Cognition 114 (2010) 253–265

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Conceptual illusions

Geoffrey P. Goodwin [a,*], P.N. Johnson-Laird [b]

[a] Department of Psychology, University of Pennsylvania, 3720 Walnut St., Solomon Lab Bldg., Philadelphia, PA 19146, USA
[b] Department of Psychology, Princeton University, Princeton, NJ 08540, USA

ARTICLE INFO

Article history:
Received 20 January 2009
Revised 22 September 2009
Accepted 25 September 2009

Keywords:
Concepts
Boolean concepts
Mental models
Mental representation
Illusions

ABSTRACT

Many concepts depend on negation and on relations such as conjunction and disjunction, as in the concept: *rich or not democratic*. This article reports studies that elucidate the mental representation of such concepts from descriptions of them. It proposes a theory based on mental models, which represent only instances of a concept, and for each instance only those properties, affirmative or negative, that the description asserts as holding in the instance. This representation lightens the demands on working memory, but it also leads to predictable conceptual 'illusions' in which individuals envisage as instances of a concept some cases that in fact are non-instances, and vice versa. Experiments 1 and 2 demonstrated the occurrence of these illusions. Experiment 3 corroborated their results, and showed that the illusions can be alleviated in a predictable way by predicates with certain meanings. These findings cannot be easily explained by alternative theories.

1. Introduction

Many concepts in everyday life depend on combining existing concepts using negation, and such logical connectives as *and* and *or* (Bruner, Goodnow, & Austin, 1956). For example, the concept of a 'ball' in baseball is defined as: a pitch at which the batter does *not* swing *and* which does *not* pass through the strike zone. Systems based on these connectives, and those that can be defined in terms of them, are known as 'Boolean' in honor of George Boole, the logician who first formulated their algebra. Even concepts that are not based on a formal definition depend in part on Boolean connectives. Consider, for example, the concept of ownership conveyed by an assertion of the form, *x owns y*. On one analysis (Miller & Johnson-Laird, 1976, p. 560), the concept means in part: *It is permissible for x to use y, and not permissible for others to prevent x from using y*. Likewise, the concept of a leg, as in *a table's leg*, whether it depends on necessary conditions (e.g., Armstrong, Gleitman, & Gleitman, 1983; Fodor, 1998; Osherson & Smith, 1981), a prototype (Hampton, 1979; Posner & Keele, 1968, 1970; Rosch & Mervis, 1975; Smith & Medin, 1981), exemplars (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986; Nosofsky & Palmeri, 1997), general knowledge (Keil, 1989; Murphy & Medin, 1985), or some other hybrid process (Love, Medin, & Gureckis, 2004; Nosofsky, Palmeri, & McKinley, 1994; Smith & Minda, 1998, 2000) cannot be grasped without access to a Boolean system, e.g., *a leg has a maximal dimension and is rigid enough that it can support part of what it is a leg of*. A major question is, therefore, how are Boolean relations represented in the mind?

One view is that Boolean relations are represented in a mental language, which contains expressions of the form, e.g., *a and b*, *not a or not b*. These expressions make explicit the logical form of propositions, and the mind contains formal rules of inference for deriving inferences from them (e.g., Rips, 1994, 2002). Likewise, the acquisition of a concept calls for individuals to set up a decision tree that yields a correct classification of instances and non-instances of the concept (e.g., Hunt, 1962), or to find a minimal description consistent with the instances of the concept (Feldman, 2000, 2003).

* Corresponding author. Tel.: +1 215 746 3579; fax: +1 215 898 7301.
E-mail address: ggoodwin@psych.upenn.edu (G.P. Goodwin).

0010-0277/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2009.09.014

An alternative possibility, however, is that individuals represent Boolean concepts in mental models. The model theory postulates that human reasoning depends, not on logical form, but on mental models of possibilities (Johnson-Laird, 2006; Legrenzi, Girotto, & Johnson-Laird, 2003). Individuals use the meaning of expressions and their knowledge to envisage what is possible, and they represent each distinct sort of possibility in a mental model. A conclusion is valid provided it holds in all models of the premises, and it is invalid if there is a counterexample, that is, a model in which the premises hold but not the conclusion. Mental models differ from other proposed sorts of mental representation, such as expressions in a mental language, because models are as iconic as possible: their structures correspond to the structure of what they represent. They can likewise unfold in time kinematically to simulate sequences of events (Goodwin, Bucciarelli, & Johnson-Laird, 2009). But, they do also contain some symbols that are not iconic, such as a symbol for negation (see Peirce, 1931–1958, Vol. 4, for an account of icons and symbols). The model theory provides an explanation of how individuals make deductions, inductions, explanatory abductions, probabilistic inferences, and inferences to default conclusions that hold in the absence of evidence to the contrary (see Johnson-Laird, 2006, for a review).

The model theory extends naturally to the representation of concepts: mental models represent the different sorts of instance of a concept. And a key assumption for concepts, which parallels an assumption about reasoning (e.g., Johnson-Laird & Savary, 1999), is the principle of *conceptual truth*:

> Mental models represent only the instances of a concept, and in each instance they represent only those properties, or their negations, that the description ascribes to the instance.

This principle minimizes both the processing load on working memory, and yields parsimonious representations. But, the principle is subtle, because it applies at two levels. At the first level, mental models represent only the instances of a concept, not its non-instances. At the second level, a mental model represents only those properties, or their negations, that the description asserts as holding in an instance. This point can be best explained by way of an example. Consider a concept based on an exclusive disjunction, such as:

Red or else not square.

Here, and for the rest of the paper, 'or else' refers to an exclusive disjunction, i.e., *A or else B* rules out the possibility of both *A* and *B* holding, whereas 'or' refers to an inclusive disjunction in which both *A* and *B* can hold. According to the principle of conceptual truth, the concept above has two mental models shown here on separate lines:

red
¬ square

where '¬' denotes the symbol for negation. Each model represents a different sort of instance of the concept. One sort consists of instances that are red, and the other sort consists of instances that are not square. Mental models accordingly do not represent non-instances of the concept, such as instances that are *red and not square*, or *not red and square*. Likewise, each instance represents only what the description asserts to hold within it. Hence, the first model does not represent that *not square* does not hold in this sort of instance, i.e., these instances are red squares. And the second model does not represent that *red* does not hold in this sort of instance, the instances are neither red nor square. An alternative description of the concept is accordingly: *red if and only if square*, but individuals do not normally realize that this equivalence holds.

