The Quarterly Journal of Experimental Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

Immediate inferences from quantified assertions

Sangeet Khemlani[a], Max Lotstein[b], J. Gregory Trafton[a] & P. N. Johnson-Laird[cd]

[a] Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC, USA

[b] Center for Cognitive Science, University of Freiburg, Freiburg, Germany

[c] Department of Psychology, Princeton University, Princeton, NJ, USA

[d] New York University, New York, NY, USA

Accepted author version posted online: 21 Jan 2015.Published online: 02 Mar 2015.

To cite this article: Sangeet Khemlani, Max Lotstein, J. Gregory Trafton & P. N. Johnson-Laird (2015): Immediate inferences from quantified assertions, The Quarterly Journal of Experimental Psychology, DOI: 10.1080/17470218.2015.1007151

To link to this article: http://dx.doi.org/10.1080/17470218.2015.1007151

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2015
http://dx.doi.org/10.1080/17470218.2015.1007151

Immediate inferences from quantified assertions

Sangeet Khemlani[1], Max Lotstein[2], J. Gregory Trafton[1], and P. N. Johnson-Laird[3,4]

[1]Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC, USA
[2]Center for Cognitive Science, University of Freiburg, Freiburg, Germany
[3]Department of Psychology, Princeton University, Princeton, NJ, USA
[4]New York University, New York, NY, USA

We propose a theory of immediate inferences from assertions containing a single quantifier, such as: *All of the artists are bakers*; therefore, *some of the bakers are artists*. The theory is based on mental models and is implemented in a computer program, *mReasoner*. It predicts three main levels of increasing difficulty: (a) immediate inferences in which the premise and conclusion have identical meanings; (b) those in which the initial mental model of the premise yields the correct conclusion; and (c) those in which only an alternative to the initial model establishes the correct conclusion. These levels of difficulty were corroborated for inferences to necessary conclusions (in a reanalysis of data from Newstead, S. E., & Griggs, R. A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and Verbal Behavior*, *22*, 535–546), for inferences to modal conclusions, such as, *it is possible that all of the bakers are artists* (Experiment 1), for inferences with unorthodox quantifiers, such as, *most of the artists* (Experiment 2), and for inferences about the consistency of pairs of quantified assertions (Experiment 3). The theory also includes three parameters in a stochastic system that predicted quantitative differences in accuracy within the three main sorts of inference.

Keywords: Quantifiers; Logic; Mental models; Reasoning; Syllogisms.

Psychologists have studied quantifiers, such as, *Some of the actors*, and *All of the bakers*, for over a century (see, e.g., Störring, 1908). They have devised at least 12 theories of inferences from pairs of quantified premises—that is, syllogisms of the sort that Aristotle was the first to analyse. A recent meta-analysis of seven of these theories showed that none of them was exemplary; the other five theories lacked sufficient information to be included in the analysis (Khemlani & Johnson-Laird, 2012). Psychologists accordingly do not yet fully understand the mental processes underlying inferences that hinge on quantifiers. Any comprehensive theory, however, must also account for immediate inferences from quantified assertions. For instance, if you are told that *none of the servers is a woman*, you might refrain from asking a woman for the check. You have tacitly inferred:

Correspondence should be addressed to Sangeet Khemlani, Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC 20375, USA. E-mail: skhemlani@gmail.com

We thank Ruth Byrne, Chand Chandrasekaran, Hua Gao, Terry Stewart, Tony Harrison, Catrinel Haught, Niklas Kunze, Gorka Navarrete, Mike Oaksford, Marco Ragni, and Tobias Sonntag for their help and criticisms.

This research was supported by a National Research Council Research Associateship to the first author; and by the National Science Foundation [grant number SES 0844851] to the fourth author to study deductive and probabilistic reasoning.

1. None of the servers is a woman.
   Therefore, none of the women is a server.

The inference is *valid*—that is, its conclusion must be true given that its premise is true: The conclusion holds in any case in which the premise holds (Jeffrey, 1981, p. 1). Psychologists have investigated these inferences for many years (e.g., Begg & Harris, 1982; Newstead & Griggs, 1983; Wilkins, 1928), but again have yet to determine how logically untrained individuals make them. The present paper aims to solve this problem.

In what follows, we outline a theory of immediate inferences from quantified assertions, which is based on mental models, and which includes three stochastic parameters yielding quantitative predictions about the accuracy of individual inferences. We report the results of five experiments testing the theory's predictions. The first two are due to Newstead and Griggs (1983), and we carried out three further experiments testing new predictions of the theory. It deals with quantifiers of the sort that occur in syllogisms, but also with unorthodox quantifiers, such as *most of the chemists*, which are outside their scope. Finally, the paper draws some general conclusions about human reasoning from quantified assertions.

A model theory of reasoning with quantifiers

The theory of mental models—the "model" theory, for short—is a general theory of reasoning, which applies to any domain of reasoning, including inferences that depend on quantifiers, on sentential connectives, on spatial and other relations, and on probabilities. It rests on several assumptions that apply to reasoning in general (see Johnson-Laird, 2006, 2010; Johnson-Laird & Khemlani, 2014). First, individuals use the meanings of sentences—their "intensional" representations, which the mental parser assembles—to construct models of the "extensions" of the sentences—that is, the situations to which they refer. Each model represents what is common to a distinct set of possibilities, and so individuals judge that a conclusion is valid if it holds in all their models of the premises. Second, mental models represent assertions in as iconic a way as possible. This concept, which is due to the nineteenth-century logician Peirce (1931–1958, Vol. 4), means that the structure of a representation corresponds to the structure of what it represents. Hence, as Peirce realized, the representation yields conclusions about relations not asserted explicitly in the premises. Third, mental models of assertions represent what is true at the expense of what is false. Mental models therefore place a reduced load on working memory, but reasoners go wrong with certain inferences as a result of failing to represent what is false (see, e.g., Khemlani & Johnson-Laird, 2009; Yang & Johnson-Laird, 2000). Individuals tend to construct just a single initial mental model, which yields intuitive inferences, but, provided the task is not too difficult, they may be able to construct an alternative model. Inferences that depend on such alternatives are more difficult (Johnson-Laird & Khemlani, 2014).

We illustrate the theory as it applies to quantified assertions of the sort under investigation here. These simple quantified assertions are "monadic"—that is, they concern the properties of entities rather than relations amongst them. The three examples in (2) illustrate different sorts of monadic assertion:

2a. All of the artists are Cubists.
2b. Most poets are obscure.
2c. Some of the musicians like John Cage.

The last of these examples, in fact, expresses a relation, but it can be treated as monadic provided that the validity of the relevant inference follows from the treatment of the relation as a property, *likes–John–Cage(musician x)* instead of its decomposition into a relation (*likes*) with two arguments (*musician x*, and *John Cage*).

Monadic assertions refer to relations between sets of entities (see, e.g., Boole, 1854; Cohen & Nagel, 1934, pp. 124–125). This conception differs from psychological theories based on logic (e.g., Rips, 1994). However, it is presupposed in various diagrammatic theories (e.g., Ceraso & Provitera, 1971; Erickson, 1974), in formal rules for inferences about sets (e.g., Geurts, 2003; Politzer, van der Henst, Luche, & Noveck,

2006), and in mental models (Bucciarelli & Johnson-Laird, 1999; Johnson-Laird & Byrne, 1991; Polk & Newell, 1995). These various systems yield the same set of valid deductions. However, we frame the present theory in terms of mental models, because they lead to predictions about errors that distinguish the theory from the other accounts (Khemlani & Johnson-Laird, 2012). Theories relying on formal rules, either logical (e.g., Rips, 1994) or set-theoretical (e.g., Geurts, 2003), have not yet been formulated for immediate inferences, especially those depending on unorthodox quantifiers and modal conclusions about possibilities, and so it is unclear what predictions such accounts would make about errors.

The quantifiers of first-order logic translate into English roughly as: *for any A* and *at least some A*, where *A* ranges over individual entities. We refer to quantifiers such as *All of the A*, as "first-order", because they can be defined in terms of first-order quantifiers. They yield four "moods" of monadic assertions, which occur in classical syllogisms:

3. All of the A are B.
   Some of the A are B.
   None of the A are B.
   Some of the A are not B.

In contrast, *most of the A* cannot be defined in terms of first-order quantifiers (Barwise & Cooper, 1981). We refer to this quantifier and its cognates as "second-order", because they can be defined only in second-order logic, in which variables can range over individuals and *sets* of individuals. We use the definite article in quantifiers, such as, *all of the artists*, to establish that the relevant individuals exist in the domain of discourse (cf. Boolos, 1984; Johnson-Laird & Bara, 1984).

