Cognitive Psychology 63 (2011) 34–59

Contents lists available at ScienceDirect

Cognitive Psychology

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

Mental models of Boolean concepts

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

[a] Department of Psychology, University of Pennsylvania, 3720 Walnut St., Solomon Lab Bldg., Philadelphia, PA 19104, United States
[b] Department of Psychology, Princeton University, Princeton, NJ 08540, United States

ARTICLE INFO

Article history:
Accepted 15 April 2011
Available online 2 June 2011

Keywords:
Concepts
Mental models
Boolean algebra
Reasoning

ABSTRACT

Negation, conjunction, and disjunction are major building blocks in the formation of concepts. This article presents a new model-based theory of these Boolean components. It predicts that individuals simplify the models of instances of concepts. Evidence corroborates the theory and challenges alternative accounts, such as those based on minimal descriptions, algebraic complexity, or structural invariance. A computer program implementing the theory yields more accurate predictions than these rival accounts. Two experiments showed that the numbers of models of a Boolean concept predict the difficulty of formulating a description of it. As mental models may also underlie deductive reasoning, the present theory integrates two hitherto separate areas of investigation.

1. Introduction

Concepts are the atoms of thought, and so they lie at the heart of cognitive science (Fodor, 1994). Individuals begin to acquire concepts in infancy and continue to acquire and to devise novel concepts throughout their lifetimes (Medin, Lynch, & Solomon, 2000; Medin & Smith, 1984). One way in which they create a new concept is to use existing ones in various combinations. A description such as *persons over the age of 18 years* defines a concept by using a property, *persons*; a relation, _ *over* _; a special sort of relation (known as a "function") that takes a single argument, *the age of* _; and a number of years, *18*. The focus of the present paper, however, is the use of various logical operations to combine existing concepts. These operators are negation (*not*), conjunction (*and*), and disjunction (*or*), and other operators that can be defined in terms of them. The operators make up what is known as a Boolean algebra, and they are a major building block for human concepts.

* Corresponding author.
E-mail address: ggoodwin@psych.upenn.edu (G.P. Goodwin).

0010-0285/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2011.04.001

Once individuals have a set of properties and relations at their disposal, they can construct novel concepts using Boolean operators. For example, a summons in a civil case can be served by *persons over the age of 18 years and not a party to the case* (Federal Civil Judicial Procedure and Rules, 1994, p. 19). This example illustrates the use of conjunction (and) and negation (not) in the formation of a new concept. We emphasize that we use the expression "Boolean concepts" to refer to concepts that are formed by Boolean operations on existing concepts, which themselves may depend on properties or relations that cannot be defined in a Boolean way. For example, no Boolean definition of *tall* is feasible, because the concept depends on a relation to a norm (Miller & Johnson-Laird, 1976, p. 325). But, individuals often form and communicate concepts, such as *tall and handsome*, which are Boolean combinations of existing concepts. There appear to be no constraints on the concepts that can enter into Boolean combinations, e.g., *rich Democrats or poor Republicans*; *buildings which are not public buildings or warehouses; a variable that is random and normally distributed.*

This paper concerns two key questions about Boolean concepts: given a description of a Boolean concept, how do individuals envisage its instances? And conversely, how do they learn to categorize and to describe its instances? In each case, we need to know what sorts of concept cause difficulty. The focus on the present paper is on the second question about the acquisition and description of Boolean concepts. Early psychological studies made progress in constraining the answer to this question, and a general theory appeared to have provided a definitive answer to it (Feldman, 2000). The theory was based on plausible assumptions, but recent studies have rendered it less convincing than it seemed at first (Lafond, Lacouture, & Mineau, 2007; Mathy & Bradmetz, 2004; Vigo, 2006). Our aim in the present article is accordingly to offer a new theory that answers this second question and that also answers the first question. The theory is based on mental models, and in this paper we report evidence supporting it.

The plan of the paper is straightforward. The first part describes Boolean concepts and previous psychological investigations of them. The second part reviews the theory of mental models, and shows how it can be extended to Boolean concepts. The third part describes the evidence corroborating the theory. And the final part draws some general lessons about concepts.

2. Studies of Boolean concepts

Consider again the description:

Persons over the age of 18 years.

The description has a meaning that all speakers of English grasp, and this meaning refers to a set of individuals. In logical terminology, the meaning is the *intension* of the description, that is, it is the concept itself, and the *extension* of the concept is the set of its possible instances to which the meaning refers. In logic, *person*, *building*, and *flower*, all denote properties, and so too do the values of variables, such as the color *red*, the shape *square*, and the texture *mottled*: they can all be treated as taking a single argument referring to what has the values. In contrast, relations such as: *over, bigger than*, and *between*, hold for two or more arguments (see e.g., Goodwin & Johnson-Laird, 2005; Halford, Wilson, & Phillips, 1998; Peirce, 1931–1958). Given a set of properties and relations, we can create concepts by establishing relations between properties, e.g.: *flowers bigger than weeds*, and *things to take on a picnic* (Barsalou, 1987). And, as we have already illustrated, we can construct complex concepts using Boolean operators:

Round, and red or mottled.

We reserve the term "simple" for concepts that do not contain Boolean operators, and the term "Boolean" for concepts that do contain them.

All Boolean operators can be defined in terms of a single primitive: *nand*, i.e., *not both _ and _*, e.g., *not red* can be defined as *red nand red*. But, psychologists have adopted three Boolean operators as primitive for concepts: *not*, *and*, and *or*. *Or* is often ambiguous in everyday English: *a or b* in its inclusive sense allows that both *a* and *b* may be the case, and so it is consistent with three possibilities: *a* and *not b*, *not a* and *b*, and both *a* and *b*. In contrast, *or* in its exclusive sense rules out the possibility of both *a* and *b*, and so it is consistent with only two possibilities. As a primitive in Boolean algebra, *or* is inclusive, and henceforth in this article, the term "disjunction" refers to an inclusive disjunction unless

otherwise stated. The three primitives can be used to define other operators. For example, exclusive disjunction, which henceforth we refer to as *or else*, can be defined as follows:

a or else *b* $=_{\text{def}}$ (*a* or *b*) and not (*a* and *b*).

An electrical circuit that yields an output for *a or else b*, and a separate output for *a and b*, is known as a half-adder, because it computes the addition of two single binary digits and the carry, if any. As readers should have gathered by now, Boolean concepts are powerful. Their realization as half-adders in silicon chips suffices for the central processing units of modern computers; they lie at the heart of genetic algorithms (see, e.g., Holland, Holyoak, Nisbett, & Thagard, 1986); and they underlie theorems about what can, and cannot, be learned in a tractable way (Valiant, 1984).

A large number of the concepts of daily life are based on Boolean operators. For instance, a strike in baseball can be defined roughly as follows: The ball must be swung at *and* missed *or* it must pass above the knees *and* below the armpits *and* over home plate without being hit *or* the ball must be hit foul *if* there are *not* two strikes (see Murphy, 2002, p. 16). This concept is replete with Boolean operators. Similarly, being trapped "leg before wicket" (*LBW*) in cricket similarly contains a large number of Boolean operators. A batsman is out *LBW*: *if* the ball hits the batsman (*and* does *not* first hit the bat *or* a hand holding the bat) *and* would have hit the wicket, but *not if* it pitches outside leg stump, *or* it strikes the batsman outside off-stump *and* the batsman was attempting to play a stroke, *or* a no-ball is called. Likewise, many of the diagnostic criteria in the Diagnostic and Statistical Manual of Mental Disorders rely on Boolean operators. For instance, one of the criteria for a major depressive episode is that a person must have experienced either a depressed mood *or* loss of interest or pleasure (American Psychiatric Association [DSM-IV-TR], 2000). Everyday less technical concepts also depend on Boolean operators such as the concept of someone's being tall *and* handsome, the concept of cruel *and* unusual punishment, or the concept of dogs *not* on a leash.

Arguments exist that aim to show that the logic of everyday concepts cannot depend solely on Boolean operators, because the underlying representation of concepts, and indeed their underlying intensions depend on graded prototypes (see e.g., Hampton, 1979; Posner & Keele, 1968; Posner & Keele, 1970; Rosch & Mervis, 1975; Smith & Medin, 1981), exemplars (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986; Nosofsky & Palmeri, 1997) relations (Miller & Johnson-Laird, 1976), or some other alternative. For example, the prototypes for everyday concepts such as a chair or a table consist of some combination of default features. How typical an item is of a particular category (and more controversially, whether an item is a member of a category) is determined by its resemblance to the prototype (i.e., the default features), and such resemblance is not computed in a Boolean way. Yet, Boolean operators play an important role even according to such views. The set of default values of a particular concept must be organized in some way, and typically cannot be grasped without access to a Boolean system. For instance, a plausible set of default values for the concept of a table would be: a worktop *and* something supporting it. A plausible set of default values for the concept of eating would be: ingesting, chewing, *and* swallowing something. Moreover, Boolean operators figure in the typical *extensions* of concepts. A typical drink at a bar is wine *or* beer, a typical household pet is a cat *or* a dog, a typical snack at the movies is popcorn *or* candy. Each of these extensions makes use of a Boolean operator.

The upshot is that psychologists need to understand how individuals represent and acquire concepts that consist of Boolean operators. Early psychological studies established that conjunctive concepts were easier to learn than disjunctive concepts (e.g., Bruner, Goodnow, & Austin, 1956), and this finding was widely corroborated (e.g., Conant & Trabasso, 1964; Haygood & Bourne, 1965). Similarly, the number of relevant variables instantiated in a concept (from one to three in these experiments) increases the difficulty of acquiring it in a roughly linear way (e.g., Bulgarella & Archer, 1962; Walker & Bourne, 1961). These studies are precursors to more recent work on relational complexity (Halford et al., 1998) and on the role of relations in concepts (Goodwin & Johnson-Laird, 2005).

Neisser and Weene (1962) extended the early treatment of Boolean concepts. They assumed that the operations of negation, conjunction, and disjunction, were primitive, and analyzed the relative complexity of other sorts of Boolean concepts in terms of them. The analysis yielded three levels of complexity: at the simplest level are what are often referred to as "literals", which are concepts defined in terms of the value of a single variable (e.g., the presence or absence of a property); one level

up are conjunctions, disjunctions, and conditionals (which can be defined in terms of disjunction); and at the most complex level are exclusive disjunctions (defined in the way we illustrated earlier), and biconditionals (defined in a similar way). Given the concepts that Neisser and Weene postulate as primitive, the three levels of concepts differ in the lengths of their minimal descriptions, e.g.:

Level 1: *not a*.
Level 2: *a or b*.
Level 3: (*a or b*) *and not*(*a and b*).

In retrospect, as Feldman (2000) pointed out, what is striking about Neisser and Weene's analysis is that it provides a putative measure of conceptual complexity – the length of a minimal description of a concept. One exception, however, is that conjunction is easier than disjunction, even though they have minimal descriptions of the same length. Moreover, given a set of primitives that includes *exclusive* disjunction, *a or else b*, minimal descriptions no longer predict difficulty.