We have written a computer program (in Common LISP) that implements the principle of conceptual truth for any Boolean concept. The program takes as input a description of a concept, which may contain negations, conjunctions, inclusive and exclusive disjunctions, and various other connectives, and its output is a set of mental models that represent the possible sorts of instance of the concept. The program also constructs fully explicit models, which represent the status of all properties in all instances. As we discovered in the output of the program, the mental models of a concept do not always correspond to the fully explicit models of the concept. If individuals follow the principle of conceptual truth they should accordingly make systematic misunderstandings of certain concepts. In order to elucidate this prediction, we consider the workings of the program in more detail.

The program uses a grammar to parse an input description, and each rule in the grammar has a corresponding semantics, so that the parser controls the process of interpretation too. Given, say, the following description:

a or b, and c

the program first constructs mental models of the inclusive disjunction, *a or b*, to yield three sorts of instance:

a
b
a b

It then forms a conjunction of each of them with a model of *c*:

a c
b c
a b c

The mechanisms for forming conjunctions of models are summarized in Table 1. Because any Boolean connective can be defined in terms of negation and conjunction, the program's mechanisms for other connectives can, in effect, be reduced to combinations of conjunctions and negations. The mechanisms in Table 1 contain some subtleties. If one model represents an instance containing the property, *a*, and another model represents an instance containing its negation, ¬ *a*, their conjunction yields the empty (or null) model of a self-contradictory and therefore impossible instance. But, what happens if two particular mental models to be conjoined contain no items in common? Examples illustrating this case occur with this description of a concept based on two exclusive disjunctions:

Table 1
The mechanisms for forming conjunctions of pairs of mental models of concepts, and pairs of fully explicit models, from the separate clauses of a Boolean description.

1. If one mental model represents a property, *a*, which is not represented in the second mental model, then if *a* occurs in at least one of the set of models from which the second model is drawn, then its absence in the second model is treated as its negation (and mechanism 2 applies); otherwise its absence is treated as its affirmation (and mechanism 3 applies). This mechanism applies only to mental models.
2. The conjunction of a pair of models containing respectively a property and its negation yields the null model (of an impossible instance), e.g.:
   a b and ¬ a b yield nil.
3. The conjunction of a pair of models that are not contradictory yields a model representing all the properties in the models, e.g.:
   a b and b c yield a b c.
4. The conjunction of a null model with any model yields the null model, e.g.:
   a b and nil yield nil.

a or else b, and b or else c.

The reader is invited to think of the possible instances of this concept. Most individuals think correctly of the following two:

a c
b

The mental models of the first exclusive disjunction are:

a
b

and the mental models of the second exclusive disjunction are:

b
c

The program forms the product of all four possible conjunctions of the models in these two sets. It makes a conjunction of the first model in the first set, *a*, with the first model in the second set, *b*. The mechanisms for conjoining mental models of concepts treat the absence of a property in a model as equivalent to its negation if that property occurs elsewhere in the set of models. The property *b* occurs in the set of models of the first disjunction from which *a* is drawn, and so the interpretative system takes the absence of *b* in the current model to mean *not b*. The conjunction thus becomes:

a ¬ b and b

Because there is a contradiction – one model contains *b* and the other its negation, ¬ *b*, the result is the null model (see Table 1).

The program next forms a conjunction of the first model in the first set, *a*, with the second model in the second set, *c*. The property *a* does not occur in the second set of models, and the property *c* does not occur in the first set of models, and so the two models are compatible with one another, and their conjunction yields:

a c

The program now forms a conjunction of the second model in the first set, *b*, with the first model in the second set, *b*. The conjunction yields:

b

The final conjunction is of the second model in the first set, *b*, with the second model in the second set, *c*. It yields the null model, because *b* occurs in the second set of models, and so its absence is treated as akin to its negation. Once again, there is a contradiction yielding the null model. Hence, the concept does indeed have the two instances shown above.

The mechanisms in Table 1 apply to the conjunction of mental models, which abide by the principle of conceptual truth and accordingly represent only what holds in each instance. But, they also apply to the conjunction of fully explicit models, which represent both what holds and what does not hold in each instance. The first mechanism in the table, however, is not relevant to fully explicit models. Here is the previous description again:

a or else b, and b or else c.

The fully explicit models of the two exclusive disjunctions are respectively:

a ¬ b b ¬ c
¬ a b ¬ b c

There are four pair-wise conjunctions, but two of them are contradictions yielding the null model. The remaining pairs yield the following models:

a ¬ b c
¬ a b ¬ c

These match the same instances as before, but now they make explicit the status of all properties in both instances of the concept. The principle of conceptual truth lightens the processing load on working memory, because it leads to the representation of only what holds in each instance according to the description. It seems innocuous, but, as we show presently, it can have striking effects on what individuals think are instances of a concept.

In fact, this principle makes a novel prediction that seems beyond the power of other current theories to make: there should be systematic failures to infer the correct instances of certain concepts from their descriptions. Consider this concept, for instance, which describes a set of possible objects based on the attributes of color and shape:

red if and only if square, or else red.

According to the mechanisms in Table 1, individuals should envisage that the concept has two sorts of instance corresponding to the mental models that the program yields:

red square (the biconditional holds)
red (‘red’ holds)

But, the fully explicit models of the concept, representing both what holds and what does not hold in each instance, show that the correct instances of the concept are:

¬ red	¬ square	(the biconditional holds, but 'red' does not)
red	¬ square	(the biconditional does not hold, but 'red' does)

Individuals should therefore suppose that one sort of instance of the concept is *red and square*. But, as the fully explicit models show, this instance is illusory. It does not correspond to the correct interpretation of the concept's description. To help readers grasp this crucial point, we present its truth table below:

Red	Square	Red if and only if square	Red if and only if square, or else red
True	True	True	False
True	False	False	True
False	True	False	False
False	False	True	True

The two fully explicit models above correspond to the two true instances in the final column.