In immediate inferences based on first-order quantifiers, there are four possible moods for the premise, and eight sorts of conclusion (4 moods by 2 figures, i.e., arrangements of terms in the second premise, *A–B*, and *B–A*). There are three principal inferential tasks. The first task is to infer a conclusion about what is necessary, as in:

4. None of the artists is a baker.
   Therefore, none of the bakers is an artist.

The second task is to infer a conclusion about what is possible, as in:

5. All of the artists are bakers.
   Therefore, possibly all of the bakers are artists.

The third task is to infer whether two assertions are consistent with one another—that is, whether they can both be true at the same time. A theory of immediate inference needs to explain the ability of individuals to make all three of these sorts of inference, and the relative difficulty of different inferences within them, in terms of accuracy and latency. We formulated such a theory based on mental models, and we have implemented it in a Common Lisp program, *mReasoner*. A text-file of its output and its source code are available (http://mentalmodels.princeton.edu/programs/mreasoner/). In what follows, we first describe the three components of the theory: (a) its use of intensional representations to build mental models; (b) its use of mental models to draw immediate inferences; and (c) its revision of these models to construct alternatives in deliberations about immediate inferences. It predicts a trend in decreasing accuracy over three sorts of monadic immediate inference, but models alone do not make predictions about relative difficulty within these sets of inferences. Hence, we then introduce three stochastic parameters governing models that yield such quantitative predictions in a system that is also implemented in the *mReasoner* program.

Intensional representations

The system for immediate inferences has access to an intensional representation of a premise, which captures its meaning in set-theoretic terms (Khemlani & Johnson-Laird, 2013; Khemlani, Trafton, & Johnson-Laird, 2013). A set-theoretic semantics covers all monadic assertions, including those based on second-order quantifiers such as, *most of the A*. We list the five main sorts of quantified assertion under investigation with their informal set-theoretic meanings, which in all cases also assert that As and Bs exist in the relevant sets:

6. *All of the A are B*: The set of As is included in the set of Bs.

At least some of the A are B: The intersection of the sets of As and Bs is not empty.
None of the A is a B: The intersection of the sets of As and Bs is empty.
At least some of the A are not B: The intersection of the sets of As and not Bs is not empty.
Most of the A are B: The intersection of the sets of As and Bs has a cardinality greater than that of half the cardinality of the As.

An intension is a blueprint for constructing a mental model, but it also constrains inferences. For example, suppose that the assertion, *All of the A are B*, is represented in a model of three individuals:

7. A B
   A B
   A B

How in this case does the system avoid inferring that three *A*s are *B*s? The answer is that the intension shows that the choice of three instances of *A* was arbitrary, and it can be changed. In contrast, a revision to a model cannot add an *A* that is not a *B*, because that would violate the intension that *A*s are included in *B*s.

Intensions, which the parser constructs, can obviate the need for model-based inferences. If a premise and a conclusion have identical intensions, the conclusion obviously follows from the premises. A numerical quantifier, such as, *All three of the Spartans*, can be explicitly represented with three tokens in a model, but that is out of the question for the quantifier, *All 300 of the Spartans*. Its intensional representation, however, allows the model-building system to represent the tokens in the model with a tag for the numeral 300. Readers may wonder: Why not base all inferences on intensions? In principle, it can be done (cf. Geurts, 2003, p. 234; Oaksford & Chater, 2009, p. 83), but the evidence counts against such systems and in favour of iconic mental models. Formal rules do not offer an account of systematic errors, and the difficulty of inferences depends, not on the length of proofs in a formal system, but on the number of models (see Johnson-Laird & Khemlani, 2014).

Mental models

Monadic assertions have a single initial mental model (an extensional representation), which represents iconically the relation between the relevant sets. The intensional representation guides its construction. These models are "canonical" in that they satisfy the set-theoretic intension of the quantified assertion but follow an overarching principle of parsimony. They tend to represent as few different sorts of individual as possible, with the proviso that both sets referred to in an assertion are represented in the model. Canonical models have lower entropy than noncanonical models and are the "preferred" mental models (corroborated in the domain of spatial inferences; see Jahn, Knauff, & Johnson-Laird, 2007). Likewise, reasoners' spontaneous diagrams and their manipulations of external models corroborate the existence of canonical models of quantified assertions (Bucciarelli & Johnson-Laird, 1999). A canonical mental model of the assertion, *All of the A are B*, represents the two sets as coextensive, and most reasoners tend to interpret the assertion in this way unless the interpretation violates their knowledge of the two sets:

8. All of the A are B: A B
   A B
   A B

Each row in this diagram represents an individual, so the first row represents an individual who is both an *A* and a *B*, the second row represents another such individual, and so on. Table 1 shows the canonical models for all the different sorts of quantified assertion in the present studies: *All of the A are B*, *Some of the A are B*, *None of the A are B*, *Some of the A are not B*, and *Most of the A are not B*. It also shows models representing noncanonical individuals too.`

The search for alternative models

Of course, initial mental models imply conclusions that may not be valid—for example, the model (8) for *All of the A are B* implies that all of the B are A. The inference is intuitive, but invalid. Hence, the theory allows that individuals can deliberate and modify an initial model as long as the revision

Table 1. *The canonical models of monadic assertions, and examples of noncanonical models, as constructed by mReasoner*

Monadic assertion	*An example of a canonical initial model*		*An example of a noncanonical model*	
All of the A are B	A	B	A	B
	A	B	¬A	B
	A	B	¬A	¬B
At least some of the A are B	A	B	A	B
	A	B	A	¬B
	A	¬B	¬A	B
None of the A is a B	A	¬B	A	¬B
	A	¬B	A	¬B
	¬A	B	¬A	B
			¬A	¬B
At least some of the A are not B	A	¬B	A	¬B
	A	¬B	A	¬B
	¬A	B	¬A	B
			¬A	B
Most of the A are B	A	B	A	B
	A	B	A	B
	A	¬B	A	¬B
			¬A	B
			¬A	¬B
Most of the A are not B	A	¬B	A	¬B
	A	¬B	A	¬B
	A	B	A	B
			¬A	B
			¬A	¬B

Note: Each row in a model denotes the properties of an individual, and "¬" denotes the negation of a property.

satisfies the intension. This idea, which contrasts intuitions with deliberations, reflects a long-standing principle of the model theory (Johnson-Laird, 1983, chapter 6). In the case of the preceding inference from *All of the A are B*, deliberations can yield an alternative model consistent with the premise's intension:

9. A B
   A B
   A B
   ¬A B

where ¬A B represents an individual that is not an A but is a B. A study of the spontaneous use of diagrams in syllogistic reasoning demonstrated three main search procedures (Bucciarelli & Johnson-Laird, 1999), and we have implemented them in *mReasoner*: The program can add new individuals or properties to a model, break individuals with multiple properties apart, and move properties from one individual to another, if the result is consistent with the premise's intension (see Khemlani & Johnson-Laird, 2013, Table 4). Because (9) refutes the conclusion above, its inference is invalid.

The model theory predicts a general trend in accuracy and latency in immediate inferences. Inferences based on identical intensions should be the fastest and most accurate; inferences based on initial mental models should be intermediate in speed and accuracy; and inferences based on alternative models should be the slowest and least accurate.

The stochastic system

The model theory predicts the preceding trend, but it makes no quantitative predictions about individual inferences. Readers may also suppose that the theory postulates that reasoning is a deterministic process—that is, it unfolds like clockwork, and reasoners build the same models with the same number of individuals with the same properties, and so on, in an invariable process. In fact, the theory makes no such assumption. It presupposes that inferential processes are almost always stochastic, and so as a consequence models differ from one occasion to another in the number of individuals that they represent and the properties that they represent. We accordingly formulated such a system, introducing stochastic parameters governing the construction of models and the search for alternatives. Our aim in the first instance was to account for the distribution of results in Newstead and Griggs's (1983) studies of immediate inferences, but we framed the system in general terms so that in principle it extends to new sorts of inference. If reasoners are using mental models for immediate inferences with quantifiers, what factors in general should

Table 2. *Examples of models of "All A are B" containing various numbers of individuals and various proportions of atypical—noncanonical—individuals.*

	Number of individuals in the model (λ)							
Presence of noncanonical individuals	*2*		*3*		*4*		*5*	
None ($\varepsilon = 0.0$)	A	B	A	B	A	B	A	B
	A	B	A	B	A	B	A	B
			A	B	A	B	A	B
					A	B	A	B
							A	B
Some ($\varepsilon = 0.5$)			A	B	A	B	A	B
			A	B	A	B	A	B
			¬A	¬B	¬A	¬B	A	B
					¬A	¬B	¬A	¬B
							¬A	¬B
Full set of possibilities ($\varepsilon = 1.0$)					A	B	A	B
					A	B	A	B
					¬A	B	¬A	B
					¬A	B	¬A	B
							¬A	¬B

affect their performance with different sorts of problem? In our view, there are three.