In a highly influential paper, Shepard, Hovland, and Jenkins (1961) investigated the acquisition of Boolean concepts defined in terms of three binary variables in which there are four instances and four non-instances of each concept. There are six logically distinct sorts of concept in this domain, and examples of them are shown in Fig. 1: the black dots represent instances of each concept. Table 1 presents the descriptions of these concepts – henceforth the "Shepard" concepts. In two of Shepard and his colleagues' experiments, the participants had to learn how to classify each of the eight possible instances according to the six concepts. The number of errors, trials to criterion, and other measures, showed a partial trend in increasing difficulty: I < II < III, IV, V < VI. This trend, which we refer to as the "Shepard" trend, has been well replicated (e.g., Feldman, 2000; Love, 2002; Nosofsky, Gluck, Palmeri, & McKinley, 1994; Smith, Minda, & Washburn, 2004), and it poses a puzzle for researchers to solve. Why should the Shepard trend hold?

Shepard and his colleagues offered two informal speculations about the cause of the trend. Like Neisser and Weene (1962), they suggested that the length of a concept's minimal description predicts its difficulty. They also suggested that the trend depends in part on the number of variables in a concept: concept I is based on a single variable, concept II is based on two variables, and concepts III through VI are based on three variables. Garner (1962) argued that the difficulty of the concepts

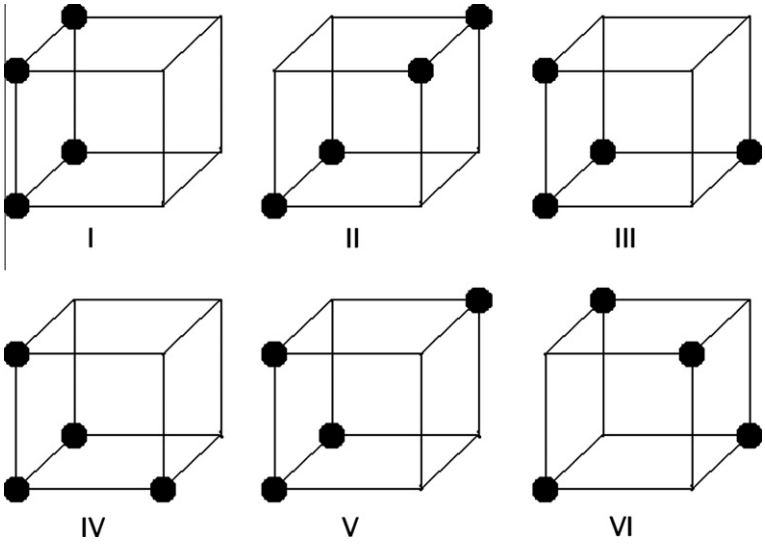

Fig. 1. Examples of the six sorts of concept used by Shepard et al. (1961). The right–left dimension corresponds to *a* and *not a*, the top-down dimension corresponds to *b* and *not b*, and the back-front dimension corresponds to *c* and *not c*. The black blobs represent the instances of the concepts.

Table 1
The six concepts in Shepard et al. (1961) and their descriptions (and lengths in parentheses) according to Feldman's (2000) heuristics, an algorithm for correct minimal descriptions, and an algorithm for correct minimal descriptions using exclusive disjunctions (*or else*).

Concept	Feldman's heuristic	Correct minimal descriptions	*Or else* version
I	*not a* (1)	*not a* (1)	*not a* (1)
II	(*a and b*) *or* (*not a and not b*) (4)	(*a and b*) *or* (*not a and not b*) (4)	*a or else b* (2)
III	(*not a and not* (*b and c*)) *or* (*a and* (*not b and c*)) (6)	(*not a and not c*) *or* (*not b and c*) (4)	(*not a and not c*) *or* (*not b and c*) (4)
IV	(*not a and not* (*b and c*)) *or* (*a and* (*not b and not c*)) (6)	(*not c or* (*not a and not b*)) *and* (*not a or not b*) (5)	(*not c or* (*not a and not b*)) *and* (*not a or not b*) (5)
V	(*not a and not* (*b and c*)) *or* (*a and* (*b and c*)) (6)	(*not a and not* (*b and c*)) *or* (*a and* (*b and c*)) (6)	(*not b or not c*) *or else a* (3)
VI	(*a and* ((*not b and c*) *or* (*b and not c*))) *or* (*not a and* ((*not b and not c*) *or* (*b and c*))) (10)	(*a and* ((*not b and c*) *or* (*b and not c*))) *or* (*not a and* ((*not b and not c*) *or* (*b and c*))) (10)	(*a or else not b*) *or else c* (3)

depends on the extent to which they have to be expressed as combinations of variables and values rather than as the value of a single variable, and he related this difficulty to the transmission of multivariate information. Likewise, Hunt, Marin, and Stone (1966, p. 134) analyzed difficulty in terms of the complexity of the decision trees that categorize the instances and non-instances of the concepts. A decision tree consists in a series of yes–no questions at nodes that classify the instances and non-instances of a concept in a parsimonious way, e.g., just a single node – a question of whether or not *a* is present in an instance – suffices for concept I, whereas the tree for concept II has three nodes, and the tree for concept VI has seven nodes (see Hunt & Hovland, 1960). Decision trees were developed with great sophistication in Artificial Intelligence (AI) by the introduction of information theory to govern the heuristics that form them (Quinlan, 1986). In our studies below, we examined how well decision trees matched the data, but they fared rather worse than recent theories, and so we spare readers any further discussion of them. Other approaches have also been examined in AI (Mitchell, 1997). But, we will not pursue these methods any further because they are much more sophisticated than humans (for a review, see, e.g., Russell & Norvig, 2003, chap. 19). They postulate a "version space" that holds all currently feasible hypotheses about a concept so that it is never necessary to backtrack and to re-examine old instances. In contrast, human concept learners tend not to hold alternative hypotheses in mind, and backtrack if the option is available to them in experiments (e.g., as in the experiments that we report below, and in Whitfield, 1951, who also estimates the "half life" of the memory for instances).

After these initial investigations, psychologists began to focus on graded concepts that do not have sharp boundaries, on prototypes (Hampton, 1979; Posner & Keele, 1968; Posner and Keele, 1970; Rosch & Mervis, 1975; Smith & Medin, 1981), on exemplars (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986; Nosofsky & Palmeri, 1997; Vandierendonck, 1995), and on the role of knowledge in the acquisition of concepts (Keil, 1989; Murphy & Medin, 1985). However, Nosofsky, Gluck, et al. (1994) replicated the Shepard trend, and examined which of several computational models of classification would generalize to fit their data. The most successful was ALCOVE (Kruschke, 1992), which is an extension of a previous exemplar-based model of classification (Medin & Schaffer, 1978; Nosofsky, 1986). It is a three-layered, feed-forward connectionist network, which represents concepts by storing individual exemplars in memory, each of which is represented as a point in a multi-dimensional space. However, in a separate paper, Nosofsky, Palmeri, and McKinley (1994) showed that a different model, RULEX, which is simpler and uses less memory, can account just as well for the partial trend. It represents categories by forming simple logical rules, and storing occasional exceptions to those rules. Hence, concept I should be easiest because it depends on a single variable; concept II should be next in difficulty because it depends on a conjunction of two variables; concepts III–V should be next in difficulty because they depend on a single variable plus an exception; and concept VI should be most difficult because it depends on the values of all three variables. A more recent model, SUSTAIN, is also capable of accounting for the Shepard trend (Love, Medin, & Gureckis, 2004).

SUSTAIN is also a connectionist model, but it is essentially a multiple prototype model. It operates by creating different "clusters" in multi-dimensional space in order to represent conceptual structure. It predicts that concepts should be more difficult to learn with an increase in the number of clusters in their representations. And this account also explains the Shepard trend – the modal number of clusters required to represent concept I is 2, the modal number for concept II is 4, the modal number for concepts III–V is 6, and the modal number for concept VI is 8.

Feldman (2000, p. 630) proposed a simple law governing the difficulty of acquiring Boolean concepts, which echoes some of the ideas we described earlier: 'the subjective difficulty of a concept is directly proportional to its minimal Boolean description length (the length of the shortest logically equivalent propositional formula) – that is, to its logical incompressibility'. All Boolean concepts have infinitely many descriptions, e.g., these two descriptions are of the same concept:

(*a* and *b*) or (*a* and not *b*) or (not *a* and *b*)

and:

(*a* or *b*)

Clearly, the second description is minimal: it cannot be compressed any further, because the concept depends on two variables, and the description refers to each of them only once. The length of a description is usually measured in terms of the number of variables it contains, i.e., values of variables, whether affirmative or negative. Hence, *not* (*a and not b*) has a length of two, because it contains the values of two literals: *a*, *not b*. As a corollary, negation has no effect on length, and so *a* and an equivalent, such as: *not not not not a*, have the same length on this metric. An alternative metric using the number of binary connectives correlates perfectly with the number of variables.

Various algorithms exist for finding a minimal description of any Boolean expression given a particular vocabulary of operators, such as *not*, *and*, and *or*, but the task is not computationally tractable, i.e., as concepts depend on an increasing number of variables and instances so ultimately it becomes infeasible to compute their minimal descriptions. Feldman (2000) relied on a set of heuristic procedures, which yielded descriptions with the following lengths for the Shepard concepts: I: 1, II: 4, III: 6, IV: 6, V: 6, VI: 10. Table 1 presents the concepts, and these lengths, which predict the psychological difficulty of the concepts. Feldman adds an auxiliary assumption, which he refers to as "parity". Apart from tautological and self-contradictory concepts, all concepts divide the universe of possibilities into instances and non-instances of the concept. Parity is the notion that concepts should be easier to learn when the instances comprise the smaller set – an idea which originates from the finding that it is easier to learn concepts from their instances than from their non-instances (Hovland & Weiss, 1953; Hunt, 1962). Conjunction and disjunction have minimal descriptions of the same length given that they are both primitives in the given vocabulary, but parity predicts that conjunction (one instance) should be easier than disjunction (three instances). As Feldman points out, minimal description length and parity together account for the Shepard trend and for about 50% of the variance in Feldman's (2000) own data on the acquisition of 76 new sorts of concept, which he created by varying both the number of variables and the number of instances in each concept. Henceforth, we refer to these 76 concepts as the "Feldman" concepts. At first sight, the key questions about Boolean concepts appeared to have received definitive answers in terms of minimal descriptions. But, as we now show, this putative solution runs into three problems.

The first problem with minimal description length is to justify the set of primitive connectives that are allowed in formulating descriptions. A standard defense of *not*, *and*, and *or*, is that they have been conventional since the work of Neisser and Weene (1962). But, why should we exclude *or else* (exclusive disjunction)? People reason better from an exclusive disjunction than from an inclusive disjunction (for a review, see Evans, Newstead, & Byrne, 1993). But, if *or else* is used as an additional primitive, the length of the minimal descriptions of the Shepard concepts changes in a striking way: I: 1, II: 2, III: 4, IV: 5, V: 3, VI: 3 (see Table 1). These lengths no longer predict the difficulty of the concepts. For instance, concept V has the following description if exclusive disjunction is excluded: (*not a and not* (*b and c*)) *or* (*a and* (*b and c*)). But, it has a much shorter minimal description if exclusive disjunction

is included: (*not b or not c*) *or else a*. The use of exclusive disjunctions does not have a uniform effect on the length of minimal descriptions across all Boolean concepts. It affects some concepts, but not others. Hence, the predictions of minimal description theory depend crucially on whether or not exclusive disjunctions occur in them; and the theory offers no principled account of why they should be excluded.