The illusion arises because individuals tend to think about the truth of one clause in the exclusive disjunction, and then the truth of the other clause, i.e., they envisage mental models of the disjunction. To understand the concept correctly, however, they need to envisage fully explicit models. That is, they need to grasp that an instance of the sort, *red and square*, makes both clauses of the exclusive disjunction true, and so it is impossible. In other words, when the biconditional holds, as it does for *red and square*, the second clause in the disjunction, 'red', must not hold; hence, *red and square* is not an instance of the concept. But, *not red and not square* is a sort of instance in which the biconditional holds and 'red' does not; likewise, *red and not square* is a sort of instance in which 'red' holds and the biconditional does not. Because individuals should tend to rely on mental models, the theory predicts that they should make a systematic error about the instances of this 'illusory' concept. The theory does not claim that such errors are common-place, or that they are likely to arise for simple or familiar Boolean concepts. However, in daily life, people do sometimes encounter complex and unfamiliar concepts, not unlike the examples above, particularly in the conveyance of technical instructions or rules. Such settings are, we argue, likely to give rise to illusory concepts.

Skeptics may argue that such concepts are highly artificial, and that errors are merely a consequence of this artificiality. A simple control problem, however, is just as artificial. It depends on a description containing an inclusive disjunction:

red if and only if square, or red.

Its mental models are:

red	square
red	
...	

where the ellipsis denotes an implicit mental model in which another sort of instance is possible. The fully explicit models of the concept show that the mental models are correct, and that *red and square* is a genuine instance of the concept:

red	square
red	¬ square
¬ red	¬ square

Hence, the theory predicts that this concept, which is just as artificial as the previous one, should yield a correct performance. Of course, skeptics might now argue that exclusive disjunction is the source of the problem, and so in order to counter this claim, we carried out an experiment using a variety of illusory and control problems, including some that did not depend on exclusive disjunctions.

2. Experiment 1: illusory concepts

The experiment investigated whether individuals were susceptible to conceptual illusions. Participants were presented with descriptions of different concepts, and were asked to write down all the possible sorts of instance of them. As we have illustrated, the mental models of some concepts do not match the fully explicit models, which are correct. The illusory problems were based on such concepts: they had at least one mental model not in the set of fully explicit models. Hence, if individuals rely on mental models, they should err on these problems. The control problems were of two sorts (see Table 2): *basic* controls had mental models matching one-to-one the fully explicit models, and *subset* controls had mental models matching a proper subset of the fully explicit models. A second prediction was that the first sort of control problems should yield better performance than the second sort of control problems.

2.1. Method

Twenty-two participants (11 female, 11 male) from Princeton University participated in the experiment for course credit. They acted as their own controls and carried out the 20 problems in Table 2. There were nine illusory problems and 11 control problems (seven basic controls and four subset controls).

The concepts were described in terms of two properties: square or not square, and red or not red. Two different assignments of these properties to the *a* and *b* variables in Table 2 were made. The first assignment was made by designating *a* as red and *b* as square for approximately half of the problems in each of the three categories, and *a* as square and *b* as red for the other half of the problems. The second assignment switched these designations for each problem. The three sorts of problem occurred in four different random orders – an initial random order and its reverse were used for the first assignment, and a second, new random order and its reverse were used for the second assignment. The different clauses of each problem were clearly separated through the use of parentheses, as Table 2 illustrates.

Table 2
The set of 20 problems, their models, and the percentage of correct descriptions in Experiment 1.

Type of problem	Description	Mental models		Correct fully explicit models		Percentage correct
Basic controls	(If a then b), and a	a	b	a	b	59
	(a and b), or else (not a and b)	a	b	a	b	91
		¬ a	b	¬ a	b	
	(a and b), or else (a and not b)	a	b	a	b	82
		a	¬ b	a	¬ b	
	(If a then b), or else (if a then not b)	a	b	a	b	50
		a	¬ b	a	¬ b	
	(If not a then b), or else (if not a then not b)	¬ a	b	¬ a	b	55
		¬ a	¬ b	¬ a	¬ b	
	(a and not b), or else (not a and b)	a	¬ b	a	¬ b	91
		¬ a	b	¬ a	b	
	(a and b), or else (not a and not b)	a	b	a	b	95
		¬ a	¬ b	¬ a	¬ b	
Subset controls	(a if and only if b), or b	a		¬ a	¬ b	56
		a	b	a	¬ b	
			...	a	b	
	(a and not b), or (not a or else b)	a	¬ b	a	¬ b	73
		¬ a		¬ a	¬ b	
			b	a	b	
	(a or else b), or else (a and b)	a	b	a	b	77
		a		a	¬ b	
			b	¬ a	¬ b	
	(a and b) if and only if a	a	b	a	b	23
			...	¬ a	b	
				¬ a	¬ b	
Illusions	(a if and only if b), or else b	a	b	¬ a	¬ b	32
			b	¬ a	b	
	(If a then b), or else a	a	b	¬ a	b	36
		a		¬ a	¬ b	
			...	a	¬ b	
	(a and b), or else a	a	b	a	¬ b	32
		a				
	(a and b), or else b	a	b	¬ a	b	27
			b			
	(If a then b) or else if (not a then b)	a	b	¬ a	¬ b	36
		¬ a	b	a	¬ b	
			...			
	(If a then b), or else (if not a then not b)	a	b	¬ a	b	17
		¬ a	¬ b	a	¬ b	
			...			
	(a and b), if and only if not b		...	¬ a	b	23
	(a or not b), or else (a or else b)		¬ b	a	b	27
		a	¬ b	¬ a	¬ b	
		a		¬ a	b	
			b			
	(a or else b), or else (a or b)	a	b	a	b	41
		a				
			b			

The participants were told to imagine that each description concerned a set of objects in a box, and they were to write down all of the possible sorts of object in the box. The meaning of each of the connectives was explained to them on an instruction sheet, and they were shown which possibilities were consistent with each connective. Before they began the experiment proper, they had to write down these possibilities on a separate sheet of paper, and they were able to consult the entire instruction sheet throughout the course of the experiment. This initial phase of the experiment served as a training phase on the meaning of the logical connectives used in the experiment. If participants misunderstood the meaning of one or other of the connectives (by writing down an incorrect set of possibilities), their error was explained to them before they continued to the main part of the experiment. Such errors happened rarely.

