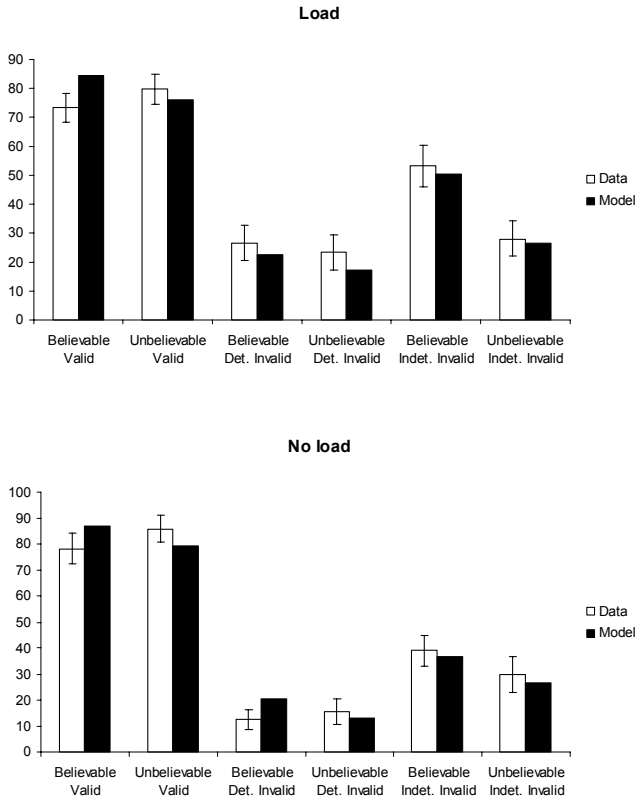

Figure 1: Percentages of conclusions accepted by condition with and without a working memory load, and model predictions

Model Fit

The concurrent working memory load was modeled by placing a chunk containing the number in declarative memory. Productions to retrieve this chunk and subvocalize the number fired whenever ACT-R was not engaged in any other memory retrieval or when it was already subvocalizing the number.

Before assessing the model fit, one further comment on the strategies used in reasoning must be made. Previous research suggested that some participants learnt to draw an invalid conclusion directly after the construction of the ambiguous model of the indeterminate problems rather than constructing the explicit mental models (Banks, 2008). However in this experiment it appeared that participants were only using this strategy in the no load condition. A better fit was found in the load condition if it was assumed that no participants used the more sophisticated strategy. It is possible that the working memory load hindered this strategic development as well as the reasoning itself.

The fit between the model and data was good, with an R^2 of 0.96. Believable conclusions were accepted more than unbelievable conclusions in all of the problem types, valid problems were accepted more than invalid problems, and the effect of belief was greater with the indeterminately invalid problems. These effects were accentuated in the working memory load condition; in particular the effect of belief on indeterminately invalid conclusions was greater than in the no load condition. The model does not replicate the unexpected finding that unbelievable valid conclusions were accepted more frequently than believable valid ones.

Discussion

Overall, the model provides good support for the activation-based account of belief bias in relational reasoning. The major belief bias effects were replicated by the model: valid conclusions were accepted more than invalid conclusions; believable conclusions were accepted more than unbelievable conclusions; and the effects of belief were stronger in the indeterminately invalid problems than the valid or determinately invalid problems.

A concurrent working memory load has the effect of disrupting and delaying the reasoning process which leads to lower activation levels of the conclusions stored in working memory. This in turn leads to an increased influence of belief on conclusion evaluation because conclusions that have merged with the prior belief chunk are more likely to remain above the retrieval threshold and be retrieved as they begin with a higher activation level. Novel conclusions that have not merged with prior belief are more likely to be forgotten. A second finding is that the influence of concurrent working memory was not solely through influencing the reasoning process, as predicted, but also altered the strategy adopted to complete the reasoning task and therefore the responses made.

General Discussion

The purpose of this paper has been to explore the link between working memory, relational reasoning, and prior knowledge to explain belief bias. An activation-based account of belief bias is proposed in which the main belief bias effects are explained through the influence that belief has on the retrieval of conclusions from working memory. This account proposes that when evaluating relational reasoning problems, mental models are constructed of the premises from which initial conclusions are drawn. These are then retrieved from memory in order to evaluate the conclusion presented in the problem. The likelihood of a conclusion being retrieved is determined by its activation level, which is raised if the conclusion matches a previously held belief. Therefore belief affects reasoning through its influence on working memory. An ACT-R model of this

theory predicts the major belief bias effects well and also the influence of a concurrent working memory load on responses.