The second problem with minimal descriptions occurs even if we grant the conventional choice of primitives. Feldman's heuristics failed to find the correct minimal descriptions for certain concepts. Mathy and Bradmetz (2004) pointed out an error for concept III, and Vigo (2006) pointed out this error and one for concept IV. We wrote a computer program (in Common Lisp) that finds minimal descriptions by recursively factoring models of concepts, and it confirmed these errors. And, indeed, Lafond et al. (2007) demonstrated that there exist descriptions that are more minimal for many of Feldman's (2000, 2003a, 2003b) 76 concepts. Table 1 shows the correct minimal descriptions for the Shepard concepts, and their minimal lengths are as follows: I: 1, II: 4, III: 4, IV: 5, V: 6, VI: 10. These true lengths predict that II and III should be of comparable difficulty, and easier than IV, which should be easier than V, which in turn should be easier than VI. The evidence fails to corroborate this prediction as neatly as it corroborated the original prediction based on putatively minimal descriptions.

The third problem with minimal description length is that little evidence exists that naïve individuals try to formulate minimal descriptions of concepts, and some evidence exists to the contrary (see the results of our experiments below). Moreover, individuals can carry out the task of categorizing instances of a concept without attempting to formulate its minimal description. In the light of these difficulties, Feldman (2006), Vigo (2009), and the present authors, have developed new theories, which we outline in turn.

Feldman (2006) developed a new theory of complexity, which rests on a formalization of "algebraic" complexity rather than the length of minimal descriptions. In this updated model, the complexity of a concept is driven by its decomposition into a set of underlying regularities. The basic idea is that any concept can be decomposed into a series of more basic concepts. At the most basic or "atomic" level, are two kinds of simple concepts: those that consist of a single constant value of a variable, and those that consist of an implication between the values of two variables (pairwise implication). These two basic sorts of concept can be algebraically combined in order to represent more complex linear concepts. This model can then be generalized further to capture any discrete-featured concept, not just those that are linear. This step is based on an analysis of a concept's "power spectrum", i.e., an analysis of the complexity of the various rules that algebraically combine to capture the concept. Thus, any concept can be decomposed into a set of underlying rules, each of differing degrees of complexity, depending on the number of variables that they instantiate. More complex rules represent more idiosyncratic patterns of data. The weighted mean of the complexity of these underlying rules then provides an overall index of the *algebraic complexity* of the concept. The model yields a complexity index for any Boolean concept, and it is not limited to concepts with only binary variables or to concepts with variables that all have the same number of values.

Feldman (2006) argued that this model is more plausible than his earlier theory of minimal descriptions, because it embodies the generally accepted principle that learners extract statistical regularities from data sets. He notes a similarity between his model and Bayesian nets, and with recent hybrid models of categorization according to which learners extract a concept's central rule, along with exceptional cases that do not fit the rule (e.g., Nosofsky, Palmeri, et al. 1994). He also notes that his model is an "analytic counterpart" of Love et al.'s (2004) SUSTAIN model (Feldman, 2006, p. 355). However, it is not clear from Feldman's account how learners represent concepts. Perhaps they represent each of the lower order rules for a concept, but this representation seems too demanding to be plausible. Nevertheless, the model is more powerful than Feldman's (2000) earlier minimal description theory. Its single parameter of algebraic complexity accounts for 50.37% of the variance in of the acquisition of the 76 Feldman concepts – substantially exceeding the earlier mark achieved by minimal descriptions alone (i.e., without the addition of a parity variable)[1].

[1] Our own simulations, based on the Matlab code on Feldman's web-page, produced a slightly lower fit, $R^2 = .493$, $\beta = -.702$, $t(74) = -8.48$, $p < .001$. In either case, the fit is high.

One further impressive feature of this model is its ability to account for "typicality" effects in learning concepts. Algebraic complexity allows each instance of a concept to be graded in terms of its typicality. The more underlying regularities of a concept that an instance satisfies, the more typical it is. Feldman analyzed the learnability of each instance of the 76 concepts he studied – a total of 1072 separate instances – and algebraic complexity accounted for 12% of the variance in the learning of these individual instances.

Vigo (2009) developed an alternative theory yielding a measure of the relative invariance of the instances of a concept. The measure depends on a series of computations, which we outline here. The first step is to compute the degree to which the instances of a concept remain the same as the values of variables are changed in every instance. This step is akin to the *partial derivative* of a function, except that the variables of the concept have only two possible values. Consider Shepard's concept III, which has these instances in which "¬" denotes negation (see Table 3 below):

$a \quad \neg b \quad c$
$\neg a \quad b \quad \neg c$
$\neg a \quad \neg b \quad c$
$\neg a \quad \neg b \quad \neg c$

To construct the partial derivative for *a*, we negate *a* in each instance:

$\neg a \quad \neg b \quad c$
$a \quad b \quad \neg c$
$a \quad \neg b \quad c$
$a \quad \neg b \quad \neg c$

We then compute the exclusive disjunction of the original instances with this new set, which yields:

$a \quad b \quad \neg c$
$a \quad \neg b \quad \neg c$
$\neg a \quad b \quad \neg c$
$\neg a \quad \neg b \quad \neg c$

The partial derivatives for *b* and *c* yield analogous "perturbed" sets of instances.

The next step is to compute the *norm*, which is the number of instances common to both the original and perturbed instances divided by the number of instances in the original category. In the example, the following instances are common to both the original and the perturbed ones for *a*:

$\neg a \quad b \quad \neg c$
$\neg a \quad \neg b \quad \neg c$

and so the norm equals 1/2. The calculations for the other two partial derivatives also yield norms of 1/2. The *manifold* for concept III is accordingly (1/2, 1/2, 1/2). A concept with the highest invariance has a zero manifold of (0, 0, 0), and so the degree of invariance of concept III is its Euclidean distance from the zero manifold. (The Euclidean distance between objects in multi-dimensional space is the square root of the sum of their squared distances on each dimension.) The Euclidean distance between the manifold for III and the zero logical manifold is: $\sqrt{(1/2-0)^2+(1/2-0)^2+(1/2-0)^2} = \sqrt{(.75)} = .866$. The final computation is of the *structural complexity* of the concept, which takes into account the number of instances in a concept, n, and is equal to: $n/(1 + \text{degree of invariance})$. The structural complexity of concept III is therefore: $4/(1 + 0.866) = 2.143$. Vigo reports that the structural complexity of the concepts in Table 1 predicts the difficulty of learning them. He also reports that his measure accounts for 42% of the variance in Feldman's (2000) data on the learning of the 76 concepts.

Vigo's (2009) account of the invariance of concepts, as he acknowledges, does not specify how individuals learn concepts. He suggests only that cognitive processes could detect invariances by comparing a set of instances to the set yielded by the partial derivative of each variable. In

contrast, we have formulated an alternative theory of the mental processes underlying the acquisition of Boolean concepts, and what causes difficulty in the process. We now turn to this alternative.

3. Models of concepts

When the values of variables in a Boolean expression, such as: *not a or else b*, are propositions, then the result is, not a concept, but an assertion that is true or false, such as the exclusive disjunction: *The switch isn't on or else the battery is dead*. Individuals are able to reason from such assertions; and the theory of mental models – henceforth, the "model" theory – offers an account of how they do so (see, e.g., Johnson-Laird, 2006; Johnson-Laird & Byrne, 1991; Johnson-Laird, Byrne, & Schaeken, 1992). The logical meaning of the assertion above can be represented in a truth table, as shown in Table 2. The table illustrates the four possible assignments of truth to the two atomic propositions and the resulting truth of the compound assertion containing them. Naïve reasoners, as Osherson (1974–1976) argued, do not rely on truth tables. According to the model theory, they rely instead on a representation that captures the possibilities consistent with an assertion's truth. The theory postulates that individuals represent each possibility consistent with the premises in a separate *mental model*, and that a conclusion is deductively valid – it follows of necessity – if it holds in all the models of the premises. Each model represents a state of affairs in as iconic a way as possible, i.e., its structure corresponds to the structure of the possibility. Models capture what is common to a set of possibilities, which allows individuals to construct the simplest possible models of assertions. In particular, mental models embody a principle of *truth*: they represent only what is possible given the truth of assertions, and within these possibilities they represent literals, simple affirmative or negative propositions in a description, only when they are true. Hence, they represent the exclusive disjunction above (*the switch isn't on or else the battery is dead*) in two mental models of the possibilities in which it is true. The following diagram denotes these two models, though actual mental models aren't strings of words but representations of situations in the world:

¬switch on
battery dead

Information about what is false is accordingly ephemeral, but if individuals hold onto it, then they can flesh out their mental models into *fully explicit* models that do represent what is false in a possibility, as shown here:

¬switch on ¬battery dead
switch on battery dead

where a true negation is used to represent a false affirmative, and a true affirmative is used to represent a false negation. These two models correspond to the fourth and first rows of the truth table in Table 2.

The evidence corroborating this account of deductive reasoning has been described elsewhere, and it falls into three main categories. First, inferences that call for multiple models are more difficult than inferences that call for only a single model: they take more time and are more likely to elicit errors, e.g., inferences from exclusive disjunctions are easier than inferences from inclusive disjunctions (e.g., Bauer & Johnson-Laird, 1993). Second, the meanings of clauses and the situations to which they refer modulate the interpretation of Boolean connectives and influence the conclusions that individuals draw from them – as a consequence the meaning of, say, conditionals transcends a purely

Table 2
The truth table for the exclusive disjunction: *The switch isn't on or else the battery is dead.*

The switch is on	The battery is dead	The switch is not on or else the battery is dead
True	True	True
True	False	False
False	True	False
False	False	True

Table 3
The instances of the six Shepard et al. (1961) concepts in Table 1 and their simplified mental models.

Concept number	Instances of the concept			Their simplified mental models		
I	¬*a*	*b*	*c*	¬*a*		
	¬*a*	*b*	¬*c*			
	¬*a*	¬*b*	*c*			
	¬*a*	¬*b*	¬*c*			
II	*a*	*b*	*c*	*a*	*b*	
	a	*b*	¬*c*	¬*a*	¬*b*	
	¬*a*	¬*b*	*c*			
	¬*a*	¬*b*	¬*c*			
III	*a*	¬*b*	*c*	¬*a*		¬*c*
	¬*a*	*b*	¬*c*		¬*b*	*c*
	¬*a*	¬*b*	*c*			
	¬*a*	¬*b*	¬*c*			
IV	*a*	¬*b*	¬*c*	¬*a*	*b*	¬*c*
	¬*a*	*b*	¬*c*	¬*a*	¬*b*	*c*
	¬*a*	¬*b*	*c*		¬*b*	¬*c*
	¬*a*	¬*b*	¬*c*			
V	*a*	*b*	*c*	*a*	*b*	*c*
	¬*a*	*b*	¬*c*	¬*a*	*b*	¬*c*
	¬*a*	¬*b*	*c*	¬*a*	¬*b*	
	¬*a*	¬*b*	¬*c*			
VI	*a*	*b*	¬*c*	*a*	*b*	¬*c*
	a	¬*b*	*c*	*a*	¬*b*	*c*
	¬*a*	*b*	*c*	¬*a*	*b*	*c*
	¬*a*	¬*b*	¬*c*	¬*a*	¬*b*	¬*c*

truth functional account (e.g., Johnson-Laird & Byrne, 2002). Third, the principle of truth leads to systematic but erroneous conclusions, that is, so-called "illusory" inferences (e.g., Johnson-Laird & Savary, 1999).