2.2. Results

One participant failed to follow the instructions, and this person's data were excluded from the analysis. We scored responses as correct if participants provided all and only the correct set of possible instances of a concept. Fig. 1 shows the robust declining trend in correct performance across basic controls, subset controls, and illusions (75% vs. 58% vs. 30% correct, Page's L, $z = 4.30$, $p < .0001$). Eleven out of the 22 participants showed the predicted trend exactly. The illusions were particularly difficult – 19 out of 22 participants performed them less accurately than the two sorts of control (Sign test, $p < .001$). These analyses were corroborated using a more sensitive scoring method ranging from 0 to 4. On this scoring method, a separate point was allocated for each separate possibility that participants correctly categorized – either by writing it down if it was a correct possibility, or by omitting it, if it was an incorrect possibility. There was again a reliable declining trend in accuracy across the three sorts of problem using this measure (3.52 vs. 3.32 vs. 2.67, Page's L, $z = 4.60$, $p < .0001$). And again, nineteen out of 22 participants performed worse on the illusions than on the controls (Sign test, $p < .001$).

A further analysis showed that the participants' erroneous models for the illusory problems were those that the model theory predicts (see Table 2). That is, the erroneous descriptions matched the mental models of the concepts, as generated by the computer program. The 21 participants who made errors constructed an illusory model predicted by the model theory on 72% of their erroneous responses. Chance performance is difficult to estimate, because it depends on the number of mental models for each problem, and because some responses may be, a priori, more likely than others. However, given that there are nine possible instances that might be listed for each problem (four conjunctive models: a b, a ¬ b, ¬ a b, ¬ a ¬ b; four singular models: a, ¬ a, b, ¬ b; and the response 'null'), .5 is a very conservative estimate of the probability that any one of these models will appear in a participant's response. And 72% is reliably greater than this conservative estimate of chance (Wilcoxon test, $z = 2.74$, $p < .01$).

2.3. Discussion

The results corroborated the occurrence of conceptual illusions arising from the principle of truth. The illusions cannot be attributed to the difficulty of particular connectives, such as *or else*, because poor performance occurred for the illusory problem in which *or else* was not present (only 23% correct, see Table 2), and good performance occurred for many control problems based on *or else*. Indeed, what appeared to make control problems more difficult was the presence of a conditional connective. Nor can the results be attributed to the number of models required by the problems. On average, the illusory problems gave rise to fewer fully explicit models than the control problems (1.78 vs. 2.27). The phenomenon is robust and extends the model theory's principle of truth to conceptual descriptions.

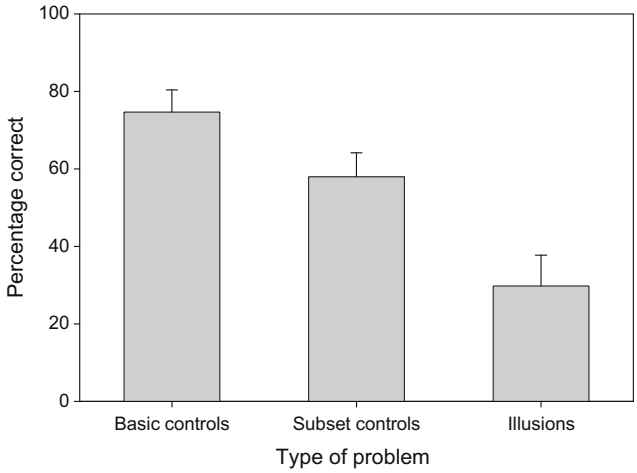

Fig. 1. The percentages of correct responses to problems (with standard errors) for the three sorts of problem in Experiment 1.

3. Experiment 2: replication without 'or else'

The results of the first experiment cannot be attributed solely to the difficulty of *or else*. However, a related concern is that participants might have misinterpreted the term *or else*, i.e., they may have misunderstood this term when it was the main connective in the problems, which led them to list erroneous possibilities. We did explain the meaning of *or else* to the participants, and indeed, their understanding of its meaning was tested in a training phase before they began the main experiment. Nevertheless, it remains possible that some participants might have been misled by the term, and treated it as equivalent to an inclusive rather than an exclusive disjunction. This objection faces the problem that performance was high on the six control problems in which *or else* was the main connective. Nevertheless, we ran Experiment 2 in order to deal with this objection in a more stringent way.

Participants performed four control problems and four illusory problems, using the same overall rubric about descriptions of objects in a box. However, unlike the previous experiment, the main connective in each problem was re-described in plain English. For instance, instead of being presented with the main connective *or else*, participants were instructed: "Only one of the following statements is true about a particular object in the box". Similarly, when the main connective was *if and only if*, participants were instructed: "If one of the following statements is true about a particular object in the box then the other statement is also true". As Table 3 illustrates, the four illusory problems consisted of two which used the "only one of the statements is true" rubric (equivalent to *or else*), and two which used the "if one statement is true, then the other is also true" rubric (equivalent to *if and only if*).

3.1. Method

Thirty-three participants (18 female, 15 male) from the University of Pennsylvania participated in the experiment for course credit. They acted as their own controls and carried out eight problems. The instructions and procedure were similar to those in the previous experiment. For each problem, participants were instructed that a set of objects had been placed in a box, and each object was either red or not red, and either square or not square. They were then presented with a description of a particular object in the box, and were asked to write down all the possibilities for the object.