The first factor is the number of individuals that they tend to represent in a model. This factor matters because if a model of *Some of the A are B* represents only three individuals, then it may yield invalid inferences because it cannot represent all four sorts of individual consistent with the premise—for example, a model such as:

A B
A B
A¬B

supports the invalid conclusion: *All of the B are A*. However, the mere number of individuals in a model does not guarantee accuracy: As Table 2 illustrates, they could all be of the same sort.`

The number of individuals represented in a model is likely to vary depending on several factors—of which the most important is likely to be the processing capacity of working memory. The number could vary according to many possible distributions, but the most plausible is a discrete Poisson distribution, because it captures the probability of a given number of events within a fixed interval of time or, in our case, a given number of entities in a bounded space. Unlike a normal distribution or the chi-squared distributions, it has the further advantage of being specified by a single parameter, λ, which states both its mean and its variance. The distribution is discrete because a set of individuals contains a discrete number of individuals—for example, no model represents four and a half artists. And it is "left-truncated" (Deshpande, Gore, & Shanubhogue, 1995, p. 199) because the quantifiers in our experiments are plural and therefore call for at least two individuals. Figure 1 shows the Poisson distributions for various values of λ that seemed appropriate a priori.

The second factor is the proportion of atypical individuals in a model. Reasoners tend to focus on typical instances of an assertion, but their reasoning is more accurate when they represent a fuller set of possibilities. For any assertion, a model is built from representations of individuals either in the *canonical* set for the assertion or in the *full* set of possible individuals consistent with the assertion's intension (see Table 1). The canonical set contains only typical instances, as shown in the diagrams that participants draw and the external models that they construct (e.g., Bucciarelli & Johnson-Laird, 1999). The full set adds to them

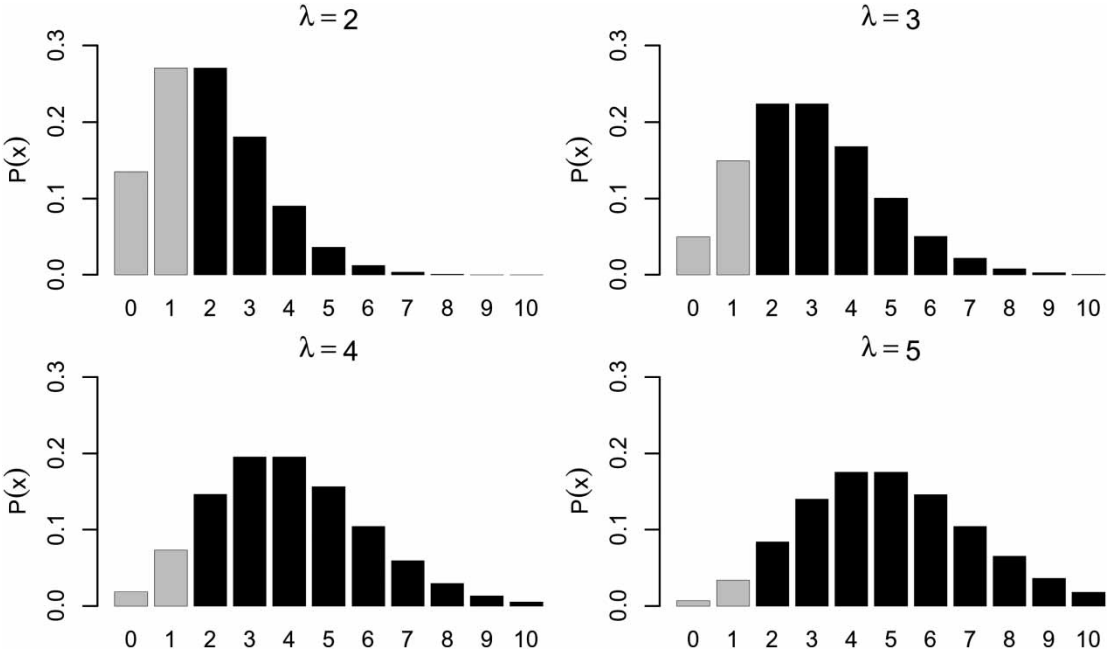

Figure 1. *Left truncated Poisson distributions for various values of λ. The distributions establish stochastically the number of individuals represented in models. Grey bars indicate values that were truncated (0 and 1) because they are impossible for plural quantifiers.*

other less typical instances. For example, the canonical set for the assertion *All of the A are B* contains only one sort of individual, A B, but the assertion's intension is consistent with two other possible sorts of individual, ¬*A B* and ¬*A* ¬*B*; and only one sort of individual is impossible, *A* ¬*B*. The parameter, ε, fixes the likelihood of constructing a representation based on the full set of individuals as opposed to the canonical set. When $\varepsilon = 0$, the model is built from canonical possibilities only, as in Table 1. When $\varepsilon = 1.0$ (its maximum), the model is built from the full set of possibilities. Table 2 illustrates models containing various numbers of individuals (depending on λ) and of varying proportions of noncanonical individuals (depending on ε). Each of these models satisfies the premise, because every *A* is also a *B*. But the different variations permit different immediate inferences.

The third factor is the probability that individuals search for an alternative to their initial model. This factor reflects the degree to which reasoners are able to go beyond their initial intuitions and to deliberate about possibilities. The initial mental model suffices for many intuitive tasks, but, as we have illustrated, a correct response may call for the construction of an alternative model. For simplicity, the parameter, σ, which is also in the unit interval [0, 1], sets the probability that such a search is successful.

In order to model the quantitative differences in the results of Newstead and Griggs (1983), we implemented the three parameters in the *mReasoner* program. The system first draws a sample from a Poisson distribution with parameter λ. It uses the sample, say, 3, to be the number of individuals in the initial model. Each individual is added progressively by drawing a random sample from one of two sets: the canonical set of possibilities, or the full set of possibilities consistent with the premise, where the parameter, ε, sets the proportional chance of sampling the set of full set of possibilities. Figure 2 summarizes these steps. The resulting model is scanned to draw or to evaluate a conclusion. Finally, either the program searches for an alternative model, which it finds, with a probability set by the parameter, σ, or it does not make such a search. As we show later,

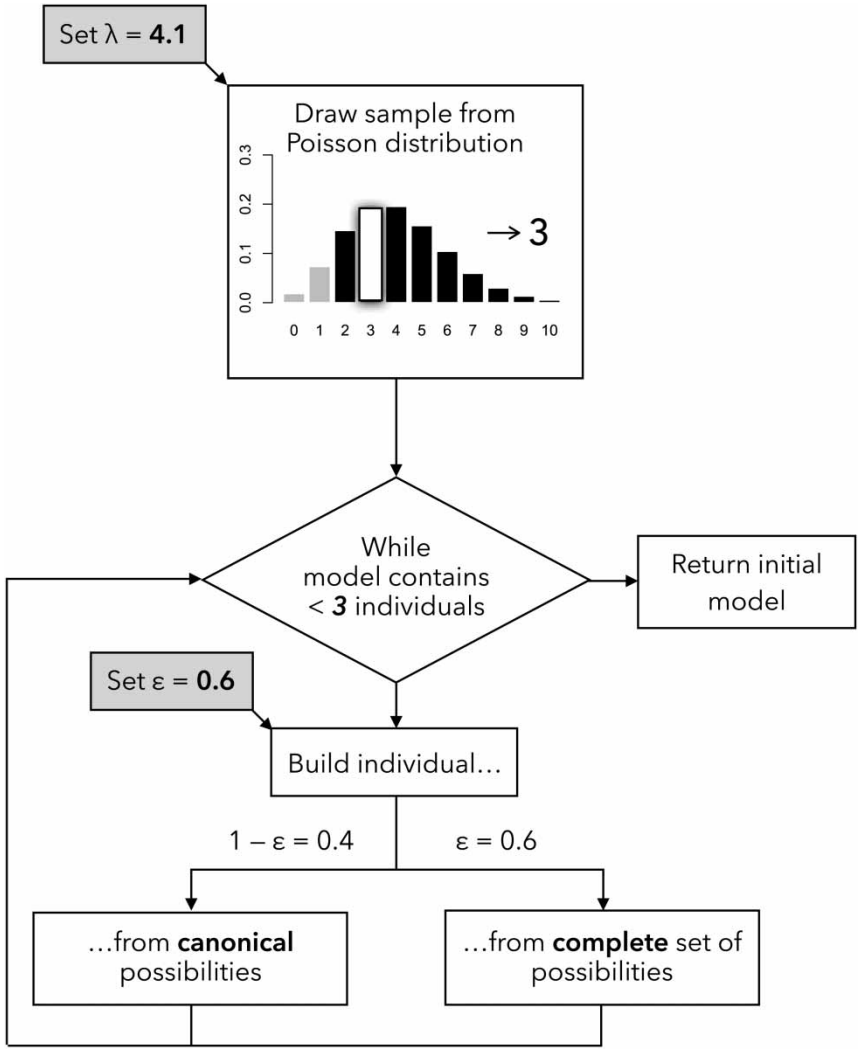

Figure 2. *The algorithm implemented in mReasoner for stochastically constructing initial mental models with* $\lambda = 4.1$ *and* $\varepsilon = 0.6$. *The system first draws a sample from a Poisson distribution with parameter* λ. *It uses the sample (in this case, 3) to set the number of individuals in the initial model. Each individual is added progressively by drawing a sample individual from one of two sets: the canonical set of possibilities, and the full set of possible individuals consistent with the premise. The resulting model is scanned to draw an initial conclusion.*

we use the program to simulate participants' inferences in order to find optimal settings that model the quantitative differences among inferences.