The theory of mental models extends in a natural way to the representation of concepts. Each instance of a concept can be represented as a conjunction of properties or relations, though some of them may be negative. The mental representation of the extension of a concept is accordingly a disjunction of its possible instances. For example, by analogy with the exclusive disjunction above, the concept *not rich or else Republican* has these fully explicit models:

¬rich ¬Republican
rich Republican

Each row represents one sort of possible instance of the concept consisting of a *conjunction* of the values of the relevant variables, and the rows are related *disjunctively*: the possibilities are alternative instances of the concept.

The model theory of concepts is based on a single fundamental principle that yields testable predictions. It concerns how individuals learn to categorize instances of a concept. The simplest process is merely to commit to memory each instance (or exemplar) of a concept (see, e.g., Medin & Smith, 1984), but for Boolean concepts this process can impose a considerable load on memory. The first principle of the theory accordingly embodies a simple idea: individuals detect those variables that are irrelevant given the values of other variables, and subsequently ignore the irrelevancies. The theory accordingly follows in Feldman's (2000) footsteps, but instead of minimizing the description of concepts, it simplifies the representation of their instances, reducing the number of models required to represent them. As we will see, the two accounts yield divergent predictions. We summarize the fundamental principle as follows:

The principle of *simplifying models*: when individuals acquire a Boolean concept from its instances, they represent its instances and can reduce the load on memory by eliminating those variables that are irrelevant given the values of the other variables. Their aim is to reduce the overall number of mental models needed to represent the concept. Individuals can represent the set of non-instances instead of the set of instances, particularly when the number of instances exceeds the number of non-instances.

As an example, consider a concept that has these two instances:

$a \quad b \quad c$
$a \quad b \quad \neg c$

The properties a and b occur in both instances, and c occurs in one instance and $\neg c$ in the other. Whether or not c is present therefore has no effect on the fact that any instance of a and b is a member of the concept – the variable is irrelevant, and so the two instances above can be simplified to:

$a \quad b$

An irrelevant variable is thus one whose value, such as the presence or absence of a property, has no effect on certain instances of the concept. The mental processes of acquiring the instances of a concept are accordingly those that lead to the elimination of variables irrelevant to instances of the concept. We spell out the details of this process in the section below describing our computer implementation of the theory.

The simplification of models is not the same as the minimization of descriptions. The two theories make divergent predictions. The simplest way to grasp the divergence is to consider the lengths of the minimal descriptions of the Shepard concepts shown in Table 1 and the numbers of models for these concepts shown in Table 3. One reason for the divergence is that the number of models of a concept does not take into account the numbers of literals in each model. But, even if one counts up the number of literals in the models in Table 3, they diverge from the numbers of literals in minimal descriptions in Table 1, for either Feldman's heuristic estimates or the correct minimal descriptions.

4. A computer model of the process of simplifying instances of concepts

To clarify the consequences of the theory, especially for large sets of concepts, we wrote a computer program in Common Lisp to simplify the instances of concepts. Its source code and output for the Feldman concepts can be accessed online at: http://mentalmodels.princeton.edu/projects/boolean-concept-learning/. The input to the program is the complete set of instances of any Boolean concept, and the program constructs, if possible, a more parsimonious set of models using a procedure that eliminates irrelevant variables according to the fundamental principle described in the previous section. Table 3 shows the instances for each of the six Shepard concepts, and the program's output for each of them. As concept VI illustrates, some concepts cannot be simplified, and so according to the theory individuals have no option but to represent all of their instances in full.

As an illustrative example of the program, consider concept IV in Table 3, which has these instances:

$a \quad \neg b \quad \neg c$
$\neg a \quad b \quad \neg c$
$\neg a \quad \neg b \quad c$
$\neg a \quad \neg b \quad \neg c$

The program detects that the pair of properties $\neg b$ and $\neg c$ occur with both a and with $\neg a$, i.e., with both possible values, $\pm a$, and so it simplifies the instances to the following three:

$\neg a \quad b \quad \neg c$
$\neg a \quad \neg b \quad c$
$\neg b \quad \neg c$

There are alternative simplifications of the original instances in terms of other variables:

$$\begin{array}{lll} a & \neg b & \neg c \\ \neg a & \neg b & c \\ \neg a & & \neg c \end{array}$$

and:

$$\begin{array}{lll} a & \neg b & \neg c \\ \neg a & b & \neg c \\ \neg a & \neg b & \end{array}$$

Each of these simplifications yields the same reduction in the number of instances: the original four models are reduced to three. With other concepts, different simplifications can yield different numbers of instances, but the program searches for and returns a simplification with the fewest number of models. Thus, the number of models returned for each problem is constant across different runs of the program. Consider this case, in which the ellipsis refers to other instances in the concept:

$$\begin{array}{llll} a & b & c & \text{d} \\ a & b & c & \neg\text{d} \\ a & b & \neg c & \text{d} \\ a & b & \neg c & \neg\text{d} \\ \ldots & & & \end{array}$$

The variables $\pm c$ and $\pm d$ can be eliminated, because all four pairs of their possible joint values occur together with a and b, and so the result is:

$$\begin{array}{ll} a & b \\ \ldots & \end{array}$$

The program can make several successive simplifications. For example, consider concept II in Table 3, which has these instances:

$$\begin{array}{lll} a & b & c \\ a & b & \neg c \\ \neg a & \neg b & c \\ \neg a & \neg b & \neg c \end{array}$$

When a and b co-occur, the value of the variable, $\pm c$, is irrelevant, and so the program simplifies the concept to:

$$\begin{array}{lll} a & b & \\ \neg a & \neg b & c \\ \neg a & \neg b & \neg c \end{array}$$

In the two lower instances in which $\neg a$ and $\neg b$ co-occur, the variable $\pm c$ is again irrelevant, and so the program makes a second simplification, resulting in a final representation of the concept in two models:

$$\begin{array}{ll} a & b \\ \neg a & \neg b \end{array}$$

The program considers the possible orders in which simplifications can be made in order to find the best simplification possible. It therefore assumes that performance should be best predicted by an optimal simplification for each problem. Results of the program are shown in Tables 3 and 5. Table 1 in the supplementary materials on the web-link above show its output for the 76 Feldman concepts.

In searching for the simplest way to represent a concept, individuals may sometimes focus on the number of non-instances that define it rather than the number of instances. This switch is likely to happen when the instances of a concept vastly outnumber its non-instances. It should also be more likely when individuals are forced to create an explicit representation of the entire concept (a "summary representation"), as for instance when they are asked to describe a concept, as opposed to simply reproducing its instances. Consider the concept, *a or not b*, defined over three variables. It has the following six instances:

$$\begin{array}{rrr} a & b & c \\ a & b & \neg c \\ a & \neg b & c \\ a & \neg b & \neg c \\ \neg a & \neg b & c \\ \neg a & \neg b & \neg c \end{array}$$

but only two non-instances:

$$\begin{array}{rrr} \neg a & b & c \\ \neg a & b & \neg c \end{array}$$

The models of the instances simplify to two:

$$\begin{array}{rr} a & \\ \neg a & \neg b \end{array}$$

whereas the non-instances have only a single model:

$$\neg a \quad b$$

It is thus more economical to represent this concept in terms of a model of its non-instances, which alleviates the burden on working memory. Hence, the theory predicts that individuals can use this strategy for such concepts, although various factors should make this strategy more or less likely (cf. Whitfield, 1951).

The underlying principle governing simplifications is the elimination of irrelevant variables, and so the most important source of difficulty in acquiring a Boolean concept should be the number of models that result from this process of simplification. Hence, the intuition that individuals seek to reduce what they have to hold in mind – as suggested by Neisser and Weene, Feldman, and others – is correct. But, what individuals try to do is, not to minimize a Boolean description of the concept, but to reduce the number of mental models of its instances. Of course, not everyone invariably discovers the optimal simplification of the models of a concept, and so the output of the program is an ideal. Nevertheless,

Table 4
Results of the separate regression analyses predicting the accuracy of performance across the 76 concepts in Feldman (2000). All correlation coefficients are reliable at $p < .001$.

Theory	Results for the 41 'up' concepts		Results for all 76 concepts	
	Pearson's r	R^2	Pearson's r	R^2
Number of models	−.80	.63	−.75	.57
Boolean complexity; Feldman (2000)	−.68	.46	−.56	.31
Algebraic complexity; Feldman (2006)	−.64	.41	−.70	.49
Invariance; Vigo (2009)	−.80	.64	−.65 (approx.)	.42
Number of clusters, SUSTAIN; Love et al. (2004)	−.43	.19	−.48	.23

Table 5
The nine problems in Experiments 1 and 2, their mental models, their minimal descriptions (and their Boolean complexity values), their algebraic complexity values, and the percentages of accurate descriptions in Experiments 1 and 2. Problems are ordered according to the number of mental models they give rise to. The problems in Experiment 2 were identical except that the polarity of the *b* variable was reversed. Their models and descriptions can be derived simply by reversing the polarity of each *b* variable – from negative to affirmative, or from affirmative to negative – in both the models and the descriptions.