There were four control problems, which were taken from the set of 'basic controls' in the previous experiment. For each of these problems, the initial mental models produced by our computer program correspond exactly to the correct set of models. There were four illusory problems, three of which were taken from the previous experiment, and one of which was new. For each of these problems, the initial mental models do not correspond to the correct set and include at least one erroneous model. Each problem was presented, as above, with the main connective

Table 3
The set of 16 problems, their respective models, and the percentage of correct descriptions for the problems in Experiment 2.

Type of problem	Description	Mental models	Correct fully explicit models	Percentage correct
Control (basic)	1. Only one of the following statements is true about a particular object in the box: a and b not a, and b	a b ¬ a b	a b ¬ a b	82
	2. Only one of the following statements is true about a particular object in the box: a and b not a, and not b	a b ¬ a ¬ b	a b ¬ a ¬ b	85
	3. Only one of the following statements is true about a particular object in the box: if not a then b if not a then not b	¬ a b ¬ a ¬ b	¬ a b ¬ a ¬ b	12
	4. Only one of the following statements is true about a particular object in the box: a and b a and not b	a b a ¬ b	a b a ¬ b	88
Illusions	5. If one of the following statements is true about a particular object in the box then the other statement is also true: a and b not b	...	¬ a b	18
	6. If one of the following statements is true about a particular object in the box then the other statement is also true: either not a or else b not a	¬ a ¬ a b	¬ a ¬ b a ¬ b	12
	7. Only one of the following statements is true about a particular object in the box: a and b b	a b b	¬ a b	15
	8. Only one of the following statements is true about a particular object in the box: if a then b if not a then not b	a b ¬ a ¬ b ...	¬ a b a ¬ b	3

re-described in plain English. Table 3 presents the complete set of problems. In this experiment, *a* was designated as red, and *b* was designated as square.

Participants received an initial instruction sheet that explained the meaning of each of the connectives. As in the previous experiment, they had to reproduce the possibilities consistent with each of these connectives before they proceeded to the main experiment. If they made an error in this stage, their error was explained to them before they proceeded to the main experiment. The problems were presented in one of four different random orders – two initial random orders, and two orders that were their reverses.

3.2. Results and discussion

Responses were again scored as correct if participants provided all and only the correct set of possible instances of a concept. As Table 3 illustrates, control problems were performed far more accurately than illusory problems (67% vs. 12% correct, Wilcoxon test, $z = 4.86$, $p < .0001$). Thirty out of the 33 participants showed the predicted trend, and there were three ties. This analysis was corroborated using the more sensitive scoring method which counts the number of separate possibilities that participants correctly categorized – either by writing down the correct possibilities, or by omitting the incorrect possibilities (control problems: 3.31 vs. illusions: 2.11; Wilcoxon test, $z = 4.88$, $p < .0001$).

A further analysis showed that the participants' erroneous models for the illusory problems were those that the model theory predicts (see Table 3). The erroneous descriptions matched the mental models of the concepts, as generated by the computer program 73% of the time which is reliably greater than a chance estimate of .5 (Wilcoxon text, $z = 4.85$, $p < .0001$).

The poor performance on control problem 3 (only 12% correct, see Table 3) was surprising. The logically analogous problem in Experiment 1 was performed correctly 55% of the time. One important feature of this problem is its use of conditional statements, which are known to be difficult (cf. Johnson-Laird & Byrne, 2002). The most common error for this problem (72% of all errors) was for participants to list all four possibilities, rather than just the two correct ones. This response may have been a result of a simple heuristic: to list all the possibilities for each of the two statements considered separately. The heuristic leads to the listing of all four possibilities for problem 3.

The results of Experiment 2 corroborate the occurrence of conceptual illusions. Moreover, they show that these illusions do not depend on participants misinterpreting the meaning of particular logical connectives such as *or else,* because the illusions still occurred when the main connectives in each problem were re-described in plain English. According to the model theory, the illusions occur because individuals construct erroneous initial models of what is possible given a description of a concept. It follows, then, that the likelihood of an individual's succumbing to an illusion should be reduced if those initial models are not constructed. Experiment 3 tested this prediction by investigating the effects of a semantic manipulation.

4. Experiment 3: the semantic modulation of illusory descriptions

The model theory postulates that the interpretation of sentences containing Boolean connectives can be modulated by the meaning and reference of clauses or by general knowledge (Johnson-Laird & Byrne, 2002). One effect is to block the construction of a particular model. For instance, given the description, 'square or circular' individuals will construct models of just two instances, because they know that an entity cannot have both shapes:

square
circular

Thus, 'semantic modulation' blocks the construction of a third model in which an object is both square and circular. In this case, the principle of conceptual truth still operates, but on the *results* of the semantic modulation. The theory accordingly predicts that modulation should reduce the conceptual illusions that we observed in Experiment 1 when it blocks the construction of an erroneous mental model. This prediction had not been examined in any domain of reasoning, and so Experiment 3 was designed to test it for concepts.

4.1. Method

Twenty-one participants (13 female, 8 male) from Princeton University participated in the experiment for course credit. They acted as their own controls and carried out 16 problems. The instructions and procedure were similar to those in the previous experiment: the participants were presented with descriptions of objects, and wrote down what sorts of object were possible given each description. There were two separate blocks of eight problems: one block of illusory problems derived from those in the previous experiment (unmodulated), and one block of the same problems was with modulated content (modulated). Half of the participants received the modulated problems first, and half of them received the unmodulated problems first. The six illusory problems in each block were presented in one of four random orders, except that adjacent problems were always of a different sort (strongly modulated, partially modulated, weakly modulated, as we explain below). There were also two control problems in each block, presented on the third and sixth trials. Table 4 below shows the full set of problems used in the experiment. In this experiment, the separate clauses of each problem were demarcated, not with parentheses, but more subtly with commas, as shown in Table 4.