A summary of the predictions

The model theory implies that there are three main levels of difficulty in immediate inferences. The easiest inferences should be those in which the conclusion is identical to, or has an identical meaning to, the premise. Reasoners do not even need to construct a model of the premises to determine that the conclusion follows in these inferences: The identity of intensional representations suffices. These inferences are "zero-model", because they do not depend on any model of the premise. The intermediate level of difficulty

should be those intuitive "one-model" inferences that can be made from the initial model of the premise. The most difficult level are those "multiple-model" inferences that depend on an alternative model of the premises. Reasoners are likely to err by failing to search for an alternative model. An alternative account, of course, would use models to represent separate relations between the sets so that, for instance, an assertion such as, *All of the A are B*, would elicit two different models—Examples (8) and (9)—from the start. A key prediction of the present theory and its computer implementation is therefore that individuals should tend to err by basing their response on the initial model of a multiple-model inference.

Previous studies suggest plausible constraints on the values of parameters in the stochastic system. Reasoners are likely to construct models containing three to five individuals, as evinced in the diagrams that participants draw in making syllogistic inferences (Bucciarelli & Johnson-Laird, 1999). The setting of the corresponding parameter, λ, for one-model and multiple-model problems should therefore have a high probability of yielding such numbers, and so its optimal value should be between 3 and 5 for immediate inferences. The ε parameter varies from 0 to 1: When it is 0, reasoners only consider canonical possibilities, and when it is 1, they consider all possibilities. When reasoners consider remote possibilities at the outset of model construction, they should be accurate and fast; but, across a wide swathe of inferential domains (Johnson-Laird & Khemlani, 2014), remote possibilities appear difficult for reasoners to consider. Hence, the optimal value of ε for most domains should be $<.5$—that is, reasoners tend to focus on canonical individuals. Of course, some tasks should encourage reasoners to consider remote possibilities and call for higher values of ε. However, values of ε values $>.8$ should be most unlikely, because they would be contrary to the model theory's assumption that individuals rely on canonical models. An analogous argument can be made for the σ parameter. When it is 0, reasoners should accept almost every inference drawn from an initial model as valid. When it is 1, reasoners should deliberate about every inference, invariably finding alternative models if they exist. They should therefore be very accurate and very slow. However, the σ parameter is likely to depend on the nature of the inferential problems under investigation. Simple problems, such as immediate inferences, should yield values of σ around .5, because a search for alternative models is not highly demanding. More complicated inferences, such as those based on three or more premises (see Ragni, Khemlani, & Johnson-Laird, 2014), should call for significantly lower values of σ in order to fit the data. In sum, the success of the model theory in accounting for general results has implications for the values of the parameters likely to be optimal in accounting for immediate inferences. The optimal value of λ should be between 3 and 5, whereas lower values should yield too few individuals in a model, and higher values should yield too many. The proportion of canonical individuals in a model should call for an optimal value of $\varepsilon < .5$. And the optimal value of σ for the discovery of alternative models should be around .5.

The empirical studies

The aim of the empirical studies that we report was to test the model theory's account of immediate inferences. We begin with a reexamination of data from two experiments in Newstead and Griggs (1983) in which the participants drew conclusions about what is necessary given a quantified assertion. We then report three new experiments. Experiment 1 examined inferences about what is possible given a first-order quantified assertion. Experiment 2 examined inferences based on a second-order quantifier, *most of the A*. Experiment 3 examined a different task in which the participants had to decide whether two assertions could both be true at the same time. This task is equivalent to the assessment of the consistency of the two assertions. It is intimately related to deduction, and some logical systems test whether a deduction is valid by assessing the consistency of the premises with the negation of the conclusion (see, e.g., Jeffrey, 1981). The model theory provides an obvious account of how individuals determine whether or not assertions are consistent. They attempt to form a model of them. If they succeed,

they evaluate the assertions as consistent; otherwise, they evaluate them as inconsistent. Other current theories of reasoning, such as those based on formal rules of inference (e.g., Rips, 1994) or on probabilistic heuristics (e.g., Chater & Oaksford, 1999), have not yet explained how reasoners evaluate consistency.

A reanalysis of Newstead and Griggs (1983): Inferences about what is necessary

Newstead and Griggs (1983) reported two experiments examining immediate inferences of conclusions that follow of necessity, such as:

10. All of the As are Bs
    Does it follow that all of the Bs are As?

Their first experiment used assertions containing letters, as in (10), and their second experiment used assertions containing letters and assertions containing terms such as "artists" and "beekeepers". Both experiments examined all 32 possible immediate inferences based on four sorts of first-order premise: *All of the As are Bs*, *Some of the As are Bs*, *None of the As are Bs*, and *Some of the As are not Bs*; and on eight sorts of conclusion: the same four moods, but each with two orders of terms in the conclusions: *A–B*, and *B–A*.

The model theory predicts that zero-model inferences should be easier than one-model inferences—that is, inferences in which the initial model yields a correct evaluation, which in turn should be easier than multiple-model inferences in which only an alternative to the initial model establishes the correct response. According to the model theory, there are four zero-model inferences, 22 one-model inferences, and six multiple-model inferences. Table 3 presents these 32 inferences, their canonical mental models, and the percentages of accurate responses in the two experiments. These results corroborated the predictions of the model theory. Their first experiment yielded the following trend: 99% correct zero-model inferences, 78% correct one-model inferences, and 55% correct multiple-model inferences (Jonckheere trend test, $z = 2.70$, $p < .005$, in a by-materials analysis). The same trend occurred in the second experiment: 97% correct zero-model inferences, 76% correct one-model inferences, and 58% correct multiple-model inferences (Jonckheere trend test, $z = 2.60$, $p < .005$).`

We used the parameterized model to simulate accuracy data for 1000 participants. We first ran a "grid" search for optimal settings of the three parameters—that is, an exhaustive search through their possible settings guided by the participants' data (see, e.g., Busemeyer & Diederich, 2010). The large number of simulated participants allowed the system to converge on stable quantitative values for the parameters. We then synthesized a set of data based on these values and computed two sorts of correlation between the synthetic data and the results in Newstead and Griggs (1983), one for the three different levels of inference (zero-, one-, and multiple-model inferences), and one for the 32 individual inferences. Table 4 presents the parameter settings and goodness-of-fit metrics for both sorts of correlation. Because of the variability in multiple data points, the coefficient of determination is bound to yield misleadingly low values for the 32 individual inferences (Khemlani & Trafton, 2012; Roberts & Pashler, 2000, p. 363; Schunn & Wallach, 2005). Hence, we report Pearson's correlation coefficient, r, in assessing the model's fit with both the three sorts of inference and the individual inferences. The table reports the root mean squared error (*RMSE*) for both sorts of comparison. The quantitative predictions of the computational model closely matched the data over the three different levels of inference ($r = .98$ and $r = .96$ for Newstead and Griggs's Experiments 1 and 2, respectively) and over the 32 individual inferences ($r = .75$ and .78 for the two experiments). Appendices A and B provide quantitative model fits for the three sorts of inference and for the 32 inferences.`

The studies did not record latencies of correct responses, but the accuracy results were sufficiently convincing not to call for replication. In order to test some new predictions of the theory, our experiments examined some different sorts of immediate inference.

`

Table 3. *The 32 problems for conclusions about what is necessary in Newstead and Griggs, along with the correct response, the number of models required to obtain the correct response, a new model for a multiple-model inference, and the accuracy percentages in the two studies*

					Newstead & Griggs (1983)	
First assertion (and canonical initial model)	*Second assertion*	*Correct response*	*No. of models*	*Alternative model*	*Experiment 1 accuracy percentage*	*Experiment 2 accuracy percentage*
All of the A are B	*All of the A are B*	Necessary	Zero	—	100	98
A B	*All of the B are A*	Not necessary	Multiple	A B	67	63
A B				A B		
A B				¬A B		
	Some of the A are B	Necessary	One	—	73	56
	Some of the B are A	Necessary	One	—	65	79
	None of the A are B	Not necessary	One	—	100	100
	None of the B are A	Not necessary	One	—	92	90
	Some of the A are not B	Not necessary	One	—	98	97
	Some of the B are not A	Not necessary	One	—	58	43
Some of the A are B	*All of the A are B*	Not necessary	One	—	100	83
A B	*All of the B are A*	Not necessary	One	—	90	90
A ¬B	*Some of the A are B*	Necessary	Zero	—	100	98
A ¬B						
	Some of the B are A	Necessary	One	—	69	90
	None of the A are B	Not necessary	One	—	100	98
	None of the B are A	Not necessary	One	—	90	92
	Some of the A are not B	Not necessary	Multiple	A B	6	3
				A B		
				A B		
	Some of the B are not A	Not necessary	Multiple	A B	37	91
				A B		
				A B		
None of the A are B	*All of the A are B*	Not necessary	One	—	100	100
A ¬B	*All of the B are A*	Not necessary	One	—	98	95
A ¬B	*Some of the A are B*	Not necessary	One	—	98	92
¬A B	*Some of the B are A*	Not necessary	One	—	79	67
	None of the A are B	Not necessary	Zero	—	98	94
	None of the B are A	Necessary	One	—	54	59
	Some of the A are not B	Necessary	One	—	69	59
	Some of the B are not A	Necessary	One	—	54	47
Some of the A are not B	*All of the A are B*	Not necessary	One	—	100	98
A ¬B	*All of the B are A*	Not necessary	One	—	87	89
A ¬B						
¬A B						
	Some of the A are B	Not necessary	Multiple	A ¬B	17	8
				A ¬B		
				A ¬B		
	Some of the B are A	Not necessary	One	—	35	12
	None of the A are B	Not necessary	Multiple	A ¬B	98	87
				A ¬B		
				A B		
	None of the B are A	Not necessary	Multiple	A ¬B	85	83
				A ¬B		
				A B		
	Some of the A are not B	Necessary	Zero	—	98	98
	Some of the B are not A	Not necessary	Multiple	A ¬B	35	19
				A ¬B		
				A B		

Note: Newstead and Griggs (1983), Experiments 1 and 2.