Problem number	Feldman Case Number	Instances	Models	Minimal descriptions without *or else*	Minimal descriptions with *or else*	Algebraic complexity	Accuracy (% correct): Experiment 1	Accuracy (% correct): Experiment 2
1	3[2]1	a ¬ b c a ¬ b ¬ c	a ¬ b	a and not b (2)	a and not b (2)	-1	96	100
2	3[4]2	a b c a b ¬ c ¬ a ¬ b c ¬ a ¬ b ¬ c	a b ¬ a ¬ b	(a and b) or (not a and not b) (4)	a or else not b (2)	0	82	89
3	3[2]1 (down parity)	a b c a b ¬ c a ¬ b c a ¬ b ¬ c ¬ a ¬ b c ¬ a ¬ b ¬ c	a ¬ a ¬ b	a or not b (2)	a or not b (2)	0	57	92
4	3[2]2	a b c a ¬ b ¬ c	a b c a ¬ b ¬ c	a and ((not b and not c) or (b and c)) (5)	a and (not b or else c) (3)	-0.5	79	100
5	3[2]3	a b ¬ c ¬ a ¬ b c	a b ¬ c ¬ a ¬ b c	((a and b) and not c) or ((not a and not b) and c)) (6)	(a or else not b) and (a or else c) (4)	-0.5	89	81
6	3[4]3	a b c a ¬ b c a ¬ b ¬ c ¬ a b c	a ¬ b b c	(a and not b) or (b and c) (4)	(a and not b) or (b and c) (4)	0	59	48
7	3[4]5	a b c a b ¬ c a ¬ b ¬ c ¬ a ¬ b c	a b a ¬ b ¬ c ¬ a ¬ b c	(a and not (not b and c)) or (not a and (not b and c)) (6)	a or else (not b and c) (3)	0.5	39	67
8	3[2]2 (down parity)	a b c a b ¬ c a ¬ b c a ¬ b ¬ c ¬ a b c ¬ a ¬ b ¬ c	a ¬ a b c ¬ a ¬ b ¬ c	a or ((not b and not c) or (b and c)) (5)	a or (not b or else c) (3)	1	64	58
9	3[2]3 (up parity)	a b c a b ¬ c a ¬ b ¬ c ¬ a b c ¬ a ¬ b c ¬ a ¬ b ¬ c	a b a ¬ b ¬ c ¬ a b c ¬ a ¬ b	not (((a and not b) and c) or ((not a and b) and not c))) (6)	(a or else not b) or (a or else c) (4)	1	68	100

Note. The symbol '¬' denotes negation. The first number in the Feldman case number column is the number of binary variables defining the concept, and the second number is the number of instances the concept gives rise to. The conjunctions of these two values define different "families" of concepts. The third number is simply a label for the distinct cases within each family, corresponding to the way Feldman (2000, 2003a) labeled the cases. The numbers in parentheses in the minimal descriptions column represent the lengths of each description (number of atoms referred to). 'Or' represents inclusive disjunction and 'or else' represents exclusive disjunction.

the program's simplifications should predict the difficulty of learning a concept. We now examine these predictions and compare them with those of alternative theories, both on a standard concept acquisition task, and on one that called for participants to formulate a description of a concept they had learned.

5. Empirical evidence

Several general theories of the difficulty of acquiring Boolean concepts exist, the most prominent of which are: ALCOVE (Kruschke, 1992), RULEX (Nosofsky et al., 1994), minimal descriptions (Feldman, 2000), SUSTAIN (Love et al., 2004), algebraic complexity (Feldman, 2006), and invariance (Vigo, 2009). Our task in the present section is to compare them with the theory of mental models. However, we do not consider ALCOVE (Kruschke, 1992), because it has already been compared unfavorably with algebraic complexity across a wide range of Boolean concepts (see Feldman, 2006). Nor do we consider RULEX, because it has no simple extension to the learning procedure of the Feldman concepts, in which all instances of each concept were displayed simultaneously rather than sequentially (Robert Nosofsky, personal communication, November 19, 2009). We therefore begin by comparing the

predictions of the model theory with those made by its four remaining rivals: minimal descriptions (Feldman, 2000), algebraic complexity (Feldman, 2006), invariance (Vigo, 2009), and SUSTAIN (Love et al., 2004).

In our assessment of SUSTAIN, we use the parameter values for fitting the Shepard et al. (1961) concepts reported in Love et al. (2004, see p. 31 for parameter values for attentional focus, cluster competition, decision consistency, and learning rate). We also rely on the modal number of clusters that SUSTAIN returns for each concept, since this number should predict learning difficulty (see e.g., Love, 2005). We used modal rather than mean clusters because doing so yielded a slightly better fit with Feldman's (2000) accuracy data.

Table 3 presents the mental models of the Shepard concepts. The numbers of mental models for each concept are as follows: I: 1, II: 2, III: 2, IV: 3, V: 3, VI: 4; and this order predicts the relative difficulty of acquiring the concepts reasonably well, $r(4) = .92$, $p < .01$. To predict difficulty perfectly, concept III would ideally give rise to three models rather than two. However, concept III is unique among the Shepard concepts in requiring two independent simplifications in order to arrive at two models, whereas each of the other concepts requires only a single simplification. The additional difficulty of III is likely to be a consequence of individuals making only one of the two simplifications, and therefore constructing three models instead of two. Nevertheless, the present theory does no worse in predicting the Shepard trend than its rivals: minimal descriptions, $r(4) = .92$, $p < .02$; algebraic complexity, $r(4) = .84$, $p < .05$; invariance, $r(4) = .85$, $p < .05$; and SUSTAIN's number of modal clusters, $r(4) = .99$, $p < .001$.

To test the model theory against these alternative accounts more rigorously, we ran our computer program on the 76 Feldman concepts, and compared its predictions with those of the alternative accounts. These concepts vary in both their number of variables and their number of instances, and were in the following families, where the first number is the number of binary variables defining a concept, and the second number is the number of instances in the concept: 3[2], 3[3], 3[4], 4[2], 4[3], 4[4]. For instance, a 3[2] concept is defined on three binary variables, such as color, shape, and size, and it has two instances out of the eight possible combinations of the three binary variables, e.g., the concept *red and square*. In Feldman's procedure, participants were presented with all of the instances and non-instances of each concept at once, with the instances spatially separated from the non-instances. This procedure differs from trial-by-trial procedures in which participants categorize each putative instance of a concept, one at a time (see e.g., Nosofsky, Gluck, et al., 1994; Shepard et al., 1961). Feldman (2000) did not observe the Shepard trend, but instead the following one: I < III < II < V < IV < VI. This trend may have resulted from the different procedure, but the deviation from the standard trend is slight. Hence, the data provide an informative test of the various theories.

In a series of correlational and regression analyses, we compared the model theory with minimal descriptions, algebraic complexity, invariance, and SUSTAIN. Minimal description lengths were obtained from Feldman (2000, 2003a),[2] algebraic complexity measures were obtained by implementing the Matlab suite on Jacob Feldman's web-site, structural invariance measures were obtained from Vigo (2009), the modal number of SUSTAIN clusters for each concept were obtained from code provided by Bradley Love, implemented in the programming language Python, and number of models were computed by our program. In Feldman's procedure, the instances of each concept were presented separately from, and above the non-instances, which should make the instances salient, and incline participants to focus on them. Thus, in our analyses, models were computed for the instances rather than the non-instances. We used Feldman's (2000) original minimal description lengths, which is conservative, since the predictive strength of minimal complexity is reduced if one corrects the errors in Feldman's estimates. The main results from these analyses are shown in Table 4.

We first examined performance over the 41 concepts that have "up" parity, i.e., that have no more instances than non-instances. Each theory reliably predicted accuracy, and number of models had the strongest predictive power, alongside structural invariance, accounting for 63% of the total variance in accuracy: number of models, $r(39) = -.80$ $p < .001$; minimal descriptions, $r(39) = -.68$, $p < .001$; alge-

[2] We used the non-minimal descriptions used in Feldman (2000, 2003a) rather than the more minimal descriptions reported in Lafond et al. (2007) since the original descriptions provided a better fit with Feldman's data (see Lafond et al., 2007).

braic complexity, $r(39) = -.64$, $p < .001$; invariance, $r(39) = -.80$, $p < .001$; modal number of clusters in SUSTAIN, $r(39) = -.43$, $p < .01$. In a comparison of these correlations, number of mental models was a reliably better predictor than minimal descriptions and modal clusters in SUSTAIN, William's ts (38) = 2.57, 6.74, respectively, both ps < .01. Number of models was also a marginally better predictor than algebraic complexity, William's $t(38) = 1.64$, $p < 0.06$, but it was not reliably better than structural invariance.

We then ran the same analyses for all 76 concepts, including those with "down" parity. Once again, number of models was the strongest predictor of accuracy, accounting for 57% of the variance in accuracy: number of models, $r(74) = -.75$, $p < .001$; minimal descriptions, $r(74) = -.56$, $p < .001$; algebraic complexity, $r(74) = -.70$, $p < .001$; modal number of clusters in SUSTAIN, $r(74) = -.48$, $p < .001$. A correlation is not reported for structural invariance, because Vigo (2009) does not provide invariance measures for the "down" concepts. However, he reports that the invariance metric accounts for approximately 42% of the variance across all 76 concepts, which therefore reflects a correlation of approximately $r = -.65$. Although number of models yielded the highest correlation with accuracy, its predictions did not differ reliably from algebraic complexity (William's $t(73) = 1.03$, $p = .15$). Algebraic complexity correlates highly with number of models, $r(74) = .78$, $p < .001$. Nevertheless, an analysis which regresses accuracy on the two predictors simultaneously shows that both number of models and algebraic complexity account for significant unique variance in accuracy, but models are the slightly stronger predictor: models, $\beta = -.52$, $t(73) = -4.42$, $p < .001$, algebraic complexity, $\beta = -.29$, $t(73) = -2.47$, $p < .02$. No statistical comparison was possible for invariance, because Vigo (2009) did not report values for all 76 concepts. However, number of models was a reliably better predictor of accuracy than minimal descriptions and modal clusters in SUSTAIN, William's ts(73) = 3.54, 4.76, respectively, both ps < .001. SUSTAIN, however, is somewhat sensitive to the ordering of instances (Bradley Love, personal communication, November 19, 2009). It sometimes returns different numbers of clusters when the instances are in a different order. It is not feasible to run all possible orderings of instances across Feldman's 76 concepts. But, we suspect that this factor does not make much of a difference on the whole. We have run SUSTAIN for the Shepard et al. (1961) problems with a variety of different orderings of the instances and found that it made no difference in terms of number of clusters – the modal number of clusters was identical in each case. Different parameter settings for this model may improve its fit across the full set of 76 Feldman concepts.

An adjustment of the mental model analysis by taking account of "parity" does not improve its fit. In fact, it considerably worsens the fit. To check this claim, we re-computed the number of models for the "down" parity problems (those that had more instances than non-instances), and treated them instead as if they were "up" parity problems, by computing the number of models for the non-instances of these problems. The resulting fit across all 76 problems was appreciably worse than that produced by the original analysis, $r(74) = -.58$, $p < .001$. What this finding reflects is that across the 76 problems, the "up" problems (mean accuracy = 86.12%) were in fact reliably more difficult than the structurally identical "down" problems (mean accuracy = 77.77%, Mann-Whitney U test, $p < .001$). Thus, the participants did not invariably encode the "down" concepts in terms of their non-instances. Some individuals may have focused on the non-instances for the "down" concepts, but there is no way to tell whether this focus improved their performance. We return to this point later.

The moral of these analyses is clear: the model theory, which is based on a plausible mental procedure for acquiring concepts, yields a single predictive parameter – the number of simplified mental models of a concept's instances – that explains the difficulty of acquiring these concepts. Taken together, the analyses demonstrate that the model theory outperforms each of its main rivals.