The participants were told that an object could be of only one shape and one color. The unmodulated problems used two terms that were compatible with one another, e.g.: 'blue' and 'circular' in order to allow a model in which both properties occurred. The modulated problems used two terms that were incompatible with one another: 'blue' and 'yellow' (for 13 participants) in order to prevent the construction of a model in which both colors occur; and 'circular' and triangular' (for 8 participants) in order to prevent the construction of a model in which both shapes occur.

Table 4
The set of 16 problems, their respective models, and the percentage of correct descriptions for the problems in Experiment 3.

Type of problem	Description	Mental models: unmodulated	Mental models: modulated	Correct fully explicit models	Percentage correct: unmodulated	Percentage correct: modulated
Control problems	If not a then b, or else, if not a then not b		¬ a b ¬ a ¬ b	¬ a b ¬ a ¬ b		38
	a and not b, or else, not a and b		a ¬ b ¬ a b	a ¬ b ¬ a b		95
	a and b, or else, a and not b	a b a ¬ b		a b a ¬ b	90	
	a and b, or else, not a and not b	a b ¬ a ¬ b		a b ¬ a ¬ b	90	
Strongly modulated illusory problems	a and b, or else, a	a b a	a	a ¬ b	0	48
	a and b, or else, b	a b b	b	¬ a b	0	57
Partially modulated illusory problems	a if and only if b, or else, a	a b a	a	¬ a ¬ b a ¬ b	0	10
	If a then b, or else, a	a b a . . .	a . . .	¬ a b ¬ a ¬ b a ¬ b	0	10
Weakly modulated illusory problems	If a then b, or else, if not a then b	a b ¬ a b . . .	¬ a b . . .	¬ a ¬ b a ¬ b	0	0
	If a then b, or else, if not a then not b	a b ¬ a ¬ b . . .	¬ a ¬ b . . .	¬ a b a ¬ b	0	0

The experiment used three sorts of illusory problems depending on the predicted effects of modulation. *Strong* modulation blocks an illusory mental model, *yielding only a correct mental model*. For example, the description: 'square and circular, or else square', blocks the otherwise illusory model:

square circular

to yield only the mental model:

square

Individuals simply need to add the missing value to this model to arrive at the correct instance:

square ¬ circular

Partial modulation blocks an illusory mental model, but does not aid in the construction of the fully explicit models, because some of them are not included in the set of mental models. For example, the description: 'square if and only if circular, or else square', again blocks the illusory model:

square circular

to yield only the mental model:

square

However, this model needs to be fleshed out and supplemented by a further model to arrive at the correct set of models:

square ¬ circular
¬ square ¬ circular

Weak modulation blocks one illusory mental model, but does not block the construction of other illusory mental models. For example, the description: 'if square then circular, or else if not square then circular', blocks the illusory mental model:

square circular

to yield only the mental model:

¬ square circular

However, this second model is still illusory, because the correct models are in fact:

¬ square ¬ circular
square ¬ circular

Hence, there should be an increasing trend in accurate performance: strong modulation should yield a greater increase than partial modulation, which in turn should yield a greater increase than weak modulation, which in fact, should have no effect on performance. There were

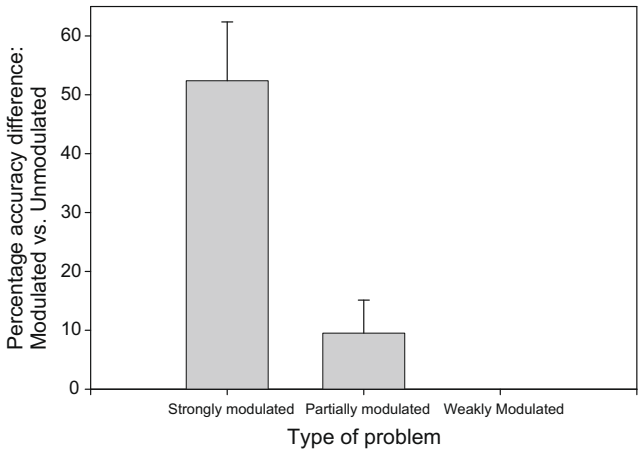

Fig. 2. The effect of modulation in Experiment 3. The histograms represent the difference in percentages of accuracy between modulated and unmodulated versions for each of the three sorts of illusory problem.

two instances of each of these three types of modulated problems within the block of modulated problems.

Participants were instructed about the basic meanings of the connectives: 'or', 'or else', 'if', and 'if and only if'. However, unlike the previous experiments they were not shown the possibilities compatible with each connective, nor did they have to write these down in a training phase.

4.2. Results

Table 4 shows the percentage of correct responses for each problem. As in the previous experiment, the illusory problems elicited a much less accurate performance than the control problems overall (10% vs. 79% correct, Wilcoxon test, $z = 4.03$, $p < .0001$). All 21 participants showed the predicted effect. The more sensitive scoring method, which assigned responses a number from 0 to 4 depending on how many possibilities were coded correctly, also demonstrated worse performance on the illusory problems than the control problems (1.87 vs. 3.63, Wilcoxon test, $z = 4.02$, $p < .0001$). As in the previous experiment, the illusory problems tended to elicit the erroneous models that the model theory predicts (on 79% of erroneous responses, which is reliably greater than a conservative chance estimate of 50%, Wilcoxon test, $z = 4.02$, $p < .0001$).

Fig. 2 shows the effect of the three sorts of modulation, and the predicted trend was reliable (Page's L, $z = 3.09$, $p < .005$). Strong modulation reliably improved performance in comparison with no modulation (52% vs. 0% correct, Wilcoxon test, $z = 3.31$, $p < .001$). Partial modulation had an effect in the right direction, but it was not reliable (10% vs. 0% correct, Wilcoxon test, $z = 1.63$, $p = .10$). And weak modulation had no effect whatsoever (0% vs. 0% correct).