The most common current explanations of belief bias are presented as dual process theories, e.g. Evans (2006) and Stanovich (2004). Essentially, these suggest that belief bias results from the interaction of heuristic and analytic processes. Belief influences the process either by a heuristic which directly cues a belief based response or through influencing the subsequent analysis. However analytic processes can override these influences, on occasion.

Although the mechanisms proposed here differ from specific dual process theories that have been suggested elsewhere, the activation-based account of belief bias described here is actually broadly consistent with a dual process account. It is proposed here that beliefs can influence reasoning in two ways, a prior belief can be retrieved directly in the evaluation of the conclusion or it can influence the retrieval of a conclusion through merging with it. The former process is comparable to a heuristic response in which the salient belief for a problem is retrieved directly. The latter process is more analytical as the initial conclusion has been inferred from the mental model – although prior belief can influence this too, through merging. There are some parallels therefore between this account of belief bias and dual process theories. The difference is that the ACT-R model presented here does not describe two distinct types of processes or systems which work relatively independently. The influence of prior experience and analytic reasoning operate much more closely together to affect the activation level of all chunks in declarative memory.

This theory also differs from previous accounts of working memory and belief bias. Working memory does serve to temporarily retain conclusions used in the reasoning process and is limited in capacity. But this capacity is a function of the time taken for an activation level to decay below a level such that a chunk cannot be recalled rather than a limit on the number of items in working memory *per se*. Working memory resources here are not described as inhibiting heuristic responses, the choice of response at any time is determined by an explicit set of production rules in the ACT-R model. However the main aspect in which this model differs from previous accounts is that belief in the items themselves affects working memory performance. Conclusions within it are not neutral and subject to invariant working memory capacities but vary in activation level according to a number of factors, including whether they match prior belief or are novel. Therefore it is proposed here that prior knowledge and working memory work together more closely than has previously been suggested in reasoning (e.g. Quayle & Ball, 2000).

Finally, the goal of this paper has been to develop a more detailed account of belief bias by investigating the role of working memory in reasoning. As a result, a large part of the explanation has rested on retrieval and memory rather than the more common explanations in relational reasoning involving mental model construction. But the claim is not that reasoning is solely about memory. Firstly, the construction of mental models is modeled in some detail and this process is important. Secondly, the aim was to make some simplifying assumptions and to test how effective the resulting model was. For example variation in the mental models constructed is not modeled and the activation levels of the mental models themselves in working memory are not validated. Future work could certainly expand on these and other areas. However, given these assumptions, the fit of the model to the data was still very good. Introducing more detail into the process would perhaps serve only to improve the model further.

References

Anderson, J.R. (2007). *How can the mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J.R., Reder, L.M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology, 30,* 221-256.

Banks, A.P. (2008). An activation level account of belief bias: Modelling the effects of belief strength. *Sixth International Conference on Thinking, Venice.*

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science, 17,* 428-433.

Evans, J.St.B.T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13,* 378-395.

Evans, J.St.B.T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59,* 255-278.

Quayle, J.D., & Ball, L.J. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. *Quarterly Journal of Experimental Psychology, 53A,* 1202-1223.

Roberts, M.J., & Sykes, E.D.A. (2003). Belief bias and relational reasoning. *Quarterly Journal of Experimental Psychology, 56A,* 131-154.

Schaeken, W., Johnson-Laird, P.N., & d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition, 60,* 205-234.

Schaeken, W., Van der Henst, J-B., & Schroyens W. (2006). The mental models theory of relational reasoning: Premises' relevance, conclusions' phrasing and cognitive economy. In W. Schaeken, A. Vandierendonck, W. Schroyens & G. d'Ydewalle (Eds.), *The mental models theory of reasoning: Refinements and extensions.* Mahwah, NJ: Erlbaum.

Stanovich, K.E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.

Vandierendonck, A., Dierckx, V., & Vooght, G.D. (2004). Mental model construction in linear reasoning: Evidence for the construction of initial annotated models. *Quarterly Journal of Experimental Psychology 57A,* 1369–1391.