One potential reinterpretation of these findings is that a set of mental models can be translated into a syntactic description – since a set of models is equivalent to a disjunction of conjunctions. Hence, skeptics might argue that it is the length of this description, i.e., a variant on the minimal descriptions, that predicts difficulty rather than number of models. Models can indeed be translated into a description, but this alternative fails as a deflationary account of our findings for two separate reasons. On the one hand, descriptions of models are not minimal Boolean descriptions, and nothing explains why this particular sort of description should best predict difficulty other than that it captures the underlying models of concepts. In contrast, true minimal Boolean descriptions can take any form, whereas such re-descriptions of models always have a particular form – they are constrained

to be disjunctions of conjuncts. On the other hand, the models of a concept can differ in the number of variables that they represent, and this factor affects the lengths of their description, but not the number of models. The model theory predicts that the number of models for a concept should predict its difficulty, rather than the length of the re-description of these models. The correlation between number of models and the length of their descriptions is very high for the Feldman concepts, although not perfect, $r(74) = .90$, $p < .0001$. And, critically, the lengths of the descriptions correlate with accuracy at $r(74) = -.62$, $p < .001$, which is reliably lower than the correlation for number of models, William's $t(73) = 3.98$, $p < .0001$. Hence, the number of models is the better predictor. Theoretically, a mental model instantiating many variables should be harder to learn than one instantiating only a small number of properties. But, no such effect occurs, because a concept based on, say, six variables is likely to be beyond human capacity to acquire. Six variables yield 64 possibilities, and most of them could be instances of the concept. For this reason, investigators restrict their studies to concepts having no more than four variables. The model theory's insensitivity to small differences in the size of models, and its focus on the number of models, give it its predictive edge for such concepts.

6. Experiment 1: Descriptions of Boolean concepts

The analyses in the previous section corroborate the principle of simplifying models with respect to a particular sort of concept learning task, which is almost exclusively to be found in the literature – one in which individuals learn to categorize instances of a concept. But, what happens when the task is not merely to distinguish between the instances and non-instances of a concept, but also to describe the concept accurately? The present section reports the results of two experiments that aimed to answer this question. The task of describing concepts has several advantages. It tests whether the predictions of the relevant theories generalize to a different task than is typically used in experiments. Moreover, as models are simplifications of instances of a concept, they might be biased to predict the accuracy of classifications better than the accuracy of descriptions. So, the task provides a more stringent test of the model theory. Descriptions also allow greater insight into the ways that individuals represent each concept. And, as the data reveal, they show that individuals use a variety of linguistic strategies that are not well accounted for by existing theories.

Experiment 1 examined concepts based on a modified "switch" task from Johnson-Laird (1983). In a computer display, three independent binary switches controlled a light, and the participants made a series of test settings of the switches to discover what turned the light on. They had to describe this concept in their own words. We selected nine concepts from the overall set of 250 possible concepts concerning three binary variables (see Feldman, 2003a) in order to examine whether number of mental models also predicted the difficulty of this task. We also examined two other potential predictors of difficulty: minimal descriptions and algebraic complexity.

6.1. Method

The participants were 28 undergraduate students (16 female, 12 male), who took part in the experiment for course credit. They acted as their own controls, and tackled a set of nine Boolean concepts in which they had to describe the conditions in which a light came on as a result of the positions of three independent binary switches. They tried out different switch positions until they could describe the concept in their own words. Table 5 presents the instances of each of the nine concepts, their minimal description lengths (both including *or else* and not including *or else*), their algebraic complexity, and their mental models. These variables are not orthogonal. For instance, concepts with more instances also tend to have more models. Nevertheless, none of the potential measures of complexity are perfectly correlated, and the concepts were chosen so that a comparison of their relative predictive strengths was feasible.

Each of the switch letters (*a*–*c*) in Table 5 could potentially correspond to the leftmost, middle or rightmost switch. With three switch letters, and three switch locations, there are six unique assignments of letters to switch locations. For each participant, the assignments were determined randomly

for each new concept, and the nine concepts were presented in a different random order. The participants pressed numbered buttons corresponding to the switch numbers to set up a configuration of the switches, and then "submitted" the configuration to see whether or not the light came on. They were permitted to make as many such submissions as they wanted within the allotted time period of five minutes. The three switches started in the "off" position, i.e., down in the USA, and when participants pressed the numbered button corresponding to a particular switch, the switch moved up to the "on" position, where it remained until its button was pressed again. The negated switch values (shown in Table 5) refer to the particular switch being in the "down" position whereas the affirmative (non-negated) switch values refer to the particular switch being in the "up" position.

The instructions stressed that participants' task was to describe the conditions that would turn the light on, and that that they should do so in their own words. They wrote down their descriptions on a sheet of paper, with space for each concept, but were not allowed to make written notes as they worked. If they were uncertain about how they should respond, they were told that they should aim to describe as clearly as possible the circumstances in which the light came on, and that the format of the descriptions was up to them. As soon as the participants were ready to write a description of a concept, they pressed a "submit" key, and they had a maximum duration of five minutes per concept (not including the time required to write the description).

6.2. Results

Although the participants' descriptions varied considerably (see below), it was straightforward to assess their accuracy. Table 5 presents the percentages of accurate descriptions for each of the nine concepts. As predicted, there was a declining trend in the accuracy of descriptions as the number of models of the concepts increased: 1 model: 96% accuracy, 2 models: 74% accuracy, three models: 52% accuracy, and 4-models: 79% accuracy (Page's L, $z = 1.96$, $p < .05$). Although the trend is reliable, accuracy for the 4-model concept was high. There was only one such concept (9), and when it is excluded from the analysis, the effect of models was substantially stronger (Page's L, $z = 4.81$, $p < .00001$). Moreover, the performance on this apparently aberrant concept can be explained by the model theory – the concept has more instances (6) than non-instances (2), and indeed nine participants focused on the non-instances and made a greater proportion of accurate descriptions (89%) than the 19 participants who focused on instances (58%; $\chi^2 = 2.69$, $p < .06$, one tailed).

The theory predicts that individuals can focus on non-instances for this sort of problem, because it has more instances than non-instances. Focusing on non-instances is more likely for the present procedure than it was for Feldman's (2000) task. In Feldman's (2000) procedure, all of the instances were presented together at the top of the screen, with the non-instances presented together at the bottom of the screen, whereas in the present procedure, the participants' tests created intermingled instances and non-instances one at a time. Moreover, when participants need to describe each concept, they are also more likely to focus on non-instances. In this case, since they have to construct an explicit "summary representation" of the concept, participants can choose to focus on either instances or non-instances, knowing that if they correctly describe either set, then they will also have correctly described the complementary set. And, in fact, several participants did use this strategy for the "down" parity problems, and benefited from it for two other concepts. For concept 3, 11 participants focused on non-instances and produced 82% accurate descriptions, and 16 participants focused on instances and produced 44% accurate descriptions ($\chi^2 = 3.91$, $p < .03$, one tailed). For concept 8, 4 participants focused on non-instances and produced 100% accurate descriptions, and 22 participants focused on instances and produced 59% accurate descriptions ($\chi^2 = 2.5$, $p < .06$, one tailed). Thus, it seems that individuals can selectively focus on non-instances in order to reduce problem difficulty.

In contrast to the model-based analysis, minimum description length, as computed by our program rather than by Feldman's heuristics (thus including *or else* as a Boolean operator), failed to yield a reliable trend in difficulty: two literals: 79% accuracy; three literals: 61% accuracy; four literals: 73% accuracy (Page's L, $z = .47$, $p > .30$). The predictions of minimum description length were not improved by taking into account whether or not a concept has more instances than non-instances, i.e., the "parity" of a concept. Nor were they improved by excluding problem 9. The predictions of minimal description length were slightly improved by relying on minimal descriptions that did not include *or else,* although the pre-

dicted trend was still not reliable: two literals: 77% accuracy; four literals: 71% accuracy; five literals; 71% accuracy, six literals: 65% accuracy (Page's L, $z = .75$, $p > .20$). However, algebraic complexity (AC) did very well in accounting for performance: AC = −1, 94% accuracy, AC = −0.5, 84% accuracy, AC = 0, 67% accuracy, AC = 0.5, 39% accuracy, AC = 1, 66% accuracy, Page's L, $z = 3.85$, $p < .001$.

The model theory's predictions were correct for the latencies of the correct descriptions, and for the numbers of tests required for the correct descriptions. As number of models increased, so too did the time participants took to formulate a correct description: one model: 104 s, two models: 122 s, three models: 140 s; 4-models: 112 s. Again, the single 4-model concept was somewhat aberrant, but when it was excluded from the analysis, the trend was highly reliable (Page's L, $z = 3.62$, $p < .001$). However, neither form of minimal description length had a reliable effect on the latencies of the correct descriptions. Algebraic complexity did not have reliable effects on correct latencies when including all problems, although this was in large part because of low power – there was only a single problem with an algebraic complexity value of 0.5, and its inclusion led to substantial loss of power in the analysis (since many participants performed it incorrectly). When this level of algebraic complexity was excluded there was a reliable effect on correct latencies: AC = −1, 116 s, AC = −0.5, 114 s, AC = 0, 158 s, AC = 1, 159 s, Page's L, $z = 3.55$, $p < .001$.

We also examined the number of tests required for correct performance – that is, for each concept that was solved correctly, the number of times participants tested a configuration in order to see whether the light came on in response. There are eight different configurations for each concept, which means that participants should test all eight configurations at least once. In fact, they generally tested each configuration more than once. The high number of tests shows that individuals had some difficulty in remembering them (see Whitfield, 1951, for analogous results). As the number of models increased, the number of tests required to yield correct descriptions also increased: one model: 16.5 tests, two models: 19.15 tests, three models: 20.96 tests, 4-models: 17.36 tests. Again, when the single 4-model concept is excluded from the analysis, the trend is reliable (Page's L, $z = 2.86$, $p < .01$). Once more, minimal descriptions did not have a reliable effect on this measure.[3] Algebraic complexity also had no reliable effect when all problems were included, although this was again due to low power (see earlier). Once the problem with an algebraic complexity of 0.5 was excluded, algebraic complexity did have a reliable effect on the number of tests: AC = −1, 16.167 tests, AC = −0.5, 15.90 tests s, AC = 0, 21.15 tests, AC = 1, 20.29 tests, Page's L, $z = 3.86$, $p < .001$.

A further interesting trend emerged from the data. The participants' descriptions were classified (by two independent judges, one of whom was the first author) as *outside* Boolean algebra if they made use of quantifiers, e.g., "all", "any two", "alone", relations between the switch settings, e.g., "same" or "different", causal relations between the switches, e.g., "switch one inhibits switch two", "switch one negates switch two", or arithmetic operators, e.g., "light comes on if the switch values sum to four". The judges agreed on the classification of 91% of the descriptions, and the first author resolved the discrepancies. The vast majority of participants' descriptions (87%) went outside Boolean language in using relations, quantifiers, or arithmetical operations. For instance, participants wrote descriptions such as, "the light will go on when switches one and two are in the same position" (concept 2), and, "switch three must be up by itself, or all switches must be up for the light to come on" (concept 4). Such descriptions are not straightforwardly accounted for by any of the theories under consideration, and we return to this issue presently.

Descriptions were also coded in terms of whether or not they were disjunctive – that is, whether or not they enumerated a set of alternative possibilities that turned the light on. The judges also agreed on 84% of occasions about whether or not the descriptions were disjunctive, and again the first author resolved the discrepancies. Overall, 70% of the descriptions were disjunctive, whereas the remaining 30% were not. One sort of disjunctive description was a complete disjunctive normal form, i.e., a complete list of instances that turned the light on; whereas other sorts of disjunctive description were more succinct. Disjunctive descriptions are compatible with the use of mental models, because they

[3] We report these analyses of minimal complexity and algebraic complexity with problem 9 included. Including problem 9 generally had little effect on the results of these analyses.

enumerate a set of *alternative possibilities* that turn the light on, i.e., the instances of the concept. In contrast, only 2% of the descriptions were minimal Boolean descriptions.