4.3. Discussion

The experiment replicated the finding that illusory problems were much more difficult than control problems. But, it also corroborated the predicted effects of modulation. When knowledge of an incompatibility between, say, square and circular ruled out an illusory model to leave only a correct model (strong modulation), performance was much improved. When the remaining model needed to be supplemented with additional correct models (partial modulation), there was a trend towards improvement. But, when the remaining model was itself illusory, there were no reliable signs of improvement. Previous investigations have found it difficult to eliminate illusions in deductive reasoning (e.g., Santamaria & Johnson-Laird, 2000; Yang & Johnson-Laird, 2000). The present experiment shows that modulation can alleviate the difficulty of illusory inferences. It does not follow, however, that it can dispel the difficulty of all illusory problems. The effect depends on the specific sort of concept, that is, on the mental models that it elicits.

The illusory problems were performed less accurately in the present experiment than in the preceding experiments. The six illusory problems in both Experiments 1 and 3 yielded a reliably poorer performance in Experiment 3 (unmodulated problems) (Experiment 3: 0% correct vs. Experiment 1: 29.6%; Mann–Whitney U-test, $z = 3.23$, $p < .01$). Similarly, the two illusory problems in both Experiments 2 and 3 yielded a marginally poorer performance in Experiment 3 (Experiment 2: 9% vs. Experiment 3: 0%; Mann–Whitney U-test, $z = 1.86$, $p < .07$). One pertinent factor is that the participants in Experiments 1 and 2 had to reproduce the meanings of each connective before tackling the main problems. Another relevant factor is that the separate clauses in the problems were demarcated with parentheses in Experiment 1, but only with commas in Experiment 3. In Experiment 2, the separate clauses were demarcated as separate assertions, owing to the redescriptions of the main connectives.

5. General discussion

The aim of the present investigation was to examine the principle of conceptual truth, which extends the mental model theory into the domain of concepts. According to the principle, mental models represent the sorts of instance of a concept, and each model of a sort of instance represents a property in the description of the concept, such as 'not

red', only when it holds in that sort of instance. Hence, a description such as:

Blue and square, or else blue

has these mental models:

blue square
blue

In contrast, fully explicit models represent the properties in each sort of instance, whether they or their negations hold. These models accordingly represent the correct instances of a concept. The concept above has just a single fully explicit model:

blue ¬ square.

In particular, the connective *or else* means that when the conjunction *blue and square* holds for a sort of instance, the second disjunct in the description, *blue*, does not hold. Hence, there cannot be a sort of instance that is blue and square. However, when the second disjunct, *blue*, holds for an instance, the first disjunct – the conjunction *blue and square* – does not hold. Thus, this case allows for one sort of instance, i.e., *blue and not square*.

The principle of conceptual truth predicts a variety of similar illusions, and all three experiments corroborated their occurrence. The participants had to list what was possible given both illusory descriptions and control descriptions for which the principle predicts correct performance. The illusions produced far fewer correct lists of possibilities than the control problems. The participants also tended to list the predicted illusory possibilities, whereas they performed almost without error with the control problems.

The poor performance on illusory problems was not because the participants had difficulty in understanding the meaning of particular logical connectives, such as *or else*. Performance was good on control problems for which *or else* connected the two main clauses of the problem. But, performance was uniformly low on illusory problems – both when *or else* was the main connective, and when it was not. Likewise, Experiment 2 showed that when the main connective in the problems was replaced with a description in plain English, the illusions were just as likely to arise. Together, this evidence suggests that the difficulty of illusions springs not from superficial errors in understanding, but rather from more deep-seated representational processes, as the principle of conceptual truth implies.

The model theory allows that meaning, reference, and general knowledge can modulate the interpretation of connectives. One effect of modulation is to block the construction of a possibility, and this phenomenon has been corroborated for conditional assertions (Johnson-Laird & Byrne, 2002; Quelhas, Juhos, & Johnson-Laird, 2009). Experiment 3 exploited a simple form of modulation, using it for the first time to alleviate illusions. The following version of the problem above:

red and green, or else green.

should not elicit the illusory model:

red green

because individuals know that the objects under description cannot have two colors. Experiment 3 corroborated this prediction. Individuals were much less likely to succumb to illusions when modulation blocked an illusory instance, leaving only a correct instance of the concept (strongly modulated problems). Modulation had a mild effect on performance when it blocked an illusory instance, but left the participant to recover the correct instances (partially modulated problems). And it had no effect whatsoever when it blocked one illusory instance but not another (weakly modulated problems).

Across all three experiments, performance on the control problems varied depending on which connectives occurred in the problems (see Tables 2–4). In particular, performance was worse for problems that included conditional or biconditional assertions than for those that did not. In Experiment 1, control problems (both basic and subset controls) with a conditional or biconditional yielded 49% correct responses, but those without these connectives yielded 85% correct responses (Wilcoxon test, $z = 3.73$, $p < .001$). In Experiment 2, the single control problem with a conditional yielded 12% correct responses, but the three other control problems yielded 85% correct responses (Wilcoxon test, $z = 4.82$, $p < .001$). The comparison is not feasible for Experiment 3, because there were no unmodulated control problems that included either a conditional or biconditional.

Conditionals give rise to a greater number of models than do conjunctions and exclusive disjunctions (though not inclusive disjunctions), and are known to be difficult to process (see e.g., Johnson-Laird & Byrne, 2002). So it is not surprising that the control problems that included them were performed more poorly than those that did not. Nevertheless, the difference between control and illusory problems remained reliable among only those problems that included either a conditional or biconditional. In Experiment 1, this difference was highly reliable (49% accuracy vs. 28% accuracy; Wilcoxon test, $z = 3.08$, $p < .01$), and in Experiment 3, it was also reliable (12% accuracy vs. 3% accuracy; Wilcoxon test, $z = 1.73$, $p < .05$, one-tailed). Hence, the difficulty of conditionals does not threaten the main argument of the present paper.