Table 4. *The experiments for which synthetic data generated by mReasoner were fitted, the logical task in the experiment, the values of the three parameters in mReasoner, and metrics for their goodness of fit over the three different levels of inference and over the 32 individual inferences*

		Values of mReasoner's parameters			*By problem type*		*By immediate inference*	
Dataset	*Task*	λ	ε	σ	r	RMSE	r	RMSE
1. Newstead & Griggs (1983) Experiment 1	Necessity	4.0	.3	.4	.98	.12	.75	.21
2. Newstead & Griggs (1983) Experiment 2	Necessity	4.0	.3	.4	.96	.12	.78	.23
3. Experiment 1	Possibility	3.8	.2	.6	.99	.07	.62	.14
4. Experiment 2	Possibility	3.8	.4	.4	.99	.03	.85	.13
5. Experiment 3	Consistency	3.5	.6	.7	.99	.06	.63	.16

Note: The λ parameter stochastically varies the size of the mental model; the ε parameter specifies the proportion of individuals in a model that correspond to a canonical mental model; and the σ parameter specifies the probability that the program carries out a search for alternative models on any particular inference.

EXPERIMENT 1: INFERENCES ABOUT WHAT IS POSSIBLE

The aim of this experiment was to test the model theory's predictions for inferences about what is possible. A typical trial was:

11. All of the artists are bakers.
    Is it possible that all of the bakers are artists?

As in Newstead and Griggs's studies, the experiment examined all 32 possible sorts of first-order inference, and in this domain valid inferences comprise four zero-model inferences, 12 one-model inferences, 6 multiple-model inferences in which the conclusion follows, and 10 multiple-model inferences in which the conclusion does not follow.

Method

Participants

Twenty-six participants completed the study on Mechanical Turk for monetary compensation (see Paolacci, Chandler, & Ipeirotis, 2010, for an analysis of the validity of results from this platform). None of the participants by their own account had received any training in logic, and they were all native speakers of English.

Design and materials

The participants carried out all 32 inferences based on four sorts of premise (*All of the A are B*, *At least some of the A are B*, *None of the A are B*, and *At least some of the A are not B*) and 8 sorts of conclusion (the four moods × two arrangements of the terms: *A–B*, and *B–A*). For each inference, the participants responded to a yes/no question about whether a conclusion was possible. The contents of the inferences were based on nouns referring to common vocations. We devised a list of 32 pairs of such vocations, which were assigned at random to the inferences to make two separate lists of inferences. The inferences were presented to each participant in a different random order.

Procedure

The study was administered using an interface written in PHP, Javascript, and HTML. On each trial, participants read the premise, and, when ready, they pressed a button marked "Next", which replaced the premise with a question concerning the immediate inference, e.g., "Is it possible that all of the bakers are artists?" They responded by pressing one of two buttons labelled, "Yes, it's possible" and "No, it's impossible". The program recorded whether or not their response was correct and its latency (to the nearest ms). The instructions stated that the task

Table 5. *The 32 inferences yielding conclusions about what is possible in Experiment 1, along with the correct response, the number of models required to obtain the correct response, a new model for a multiple-model inference, the accuracy percentage, and the mean latency*

First assertion (and canonical initial model)	*Second assertion*	*Correct response*	*No. of models*	*Alternative model*	*Accuracy percentage*	*Latency (in s)*
All of the A are B	*All of the A are B*	Possible	Zero	—	100	3.98
A B	*All of the B are A*	Possible	One	—	88	5.25
A B	*Some of the A are B*	Possible	One	—	96	5.80
A B	*Some of the B are A*	Possible	One	—	96	6.01
	None of the A are B	Not possible	One	—	96	3.60
	None of the B are A	Not possible	One	—	81	6.71
	Some of the A are not B	Not possible	One	—	92	4.89
	Some of the B are not A	Possible	Multiple	A B A B ¬A B	58	5.70
Some of the A are B	*All of the A are B*	Possible	Multiple	A B A B A B	69	5.67
A B A ¬B A ¬B	*All of the B are A*	Possible	Multiple	A B A B A B	73	5.32
	Some of the A are B	Possible	Zero	—	100	4.19
	Some of the B are A	Possible	One	—	100	4.75
	None of the A are B	Not possible	One	—	96	4.35
	None of the B are A	Not possible	One	—	96	5.71
	Some of the A are not B	Possible	One	—	96	4.85
	Some of the B are not A	Possible	One	—	96	5.91
None of the A are B	*All of the A are B*	Not possible	One	—	100	4.13
A ¬B A ¬B ¬A B	*All of the B are A*	Not possible	One	—	85	5.95
	Some of the A are B	Not possible	One	—	92	5.56
	Some of the B are A	Not possible	One	—	73	5.88
	None of the A are B	Possible	Zero	—	96	4.35
	None of the B are A	Possible	One	—	92	6.60
	Some of the A are not B	Possible	One	—	77	5.56
	Some of the B are not A	Possible	One	—	85	8.71
Some of the A are not B	*All of the A are B*	Not possible	One	—	85	5.60
A ¬B A ¬B ¬A B	*All of the B are A*	Possible	Multiple	A ¬B A ¬B A B	38	6.63
	Some of the A are B	Possible	Multiple	A ¬B A ¬B A B	92	4.72
	Some of the B are A	Possible	Multiple	A ¬B A ¬B A B	92	5.84
	None of the A are B	Possible	One	—	62	6.40
	None of the B are A	Possible	One	—	38	7.62
	Some of the A are not B	Possible	Zero	—	96	4.59
	Some of the B are not A	Possible	One	—	92	5.95

was to respond to questions about a series of assertions concerning what was possible given the truth of an assertion. The participants carried out three practice trials in order to familiarize themselves with the task before they proceeded to the experiment proper.

Results and discussion

Table 5 presents the percentages of correct responses for all 32 inferences and the latencies of the correct responses. The results corroborated the theory's predicted trend in accuracy: 98% correct zero-model inferences, 84% correct one-model inferences, and 71% correct multiple-model inferences (Page's trend test, $L = 340.0$, $z = 3.88$, $p < .0001$). To assess that the trend did not depend on just one sort of inference, we examined the pairwise differences between the accuracies. Participants made more correct responses for zero-model inferences than for one-model inferences (Wilcoxon test, $z = 3.69$, $p < .0005$), and more correct responses for one-model inferences than for multiple-model inferences (Wilcoxon test, $z = 2.73$, $p < .01$).`

The mean latencies for correct responses yielded an analogous trend: 4.26 s for zero-model inferences, 5.26 s for one-model inferences, and 5.11 s for multiple-model inferences (Page's trend test, $L = 333.0$, $z = 2.91$, $p < .005$). Pairwise analyses showed that zero-model inferences yielded shorter latencies than one-model inferences (Wilcoxon test, $z = 3.72$, $p < .0002$), but the difference in latencies between one-model and multiple-model inferences was not reliable (Wilcoxon test, $z = .52$, $p = .60$) and in a direction opposite to the prediction. The latencies therefore failed to reveal the difference in difficulty reflected in the accuracy data. It may be that latencies are sensitive to more than just the number of models necessary for an inference. Table 5 suggests one such factor: Five of the 12 one-model inferences contained the negative quantifier, *None of the* __, whereas this quantifier did not occur in the multiple-model inferences. When inferences containing the quantifier were dropped from the analysis, the mean latency for one-model inferences fell to 5.04 s, but the difference between the two conditions was still not reliable (Wilcoxon test, $z = 0.42$, $p > .5$).