7. Experiment 2: A replication

Experiment 2 was a replication of the previous experiment using an analogous set of nine concepts, which were constructed by negating the *b* variable in the minimal descriptions in Table 5: *b* in a description in the table was switched to *not b*, and vice versa. The participants were 27 undergraduate students (12 male, 14 female, one unreported), who took part in the experiment for course credit. The procedure was the same as the previous experiment.

7.1. Results

Table 5 above presents the percentages of accurate descriptions for the nine concepts. Concept 9 was aberrant: it should have been the most difficult according to all the theories under consideration, but it was performed with 100% accuracy. With hindsight, the reason was obvious: the light came on in every case except when all of the switches were on or when all of them were off. These configurations were easy to detect, and easy to describe using quantifiers. Indeed, quantifiers were used in 92% of descriptions of this concept in comparison with 55% of descriptions for the remaining eight concepts; Wilcoxon test, $z = 4.05$, $p < .001$). Hence, we excluded this concept from the remaining analyses of participants' performance (accuracies, latencies, and number of tests required).

Overall, the results were similar to those of the previous experiment. There was a declining trend in the accuracy of descriptions as the number of models increased: one model: 100%, two models: 82%, three models: 65% (Page's *L*, $z = 3.33$, $p < .001$). For the three problems that had fewer non-instances than instances, unlike the previous experiment, there was no reliable difference between participants who focused on instances as opposed to non-instances. Minimal description length predicted accuracy to a greater extent than in the previous experiment, and was reliable: two literals: 93%, three literals: 75%, four literals: 67% (Page's *L*, $z = 2.44$, $p < .01$). Minimal descriptions also predicted accuracy when *or else* was not included in the language of description, although the trend was only marginal in this case: two literals: 94%, four literals: 70%, five literals: 81, six literals: 74% (Page's *L*, $z = 1.67$, $p < .10$). Algebraic complexity predicted accuracy reliably too: AC = −1, 100% accuracy; AC = −0.5, 89% accuracy; AC = 0, 76% accuracy; AC = 0.5, 65% accuracy; AC = 1, 57% accuracy; Page's *L*, $z = 3.11$, $p < .001$.

The results for the latencies of correct descriptions and for the numbers of tests required for the correct descriptions corroborated the model theory. As number of models increased, so too did the time required for a correct description: one model: 92 s, two models: 100 s, three models: 141 s (Page's *L*, $z = 2.93$, $p < .01$). Minimal descriptions did not predict latencies, whereas algebraic complexity did: AC = −1, 105 s; AC = −0.5, 97 s; AC = 0, 111 s; AC = 0, 111 s; AC = 0.5, 143 s; AC = 1, 136 s, Page's *L*, $z = 2.34$, $p < .05$. As the number of models increased, the number of tests required for correct descriptions also increased: one model: 14.2 tests, two models: 13.9 tests, three models: 16.6 tests (Page's *L*, $z = 2.55$, $p < .02$). Minimal descriptions did not reliably predict number of tests. Nor did algebraic complexity with all problems included, although it did do so when the single problem with algebraic complexity of 0.5 was excluded, AC = −1, 14.28 tests; AC = −0.5, 12.17 tests; AC = 0, 13.90 tests; AC = = 1, 16.22 tests, Page's *L*, $z = 2.75$, $p < .01$.

The descriptions again tended to go outside the language of Boolean algebra (72% of all descriptions), and only 17% of them were correct minimal descriptions. Overall, 70% of the descriptions were disjunctive, whereas the remaining 30% were not. Disjunctive descriptions were again expressed either as complete disjunctive normal forms, or more succinctly.

7.2. Discussion of Experiments 1 and 2

Number of models consistently predicted accuracy, correct latencies, and number of tests. Minimal complexity was able to predict accuracy in Experiment 2, though it did not do well predicting correct latencies or number of tests in either experiment. Algebraic complexity performed very well, predict-

ing accuracy, latencies and number of tests well across both experiments. There is thus very little to separate number of models and algebraic complexity in these data, although the slight edge may go to number of models because it predicts the use of disjunctive descriptions. The frequent use of relations and quantifiers in individuals' descriptions goes beyond the scope of any of the theories discussed. However, although the use of such descriptions poses a substantial challenge for all of the theories under consideration, the mental model theory is in principle capable of accommodating them, which we discuss further in Section 9.

8. General discussion

Negation (*not*), conjunction (*and*), and disjunction (*or*) play major roles in the formation of concepts from existing properties, and relations. Fifty years ago psychologists discovered that conjunctive concepts are easier to acquire than disjunctive concepts, and then that robust differences also occur among more complex concepts (Shepard et al., 1961). Recent studies have led to a resurgence of interest in Boolean concepts, particularly as a result of Feldman's (2000) bold conjecture that the length of a concept's minimal description predicts the difficulty of learning to categorize its instances and non-instances. The evidence in favor of this theory is weaker than it first appeared. The theory correctly identified the importance of simplification, but the present research suggests that what is simplified are models of the instances of concepts.

Theorists have improved upon the hypothesis of minimal descriptions. First, Feldman himself has proposed a theory of algebraic complexity, which posits that a weighted mean of the complexity of the rules into which a concept can be decomposed predicts the difficulty of its acquisition (Feldman, 2006). Second, Vigo (2009) has proposed that the degree of invariance across the instances of a concept predicts how easy the concept is to learn. Third, Love et al. (2004) have proposed that SUSTAIN, a connectionist model of prototypes, extends to Boolean concepts. It does less well, but it is not strictly applicable to the results for Feldman's concepts, which depended on a simultaneous presentation of instances and non-instances. Fourth, we have proposed a theory based on mental models in the present paper. Mental models were invoked in a theory of reasoning with Boolean connectives (Johnson-Laird, 2006). The theory readily extends to concepts. The key idea is that individuals seek to categorize the instances of a concept using models that do not represent those variables that are irrelevant given the values of the other variables (the principle of simplifying models). We devised a computer program that implemented this principle. The numbers of models that the program produced explained the relative difficulty of both the Shepard concepts and the 76 novel Feldman concepts. Although the model theory shares the same broad idea as Feldman's minimal description theory, it rests on a different theoretical basis, and yields empirically distinguishable predictions. Indeed, it appears to give a better account of the Feldman concepts than the rival accounts of categorical invariance, and the SUSTAIN model. Algebraic complexity is the model theory's closest rival for these concepts, as it does almost as well as in accounting for these data. In the present experiments, the participants formulated their own *descriptions* of concepts that they acquired as a result of testing potential instances, and their results in terms of accuracy, latency, and number of tests, again showed that the model theory gave a better account than all of its rivals except algebraic complexity.

The model theory of Boolean concepts is a theory at the computational level, but it also embodies a plausible procedure at the algorithmic level (cf. Marr, 1982): Its implementation in a computer program yields an ideal simplification of the instances of a Boolean concept granted that individuals eliminate irrelevant variables by discovering that their possible values – the presence or absence of a property – occur in instances that otherwise have properties in common. You observe, say, that members of a certain elite club are either rich or Republican, but in either case may be male or female, and so you can eliminate gender as relevant to membership. Our best evidence for this process is the participants' frequent use of succinct disjunctive descriptions of concepts.

The program at present is quite insensitive to the order in which it receives the instances. However, if they were presented one at a time, a reasonable assumption is that individuals would be more likely to simplify instances that occur near to one another in the ordering than those that are distant from one another. For instance, a concept with the following three instances:

$$\begin{array}{ccc} a & b & \neg c \\ \neg a & \neg b & c \\ \neg a & \neg b & \neg c \end{array}$$

can be represented in the following two models:

$$\begin{array}{ccc} a & b & \neg c \\ \neg a & \neg b & \end{array}$$

once it is detected that *c* is irrelevant in the presence of *not a* and *not b*. Consequently, the model theory predicts that this concept should be easier to learn when the instances containing the irrelevant attribute are presented one after the other, thus facilitating the detection of the irrelevancy (see Mathy & Feldman, 2009, for a related idea regarding item presentation order for concepts with a "rule plus exception" structure). This notion gives rise to predictions about which particular simplifications individuals should make when more than one is available. The more contiguous the instances containing the irrelevant attributes are, the more likely it should be that the particular simplification is made. No current data exist to test these predictions systematically, so they call for future research.

Another topic for future research is to investigate the possibility of a rapprochement with rival theories in this domain. There may be ways to exploit the strengths of various models within a model-based framework. The most likely candidates are SUSTAIN, which, given its representation of concepts in terms of separable clusters, shares a broad conceptual similarity with the model theory, and algebraic complexity, which Feldman (2006) describes as the analytic counterpart of SUSTAIN. A recent rule-based approach to concept learning is due to Goodman, Tenenbaum, Feldman, and Griffiths (2008). It uses rational rules and also has some similarity to the model theory. It relies on a disjunctive normal form grammar, and thus explicitly constrains the rules it constructs to be disjunctions of conjuncts. It is capable of handling Boolean concepts, as well as concepts with continuous features. This model provides a good fit to the six Shepard et al. (1961) concepts, but has not yet been extended to Feldman's (2000) concepts, and so it is impossible to assess its performance with them (Noah Goodman, personal communication, July 7, 2010). Given its prioritization of disjunctive normal form representations, it is more similar to the model-based approach than previous rule-based approaches, and so a combination of the two approaches might yield a better theory than either of them.

Many studies have investigated the acquisition of concepts from their instances, but few have examined how individuals formulate *descriptions* of the resulting concepts (though cf. Shepard et al., 1961; Wason & Johnson-Laird, 1972). Of the rivals to the model theory, only minimal descriptions in principle provide an account of this task. But, the two experiments presented here showed that number of models gives the best account of participants' performance on this task.

Participants' descriptions also revealed that they seldom formulated minimal descriptions, and did not restrict themselves to the linguistic analogs of the operators in Boolean algebra, but instead freely used quantifiers, such as "at least one", and "any two", and relations, such as, "switch 1 has to be in the *same* position as switch 2". Such descriptions are powerful, and they go beyond all of the theories under consideration, including the formulation of the model theory as it applies to Feldman's (2000) data.

In principle, however, the model theory is capable of accounting for this tendency. The theory has been shown in the past to explain deductive reasoning with relations (e.g., Goodwin & Johnson-Laird, 2005) and quantifiers (Bucciarelli & Johnson-Laird, 1999). These prior applications of the theory concerned reasoning *from* quantified or relational assertions, whereas the present data call for an account of how individuals reason *to* such assertions. Several critical questions need to be answered about this process. One question concerns the circumstances in which people search for such descriptions. We suspect that a critical factor is the perceptual similarity of the attributes under description. The use of quantified and relational expressions is a kind of conceptual *chunking*, which simplifies the problems under description by relating the attributes of the concept together succinctly (e.g., Halford et al., 1998). But, in order for this chunking to occur, the attributes must be relatable in some way. Thus, it should be much easier to use relational and quantified expressions when the attributes constituting a concept are perceptually similar, which makes it possible to formulate descriptions such as: "all switches on/ off", "only one switch needs to be on", or "switches A and B need to be in the same

position". Indeed, we have shown that individuals are more likely to use quantified and relational descriptions when concepts depend on homogeneous variables, such as three switches of the same sort, as opposed to three different "toggles" (see e.g., Goodwin & Johnson-Laird, 2011). The use of such descriptions leads to a substantial change in the difficulty ordering of the Shepard et al. (1961) concepts (see Goodwin & Johnson-Laird, 2011).