The present investigation has yielded three novel outcomes. First, it demonstrated the existence of a new class of illusory problems that occur in the understanding of concepts. It thus extends the scope of the model theory, in line with similar recent work that has extended the model theory to the acquisition of concepts from their instances (Goodwin & Johnson-Laird, 2009). Second, it has established a more fine-grained categorization of problems than has previously been possible. As the theory predicted, problems for which the mental models exactly match the correct models are easiest, problems for which the mental models are a subset of the correct models are of intermediate difficulty, and problems for which the mental models contain an entirely erroneous model are the most difficult. And Experiment 1 strongly corroborated this trend. Third, the investigation has shown how semantic modulation can substantially alleviate the difficulty of illusory problems. Illusory inferences have been notoriously difficult to remediate (see e.g., Santamaria & Johnson-Laird, 2000;

Yang & Johnson-Laird, 2000), but the present investigation has established one way in which it can be done. The effects of modulation were also fine-grained, because semantic modulation does not always yield only the correct models. Experiment 3 showed that as the effects of modulation weaken, so too does its power to block illusory inferences.

Formal theories do not predict illusions in reasoning, because they do not concern themselves with the possibilities that premises give rise to. Yet illusions are not readily explained without reference to the divergence between the initial possibilities (mental models) that individuals represent, and the fully explicit set of possibilities. The conceptual illusions that we investigated here are equally difficult to explain using formal rules. Likewise, theories of concepts based on a mental language, such as those that rely on the acquisition of decision trees (Hunt, 1962) or minimal descriptions (e.g., Feldman, 2000, 2003) also have no machinery to explain the present results. Existing theories of concepts, whether they are based on prototypes, exemplars, or some other sort of representation, have tended to focus more on graded concepts rather than Boolean concepts, and do not have any apparent mechanisms for dealing with conceptual illusions. This claim holds even for the most recent theories, which focus on the Bayesian inferences underlying concept learning and representation (e.g., Kemp & Tenenbaum, 2008; Tenenbaum, Griffiths, & Kemp, 2006). In principle, Bayesian theories have no machinery giving rise to radical misrepresentations of problems. However, granted that mental models form the basis of mental representations, a Bayesian approach could make use of them, and therefore be reconciled with the occurrence of illusions.

Further questions remain about the occurrence of the conceptual illusions documented in this paper. We make no strong claim that their occurrence is common. In contrast, our aim is to document the possibility of their occurrence. Many concepts in daily life, such as *bird* or *table*, seem highly unlikely to produce illusions, because their logical structure is relatively simple, and they are highly familiar. How could an individual fall prey to a conceptual illusion with this sort of concept? We accept that illusions are unlikely to arise for this sort of concept, although we make two further observations. First, such concepts do contain some underlying Boolean structure, in addition to a more complex relational structure. A table for instance, can be defined as: "an article of furniture supported by one or more vertical legs and having a flat horizontal surface on which objects can be placed" (Miller & Johnson-Laird, 1976, p. 222–223), which includes both a disjunction and a conjunction. The example thus makes clear the importance of understanding the representation of Boolean components for even seemingly straightforward concepts. Of course, this concept also consists of more complex relational elements: *supported by*, *on which*, etc., which go beyond the scope of the present paper. Second, many concepts in daily life are more complex than these examples, and also much less familiar – particularly those encountered in more technical settings (e.g., in the context of games, legal proceedings, or other contexts in which detailed procedures are stipulated). Concepts that occur in these sorts of domains seem to us to be prime candidates for the production of illusions. We conclude that the model theory is at present unique in predicting conceptual illusions. People fall prey to these illusions because they focus on what holds in the instances of a concept, and overlook what does not hold. The moral we draw is that in order to understand how people represent concepts, it is crucial to know what possibilities they represent.

Acknowledgements

This research was supported in part by a National Science Foundation grant, SES 0844851 to the second author to study deductive and probabilistic reasoning. We thank our colleagues for their advice: Jeremy Boyd, Adele Goldberg, Sam Glucksberg, Sangeet Khemlani, Greg Murphy, and Paula Rubio, and we thank Yael Diamond Schlenger for valuable research assistance. We also thank Ray Nickerson and two anonymous reviewers for their helpful comments.

References

Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13*, 263–308.
Bruner, J. S., Goodnow, J. A., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.
Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633.
Feldman, J. (2003). A catalog of Boolean concepts. *Journal of Mathematical Psychology, 47*, 75–89.
Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Clarendon Press.
Goodwin, G. P., Bucciarelli, M., & Johnson-Laird. (2009). *Everyday reasoning and mental simulation*. University of Pennsylvania. Unpublished manuscript.
Goodwin, G. P., & Johnson-Laird, P. N. (2009). *Mental models as representations of Boolean concepts*. University of Pennsylvania. Submitted for publication.
Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*, 441–461.
Hunt, E. B. (1962). *Concept learning: An information processing problem*. New York: Wiley.
Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.
Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109*, 646–678.
Johnson-Laird, P. N., & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition, 71*, 191–229.
Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.
Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*, 10687–10692.
Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.
Legrenzi, P., Girotto, V., & Johnson-Laird, P. N. (2003). Models of consistency. *Psychological Science, 14*, 131–137.
Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.
Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.
Murphy, G., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded categorization. *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.

Osherson, D., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition, 9*, 35–58.

Peirce, C. S. (1931). In C. Hartshorne, P. Weiss, & A. Burks (Eds.). *Collected papers of Charles Sanders Peirce* (Vol. 8). Cambridge, MA: Harvard University Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353–363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*, 304–308.

Quelhas, A. C., Juhos, C., & Johnson-Laird, P. N. (2009). *The modulation of conditional assertions and its effects on reasoning*. Instituto Superior de Psicologia Aplicada.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Rips, L. J. (2002). Reasoning. In D. Medin (Ed.), *Stevens' handbook of experimental psychology, Vol. 2: Memory and cognitive processes* (3rd ed., pp. 317–362). New York: John Wiley.

Rosch, E., & Mervis, C. B. (1975). Family resemblance. Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Santamaria, C., & Johnson-Laird, P. N. (2000). An antidote to illusory inferences. *Thinking and Reasoning, 6*, 313–333.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1411–1436.

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.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive reasoning and learning. *Trends in Cognitive Sciences, 10*, 309–318.

Yang, Y., & Johnson-Laird, P. N. (2000). How to eliminate illusions in quantified reasoning. *Memory and Cognition, 28*, 1050–1059.