We used the parameterized system to generate a synthetic dataset of 1000 participants and to find optimal settings for the three parameters: the size of models, their canonicity, and the search for alternative models. Table 4 presents the parameter settings and their goodness-of-fit metrics both over the three levels of inference ($r = .99$, $RMSE = .07$), and over the 32 individual inferences ($r = .62$, $RMSE = .14$). Appendices A and B provide quantitative model fits by the levels of inference and the individual inferences.

EXPERIMENT 2: SECOND-ORDER QUANTIFIERS

The previous tests of the theory have depended on first-order quantifiers. Experiment 2 extended the investigation to include the second-order quantifier, *most of the* __, which cannot be defined using the quantifiers of first-order logic (Barwise & Cooper, 1981), and it examined immediate inferences of conclusions about what is possible.

Method

Participants

Forty participants from the same population as those in Experiment 1 were recruited on Mechanical Turk.

Design, materials, and procedure

On each trial, participants evaluated an inference based on a quantified premise and a conclusion about a quantified possibility, for example:

12. Most of the artists are barbers.
    Is it possible that all of the barbers are artists?

The experiment examined 32 sorts of inference, consisting of a premise that was in one of four moods: *All of the A are B*, *Most of the A are B*, *None of the A are B*, and *Most of the A are not B*. The conclusions were of the same sort, but also

`

Table 6. *The 32 inferences yielding conclusions about what is possible in Experiment 2 along with the correct response, the number of models required to obtain the correct response, a new model for a multiple-model inference, the accuracy percentage, and the mean latency*

First assertion (and canonical initial model)	*Second assertion*	*Correct response*	*No. of models*	*Alternative model*	*Accuracy percentage*	*Latency (in s)*
All of the A are B	*All of the A are B*	Possible	Zero	—	98	5.72
A B	*All of the B are A*	Possible	One	—	78	11.14
A B	*Most of the A are B*	Not possible	One	—	63	8.48
A B	*Most of the B are A*	Possible	Multiple	A B A B ¬A B	80	8.22
	None of the A are B	Not possible	One	—	93	6.78
	None of the B are A	Not possible	One	—	93	9.71
	Most of the A are not B	Not possible	One	—	95	11.33
	Most of the B are not A	Possible	Multiple	A B A B ¬A B ¬A B ¬A B	40	10.28
Most of the A are B	*All of the A are B*	Not possible	One	—	40	8.20
A B	*All of the B are A*	Possible	One	—	55	8.83
A B						
A ¬B						
	Most of the A are B	Possible	Zero	—	100	6.46
	Most of the B are A	Possible	One	—	90	8.13
	None of the A are B	Not possible	One	—	90	7.87
	None of the B are A	Not possible	One	—	85	8.57
	Most of the A are not B	Not possible	One	—	93	6.20
	Most of the B are not A	Possible	Multiple	A B A B ¬A B ¬A B ¬A B	40	15.69
None of the A are B	*All of the A are B*	Not possible	One	—	98	5.75
A ¬B	*All of the B are A*	Not possible	One	—	85	12.19
A ¬B						
¬A B						
	Most of the A are B	Not possible	One	—	95	7.21
	Most of the B are A	Not possible	One	—	93	11.28
	None of the A are B	Possible	Zero	—	100	6.60
	None of the B are A	Possible	One	—	98	19.74
	Most of the A are not B	Not possible	One	—	40	11.57
	Most of the B are not A	Not possible	One	—	83	9.82
Most of the A are not B	*All of the A are B*	Not possible	One	—	90	8.80
A ¬B	*All of the B are A*	Possible	One	—	38	11.21
A ¬B						
A B						
	Most of the A are B	Not possible	One	—	95	7.62

(*Continued overleaf*)

Table 6. *Continued*

First assertion (and canonical initial model)	*Second assertion*	*Correct response*	*No. of models*	*Alternative model*	*Accuracy percentage*	*Latency (in s)*
	Most of the B are A	Possible	Multiple	A ¬B A ¬B A ¬B A B A B ¬A B	40	8.21
	None of the A are B	Not possible	One	–	38	11.69
	None of the B are A	Not possible	One	–	53	19.91
	Most of the A are not B	Possible	Zero	–	100	7.25
	Most of the B are not A	Possible	Multiple	A ¬B A ¬B A ¬B A B ¬A B ¬A B	93	10.93

varied in the order of the two terms. The contents used the same list of 32 pairs of vocations as that in Experiment 1, which were assigned at random to the forms of inference for each participant. Each participant received the inferences in a random order.

Results and discussion

Table 6 presents the percentages of correct responses for each of the inferences. The results yielded the following trend: 99% correct zero-model inferences, 85% correct one-model inferences, and 67% correct multiple-model inferences (Page's trend test, $L = 407.0$, $z = 8.16$, $p < .0001$). The participants were more accurate with zero-model inferences than with one-model inferences (Wilcoxon test, $z = 5.31$, $p < .0001$), and more accurate with one-model inferences than with multiple-model inferences (Wilcoxon test, $z = 4.75$, $p < .0001$).`

The mean latencies for correct responses yielded a reliable trend: 6.67 s for zero-model inferences, 8.61 s for one-model inferences, and 9.29 s for multiple-model inferences (Page's trend test, $L = 500$, $z = 5.05$, $p < .0001$). The zero-model inferences had shorter latencies than the one-model inferences (Wilcoxon test, $z = 3.71$, $p < .0005$), and the one-model inferences had shorter latencies than the multiple-model inferences, but the difference was only marginally reliable (Wilcoxon test, $z = 1.56$, $p = .06$). But, when problems containing the negative quantifier, *None of the* __, were excluded from the analysis, the mean latency for one-model inferences dropped to 8.06 s, and the difference between one-model and multiple-model inferences was reliable (Wilcoxon test, $z = 2.26$, $p < .01$).

As in the previous study, the parameterized model was used to generate a synthetic dataset of 1000 participants and to carry out a search for optimal settings for the three parameters. Table 4 presents these values and metrics of their goodness-of-fit both across the three sorts of inference ($r = .99$, $RMSE = .04$) and across the individual inferences ($r = .85$, $RMSE = .13$). Appendices A and B provide quantitative model fits for the three sorts of inference and the 32 individual inferences.

In syllogistic reasoning, individuals use heuristics that depend on the polarity of the premises—that is, whether or not they are both affirmative or contain at least one negative—and on the quantifiers—that is, whether or not they are both universal (*all* and *no*) or contain at least one existential (*some*), and on the arrangement of the terms in the premises (see Chater & Oaksford, 1999;

Khemlani et al., 2013). The present results suggested that reasoners also rely on a heuristic depending on polarity. When the premise and conclusion have the same polarity, affirmative or negative, participants tend to accept the conclusion as possible; but otherwise they tend to reject it as impossible (80% of all responses could be predicted on this basis, Wilcoxon test, $z = 5.52$, $p < .0001$). The model theory explains the origin of the heuristic. The initial model of any sort of premise yields only conclusions of the same polarity as the premise. Individuals may therefore learn this pattern. Those conclusions that do not hold in the initial model, including those of a different polarity, call for a search for an alternative model. They are multiple-model deductions and are therefore difficult.

The results also showed that participants vacillate in their interpretation of "most". Consider this inference:

13. All of the doctors are golfers.
    Is it possible that most of the doctors are golfers?

Many participants rejected this inference, whereas many accepted the converse inference from *most* to the possibility of *all*. What is at stake here is whether individuals accept the Gricean implicature that *most* implies *not all* (see Grice, 1989). The results suggest that naïve individuals tend not to make the implicature: They accept that *most* does not rule out *all*, but they do take *all* to rule out *most*. However, not all participants concur, and so no uniform view exists about the matter.

EXPERIMENT 3: JUDGEMENTS OF CONSISTENCY

As we pointed out in the introduction, the evaluation of the consistency of a set of assertions is closely related to valid deduction. According to the model theory, it calls for individuals to try to construct a model of the assertions, and if they can so then they evaluate the assertions as consistent. In the case of a pair of assertions, if the second assertion has the same meaning as the first (a zero-model inference) then the task should be the easiest of all, and if the second assertion holds in the initial model of the first assertion (a one-model inference) the task should be easier than if the second assertion holds only in an alternative model of the first assertion (a multiple-model inference). The experiment tested this predicted trend with the 32 possible pairs of assertions. The model theory predicts that four are zero-model inferences, 12 are one-model inferences, and 6 are multiple-model inferences (see Table 7). The remaining 10 inferences are those in which the two assertions are inconsistent.`

Method

Participants

Twenty-four participants from the same population as before were recruited on Mechanical Turk.

Design, materials, and procedure

The experiment examined all 32 possible pairs of quantified assertions: four sorts of first assertion, *All of the A are B*, *At least some of the A are B*, *None of the A are B*, and *At least some of the A are not B*, and eight sorts of second assertion (the same four moods and two orders of terms in the second premise). Because naïve individuals are often uncertain about the meaning of "consistent", the participants' task was to answer an equivalent question, "Can both of these statements be true at the same time?" They responded by pressing a button marked "Yes, they can" or "No, they cannot". Each participant received the 32 inferences in a different random order. The experiment used the same pairs of occupations as those in the previous experiments.