The likelihood of representing a concept in terms of relations and quantifiers should also increase when individuals need to formulate a *description* of a concept, as was required in our experiments. The need to describe concepts should increase the accessibility of linguistic resources that may not immediately spring to mind when participants need only to classify individual instances (the more standard procedure). Moreover, when individuals have to provide descriptions of the *whole* concept they have learned, as opposed to just descriptions of its separate instances, they may be more inclined to seek such linguistic simplifications. Both of these factors – the homogeneity of the variables, and the need for descriptions – were present in our experiments, but not in Feldman's (2000) task. This difference may explain why models do not need to represent relations and quantifiers in Feldman's (2000) task, but do need to represent them in our description task.

How can the model theory be extended to account for the use of quantifiers and relations? Two main developments of the theory are necessary. First, there is a need for a parameter that represents the likelihood that an individual will represent a concept in terms of relations and quantifiers. Only some of our participants used such descriptions, and even those who did, did not do so for all concepts. Hence, the need for some kind of probabilistic gating on this strategy – it seems to consist of a kind of "insight" that may be hard to predict a priori. Second, the model theory needs to be extended so as to be able to represent descriptions that integrate Boolean and non-Boolean expressions, for instance: "Light comes on when all of the switches are off, or when switch 1 is on", "Light comes on when switches 1 and 2 are the same or when switch 3 is on by itself". How are such descriptions represented in mental models, and how many models do they give rise to? There may be a need for some new notational machinery here. For instance, sameness between two switches can be represented by an equivalence relation in a single mental model:

switch 1 ≡ switch 2.

The same is true for the representation of the relation of difference, i.e., the negation of equivalence:

Switch 1 / ≡ switch 2.

Quantified assertions, such as "all switches on," can similarly be represented in a single model representing the three switches (see, e.g., Johnson-Laird, Byrne, & Tabossi, 1989):

Switch 1 on.
Switch 2 on.
Switch 3 on.

Thus, from a conceptual point of view, the task of extending the model theory in this way does not require radically new assumptions about mental representation. Indeed, similar simplifying mechanisms and notations have been employed previously by the model theory in its use of "isomeric" models in spatial and temporal reasoning (e.g., Schaeken, Van der Henst, & Schroyens, 2007). However, there do remain significant technical and implementational challenges in extending the model theory in this way, which are beyond the scope of the present paper.

In Section 1, we formulated two key questions that psychological theories of concepts need to answer. One question is: *How do individuals envisage instances of a Boolean concept given its description?* The answer to this question comes from the principle of truth, which we outlined earlier (see the section on the model theory of concepts). At a high level, mental models represent only the instances of a concept (or, in the case of a change in focus, only its non-instances). At a low level, each mental model of an instance (or non-instance) represents only the values of variables that the description asserts as

holding in the instance. A computer program (in Common Lisp) implements the principle. It takes as input a *description* of any Boolean concept, which may contain conjunctions, inclusive and exclusive disjunctions, and various other connectives, and its output is a set of mental models representing the instances of the concept. It also outputs the instances of the concept when the principle of truth does not constrain its performance. The program predicts a systematic misunderstanding of certain concepts. Given the description, e.g.:

blue and large, or else blue.

the program yields two mental models of the instances of the concept:

blue	large	(the conjunction holds)
blue		("blue" holds)

But, as the program showed, the first of these models is erroneous. The conjunction in the description cannot hold, because when it does, "blue" in the second clause of the description also holds, which violates the exclusive disjunction between the two clauses. Thus, the single correct and fully explicit model of the concept is:

blue	¬large

Indeed, individuals make exactly this sort of error, and many others like it (see Goodwin & Johnson-Laird, 2010). They perform much better on the control concepts, in which the principle of truth does not yield errors. The errors cannot be readily explained by any of the rival theories of concepts that we have considered in this paper, and especially not by any theory, such as algebraic complexity or invariance, which presupposes an accurate Boolean account of concepts.

Hence, individuals rely on mental models of descriptions when they envisage the instances of a concept from its verbal description.

The second key question is: *How do individuals learn to categorize instances of a concept, and to describe these instances?* The research that we have reported here enables us to answer this question. When individuals acquire a concept, they seek to reduce the number of mental models representing its instances. They eliminate any variable that is irrelevant given the values of others. The number of resulting mental models provides a good predictor of the difficulty of the task. Likewise, in their descriptions of a concept, they tend to formulate disjunctive descriptions of their simplified models, but they also use linguistic resources such as quantifiers and relations. Further research is needed to account for the use of such non-Boolean descriptions.

Boolean concepts, such as "tall, dark, and handsome", have an obvious similarity to *assertions* using the same operators, such as "She was tall, dark, and handsome". So, perhaps it should not be too surprising that a theory developed for one domain applies to the other domain too. Mental models unite two hitherto separate and unrelated domains.

Acknowledgments

This research was supported by a fellowship awarded to the first author by the Woodrow Wilson School, Princeton University, and by National Science Foundation Grant SES 0844851 to the second author to study deductive and probabilistic reasoning. We thank Jacob Feldman for providing us with his experimental data, and for his advice, Rob Nosofsky for his advice on RULEX, Brad Love for his advice on SUSTAIN, John Kruschke for his advice on ALCOVE, and Noah Goodman for his advice on Rational Rules. For their helpful comments, we also thank Greg Detre, Adele Goldberg, Sam Glucksberg, Sangeet Khemlani, Louis Lee, Greg Murphy, Dan Osherson, and three anonymous reviewers.

References

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author (Revised).

Barsalou, L. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 101–140). Cambridge, UK: Cambridge University Press.

Bauer, M. I., & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychological Science, 4*, 372–378.
Bruner, J. S., Goodnow, J. S., & Austin, G. G. (1956). *A study of thinking*. New York: Wiley.
Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science, 23*, 247–303.
Bulgarella, R., & Archer, E. J. (1962). Concept identification of auditory stimuli as a function of amount or relevant and irrelevant information. *Journal of Experimental Psychology, 63*, 254–257.
Conant, M. B., & Trabasso, T. (1964). Conjunctive and disjunctive concept formation under equal-information conditions. *Journal of Experimental Psychology, 67*, 250–255.
Evans, J. S. P. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hillsdale, NJ: Erlbaum.
Federal Civil Judicial Procedure and Rules. (1994). Minneapolis-St. Paul, MN: West Publishing Co.
Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633 (5 October).
Feldman, J. (2003a). A catalog of Boolean concepts. *Journal of Mathematical Psychology, 47*, 75–89.
Feldman, J. (2003b). Simplicity and complexity in human concept learning. *The General Psychologist, 38*, 9–15.
Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology, 50*, 339–368.
Fodor, J. (1994). Concepts: A potboiler. *Cognition, 50*, 95–113.
Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.
Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*, 108–154.
Goodwin, G. P., & Johnson-Laird, P. N. (2011). *The use of relations and quantifiers in descriptions of concepts*. Manuscript in preparation. University of Pennsylvania.
Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. *Psychological Review, 112*, 468–493.
Goodwin, G. P., & Johnson-Laird, P. N. (2010). Conceptual illusions. *Cognition, 114*, 253–265.
Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*, 803–831.
Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18*, 441–461.
Haygood, R. C., & Bourne, L. E. Jr., (1965). Attribute and rule learning aspects of conceptual behavior. *Psychological Review, 72*, 175–195.
Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.
Hovland, C. I., & Weiss, W. (1953). Transmission of information concerning concepts through positive and negative instances. *Journal of Experimental Psychology, 45*, 175–182.
Hunt, E. B. (1962). *Concept learning: An information processing problem*. New York: Wiley.
Hunt, E. B., & Hovland, C. I. (1960). Order of consideration of different types of concepts. *Journal of Experimental Psychology, 59*, 220–225.
Hunt, E. B., Marin, J., & Stone, P. J. (1966). *Experiments in induction*. New York: Academic Press.
Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, MA: Harvard University Press.
Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.
Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
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., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99*, 418–439.
Johnson-Laird, P. N., Byrne, R. M. J., & Tabossi, P. (1989). Reasoning by model: The case of multiple quantification. *Psychological Review, 96*, 658–673.
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.
Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.
Lafond, D., Lacouture, Y., & Mineau, G. (2007). Complexity minimization in rule-based category learning: Revising the catalog of Boolean concepts and evidence for non-minimal rules. *Journal of Mathematical Psychology, 51*, 57–74.
Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin and Review, 9*, 829–835.
Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science, 14*, 195–199.
Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.
Marr, D. (1982). *A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.
Mathy, F., & Bradmetz, J. (2004). A theory of the graceful complexification of concepts and their learnability. *Current Psychology of Cognition, 22*, 41–82.
Mathy, F., & Feldman, J. (2009). A rule-based presentation order facilitates category learning. *Psychonomic Bulletin & Review, 16*, 1050–1057.
Medin, D. L., & Smith, E. E. (1984) Concepts and concept formation. In M. R. Rosensweig & L. W. Porter (Eds.), *Annual review of psychology* (Vol. 35, pp. 113–118).
Medin, D. L., Lynch, E. B., & Solomon, K. O. (2000). Are there kinds of concepts? *Annual Review of Psychology, 51*, 121–147.
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.
Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.
Murphy, G., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.
Neisser, U., & Weene, P. (1962). Hierarchies in concept attainment. *Journal of Experimental Psychology, 64*, 640–645.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., & McKinley, S. C. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition, 22*, 352–369.
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. (1974–1976). Logical ability in children (Vols. 1–4). Hillsdale, NJ: Erlbaum.
Peirce, C. S. (1931–1958). 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.
Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.
Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.
Russell, S., & Norvig, P. (2003). *Artificial intelligence. A modern approach* (2nd ed.). Saddle River, NJ: Prentice Hall.
Schaeken, W., Van der Henst, J.-B., & Schroyens, W. (2007). 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* (pp. 129–151). Mahwah, NJ: Erlbaum.
Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75*, 1–42.
Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: A study of the Shepard, Hovland, and Jenkins (1961) tasks. *Journal of Experimental Psychology: General, 133*, 398–414.
Valiant, L. G. (1984). A theory of the learnable. *Communications of the Association for Computing Machinery, 27*, 1134–1142.
Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin and Review, 2*, 442–459.
Vigo, R. (2006). A note on the complexity of Boolean concepts. *Journal of Mathematical Psychology, 50*, 501–510.
Vigo, R. (2009). Categorical invariance and structural complexity in human concept learning. *Journal of Mathematical Psychology, 53*, 203–221.
Walker, C. M., & Bourne, L. E. Jr., (1961). Concept identification as a function of amount of relevant and irrelevant information. *American Journal of Psychology, 74*, 410–417.
Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.
Whitfield, J. W. (1951). An experiment in problem solving. *Quarterly Journal of Experimental Psychology, 3*, 184–197.