Results

Table 7 presents the percentages of correct responses and latencies for all 32 inferences in Experiment 3. The participants were correct for 99% zero-model inferences, for 87% one-model inferences, and for 74% multiple-model inferences (Page's trend test, $L = 317.5$, $z = 4.26$, $p < .0001$).

Table 7. *The 32 inferences about the consistency of the assertions in Experiment 3, along with the correct response, the number of models required to obtain the correct response, a new model for a multiple-model inference, the accuracy percentage, and the mean latency*

First assertion (and canonical initial model)	*Second assertion*	*Correct response*	*No. of models*	*Alternative model*	*Accuracy percentage*	*Latency (in s)*
All of the A are B	*All of the A are B*	Consistent	Zero	—	100	4.98
A B	*All of the B are A*	Consistent	One	—	100	5.62
A B						
A B						
	Some of the A are B	Consistent	One	—	79	5.85
	Some of the B are A	Consistent	One	—	83	7.56
	None of the A are B	Inconsistent	One	—	100	4.86
	None of the B are A	Inconsistent	One	—	96	8.06
	Some of the A are not B	Inconsistent	One	—	96	8.24
	Some of the B are not A	Consistent	Multiple	A B	54	9.30
				A B		
				¬A B		
Some of the A are B	*All of the A are B*	Consistent	Multiple	A B	88	5.63
				A B		
				A B		
A B	*All of the B are A*	Consistent	Multiple	A B	88	9.37
A ¬B				A B		
A ¬B				A B		
	Some of the A are B	Consistent	Zero	—	100	5.77
	Some of the B are A	Consistent	One	—	100	6.89
	None of the A are B	Inconsistent	One	—	100	5.93
	None of the B are A	Inconsistent	One	—	83	8.66
	Some of the A are not B	Consistent	One	—	100	6.92
	Some of the B are not A	Consistent	One	—	96	7.49
None of the A are B	*All of the A are B*	Inconsistent	One	—	100	5.42
A B	*All of the B are A*	Inconsistent	One	—	88	9.33
A ¬B						
¬A ¬B						
	Some of the A are B	Inconsistent	One	—	96	6.28
	Some of the B are A	Inconsistent	One	—	79	8.47
	None of the A are B	Consistent	Zero	—	96	4.58
	None of the B are A	Consistent	One	—	96	5.90
	Some of the A are not B	Consistent	One	—	63	13.35
	Some of the B are not A	Consistent	One	—	83	8.80
Some of the A are not B	*All of the A are B*	Inconsistent	One	—	96	7.92
A B	*All of the B are A*	Consistent	Multiple	A ¬B	25	12.88
A ¬B				A ¬B		
¬A ¬B				A B		
	Some of the A are B	Consistent	Multiple	A ¬B	96	8.18
				A ¬B		
				A B		
	Some of the B are A	Consistent	Multiple	A ¬B	96	10.48
				A ¬B		
				A B		
	None of the A are B	Consistent	One	—	75	9.42
	None of the B are A	Consistent	One	—	71	11.93
	Some of the A are not B	Consistent	Zero	—	100	6.34
	Some of the B are not A	Consistent	One	—	100	7.78

The participants were more accurate with zero-model inferences than with one-model inferences (Wilcoxon test, $z = 3.19$, $p < .005$) and more accurate with one-model inferences than with multiple-model inferences (Wilcoxon test, $z = 2.57$, $p < .05$). Likewise, the mean latencies for the correct "yes" responses were 5.43 s for zero-model inferences, 6.91 s for one-model inferences, and 8.27 s for multiple-model inferences (Page's trend test, $L = 323$, $z = 5.05$, $p < .0001$). The latencies for zero-model inferences were shorter than those for one-model inferences (Wilcoxon test, $z = 4.05$, $p < .0001$) and shorter for one-model inferences than for multiple-model inferences (Wilcoxon test, $z = 2.71$, $p < .01$). Like the previous experiment, the latencies for one-model inferences dropped to 6.55 s when inferences that included the negative quantifier, *None of the* __, were removed from the analysis.

Once again, the parameterized model was used to generate a synthetic dataset of 1000 participants and to carry out a search for optimal settings for the three parameters. Table 4 presents these values and metrics of their goodness-of-fit both over the three levels of inference ($r = .99$, $RMSE = .03$) and over the 32 individual inferences ($r = .63$, $RMSE = .16$). Appendices A and B provide quantitative model fits by the levels of inference and the individual inferences. These results corroborated the model theory's qualitative and quantitative predictions about the evaluation of the consistency of pairs of quantified assertions.

GENERAL DISCUSSION

Aristotle was the first Western logician to formulate the logic of immediate inferences from quantified assertions, for example:

14. Some of the artists are bakers. Therefore, some of the bakers are artists.

Yet, despite many years of investigation (e.g., Wilkins, 1928), no prior theory accounts for the mental processes underlying inferences from them about what is necessary, possible, or consistent. Indeed, previous studies focused on the first task alone (e.g., Begg & Harris, 1982; Newstead & Griggs, 1983). A logical view of inference implies the use of formal rules of inference for quantifiers, but such theories have been framed only for the first-order predicate calculus (e.g., Rips, 1994), which treats quantifiers, such as *for any artist*, as ranging over individuals. Syllogistic reasoning similarly can be based on the properties of a single representative individual (e.g., Politzer, 2011). Quantifiers such as *most of the artists*, and *fewer than half of the artists* cannot be defined in first-order logic (Barwise & Cooper, 1981), and inferences from them cannot be based on representative individuals. They call for quantification over properties (second-order logic), or, equivalently for set-theoretic relations, and theories of reasoning with them need to take into account relations between sets. Likewise, neither logical theories of reasoning nor probabilistic theories (e.g., Oaksford & Chater, 2009) at present offer an account of how naïve individuals reason about possibility or consistency.

The model theory accounts for performance in all these tasks with first-order and second-order quantifiers, and its computational implementation carries them out, yielding predicted errors and correct performance. A conclusion about what is necessary, possible, or consistent follows at once if the conclusion is synonymous with the premise. Otherwise, individuals construct a mental model of the premise, and the conclusion about what is necessary follows if it holds in this initial model and in any alternative model. A conclusion about what is possible follows if it holds in the initial model, or, failing that, in an alternative model. And assertions are consistent if and only if there is a model in which they each hold.

The results of two experiments due to Newstead and Griggs (1983) corroborated the predicted trend in difficulty for inferences about what is necessary. Zero-model inferences—those for which the premise and conclusion were identical—yielded a greater percentage of correct conclusions than those for which the initial model—one representing only canonical individuals (see Table 1)—yielded the correct conclusion, which in turn yielded a greater percentage of correct conclusions than those for which the correct conclusions

depended on finding an alternative model, which represents atypical (noncanonical) individuals consistent with the meaning of the premise. With inferences about what is possible, as in:

15. Some of the artists are bakers. Is it possible that all of the bakers are artists?

the same trend occurred in both the accuracy and speed of responses (Experiment 1). Models can represent the relations between sets needed for second-order quantifiers, such as *most of the artists* (Johnson-Laird, 1983, p. 137 et seq.), and the same trend over zero-, one-, and multiple-model inferences occurred with this quantifier in both affirmative and negative premises (Experiment 2). The evaluation of the consistency of pairs of quantified assertions, which our participants carried out by answering the question *Can both of these statements be true at the same time?* also bore out the predicted trend in difficulty over zero-, one-, and multiple-model inferences (Experiment 3).

Four of the 32 inferences in each study were "zero-model" inferences—that is, inferences in which the conclusion is identical to the premise. The equivalence obviates the need to build a model, and so the model theory predicts that participants should be the fastest and most accurate on zero-model inferences. The robust trends of increasing difficulty over zero-, one-, and multiple-model inferences did not depend solely on the ease of zero-model inferences. Reasoners are consistently more accurate for one-model inferences than multiple-model inferences, and they tend to have shorter latencies of correct responses for one-model inferences than for multiple-model inferences, especially when the negative quantifier, *None of the* __, is removed from the analysis. In the resulting analyses, the difference was not reliable in Experiment 1, but reliable in Experiments 2 and 3.

The theory of mental models postulates that the meanings of quantified assertions—their intensions—can play a role in heuristics. The present results may also reflect the use of heuristics, and the possibility seemed very likely in Experiment 2, which included assertions based on the quantifier, *most of the* __. Participants tended to accept a conclusion as possible only if it had the same polarity, affirmative or negative, as the premise. This heuristic, however, is compatible with the model theory: The initial models of premises yield as possible only conclusions of the same polarity as the premise. Indeed, this factor could explain how individuals acquire the heuristic. A corollary is that those conclusions that are possible, but that do not hold in the initial model, call for a search for an alternative model, and these multiple-model inferences are the most difficult of all to draw.

When reasoners are given an inference with a conclusion to be evaluated, they often work backwards by testing whether the given conclusion follows from the premises (Van der Henst, Yang, & Johnson-Laird, 2002). Participants in the present studies may have used a similar strategy. For example, given the following sort of immediate inference:

16. All of A are B. Therefore, is it possible that most of the B are not A?

Reasoners might build a canonical model of the putative conclusion, such as:`

B ¬A
B ¬A
B A

The premise is true in this model, and so by working backwards, the strategy requires only one model in order to obtain the correct inference. However, reasoners who work forwards will tend to start with a canonical model of the premise, such as:`

A B
A B
A B

They now need to search successfully for an alternative model in order to reach a correct conclusion. The stochastic system could include different inferential strategies, and it may then yield an even closer fit to the data.

Could the stochastic system rely on a better set of parameters than the present ones? The current system uses two parameters to control the construction of models: The λ parameter constrains the

number of individuals in a model, and the ε parameter constrains the chances that the model represents atypical (noncanonical) individuals drawn from the full set satisfying the premise. The σ parameter constrains the chances that the system finds an alternative to the initial model. The mathematician, von Neumann, remarked that with four parameters he could fit an elephant, and with five he could make it wiggle its trunk (Dyson, 2004). So, the question is whether a *smaller* number of parameters could fit the data as well. For example, perhaps we could use ε parameter to select the proportion of different individuals and then construct a model with the smallest number of individuals that can represent them all. However, we suspect that if a stochastic system is to model syllogistic reasoning, it may be necessary to split the σ parameter into two, with one parameter determining whether individuals search for alternative models and another determining the characteristics of the search including the likelihood of it finding a refutation of a putative conclusion. Although the present system fitted the data well, its characteristic error was to overestimate the accuracy of the participants' performance, especially on problems concerning negative assertions (see Appendix B). The one exception to this trend is in the fit with Experiment 2, which examined the quantifier, *Most of the* _, and which did not include assertions containing the quantifier, *Some of the* _. Hence, a parameter concerning the different moods of assertion might improve the system's fit. It could be relevant to a quantifier's polarity in a conclusion, and to the slower responses that *None of the* _ tended to elicit in our studies.

In sum, the model theory explained the participants' performance in making 160 separate immediate inferences (32 inferences in five experiments concerning what is necessary, possible, and consistent). It predicted the participants' relative accuracy both over three main sorts of inference (zero-, one-, and multiple-models) and over the individual inferences in these sorts. It also predicted the relative latencies of correct responses. Other sorts of theory may be able to accommodate these data, including the results with the unorthodox quantifier, *most of the* _, and those in evaluating the consistency of assertions. However, the theory of mental models seems to be the only current account that offers an explanation of the empirical phenomena of immediate inferences with quantifiers. It implies that naïve reasoners—those who have not mastered logic or set theory—do not manipulate the logical forms of premises, but instead envisage the situations to which the premises refer. Reasoning is therefore mental simulation rather than formal manipulation. Likewise, if the theory is correct, probabilities govern the process of reasoning with *all*, *most*, *some*, and other quantifiers, and probabilities can be derived from them (Khemlani, Lotstein, & Johnson-Laird, 2014), rather than occurring as an intrinsic part of their mental representation.

Original manuscript received 23 July 2014
Accepted revision received 23 December 2014

REFERENCES

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*, 159–219.

Begg, I., & Harris, G. (1982). On the interpretation of syllogisms. *Journal of Verbal Learning and Verbal Behavior*, *21*, 595–620.

Boole, G. (1854). *An investigation of the laws of thought*. London: Macmillan.

Boolos, G. (1984). On "syllogistic inference". *Cognition*, *17*, 181–182.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*, 247–303.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. New York, NY: Sage.

Ceraso, J., & Provitera, A. (1971). Sources of error in syllogistic reasoning. *Cognitive Psychology*, *2*, 400–410.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*, 191–258.

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method*. London: Routledge & Kegan Paul.

Deshpande, J. V., Gore, A. P., & Shanubhogue, A. (1995). *Statistical analysis of nonnormal data*. New York: Wiley.

Dyson, F. (2004). A meeting with Enrico Fermi. *Nature*, *427*, 297.

Erickson, J. R. (1974). A set analysis theory of behavior in formal syllogistic reasoning tasks. In R. Solso (Ed.), *Loyola symposium on cognition* (Vol. *2*, pp. 305–330). Hillsdale, NJ: Lawrence Erlbaum Associates.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, *86*, 223–251.

Grice, H. P. (1989). *Studies in the ways of words*. Cambridge, MA: Harvard University Press.

Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, *35*, 2075–2087.

Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York: McGraw-Hill.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford, UK: Oxford University Press.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*, 18243–18250.

Johnson-Laird, P. N., & Bara, B. G. (1984). Logical expertise as a cause of error: A reply to Boolos. *Cognition*, *17*, 183–184.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P. N., & Khemlani, S. (2014). Toward a unified theory of reasoning. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. *59*. pp. 1–42). New York: Elsevier.

Khemlani, S., & Johnson-Laird, P. N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory & Cognition*, *37*, 615–623.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*, 427–457.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*, 4–20.

Khemlani, S., Lotstein, M., & Johnson-Laird, P. N. (2014). Naive probability: Model-based estimates of unique events. *Cognitive Science*. doi:10.1111/cogs.12193.

Khemlani, S., & Trafton, J. G. (2012). mReactr: A computational theory of deductive reasoning. In N. Miyake, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 581–586). Austin, TX: Cognitive Science Society.

Khemlani, S., Trafton, J. G., & Johnson-Laird, P. N. (2013). Deduction as stochastic simulation. In R. West & T. Stewart (Eds.), *Proceedings of the 12th international conference on cognitive modeling* (pp. 297–302). Retrieved from http://iccm-conference.org/2013-proceedings/

Newstead, S. E., & Griggs, R. A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. *Journal of Verbal Learning and Verbal Behavior*, *22*, 535–546.

Oaksford, M., & Chater, N. (2009). Precis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral & Brain Sciences*, *32*, 69–84.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Peirce, C. S. (1931–1958). C. Hartshorne, P. Weiss, & A. Burks (Eds.), *Collected papers of Charles Sanders Peirce* (Vols. *8*). Cambridge, MA: Harvard University Press.

Politzer, G. (2011). Solving natural syllogisms. In K. Manktelow, D. Over, & S. Elqayam (Eds.), *The science of reason: A Festschrift for Jonathan St BT Evans* (pp. 19–36). London: Psychology Press.

Politzer, G., van der Henst, J. B., Luche, C. D., & Noveck, I. A. (2006). The interpretation of classically quantified sentences: A set-theoretic approach. *Cognitive Science*, *30*, 691–723.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*, 533–566.

Ragni, M., Khemlani, S. S., & Johnson-Laird, P. N. (2014). The evaluation of the consistency of quantified assertions. *Memory & Cognition*, *42*, 53–66.

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Schunn, C., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), *Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115–154). Saarbrücken, Germany: University of Saarland Press.

Störring, G. (1908). Experimentelle Untersuchungen über einfache Schlussprozesse [Experimental investigations of simple inference processes]. *Archiv für die gesamte Psychologie*, *11*, 1–27.

Van der Henst, J.-B., Yang, Y., & Johnson-Laird, P. N. (2002). Strategies in sentential reasoning. *Cognitive Science*, *26*, 425–468.

Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology*, *102*, 5–83.

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

APPENDIX A

The goodness of fit of the model theory for three sorts of inference in five experiments.

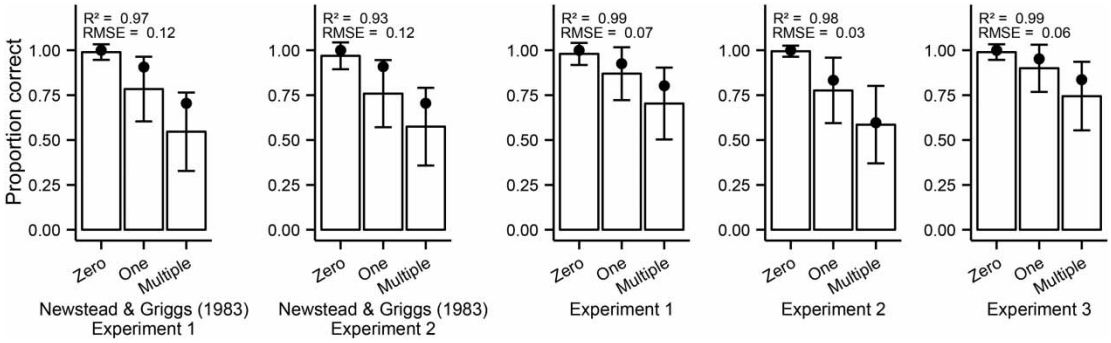

Figure A1. *Observed (histograms with error bars) and predicted (circles) proportions of correct inferences for the three sorts of inference (i.e., zero-, one-, and multiple-model) across the five datasets under investigation. Error bars show 95% confidence intervals. RMSE = root mean squared error.*

APPENDIX B

The goodness of fit of the model theory for the 32 individual sorts of inference in five experiments.

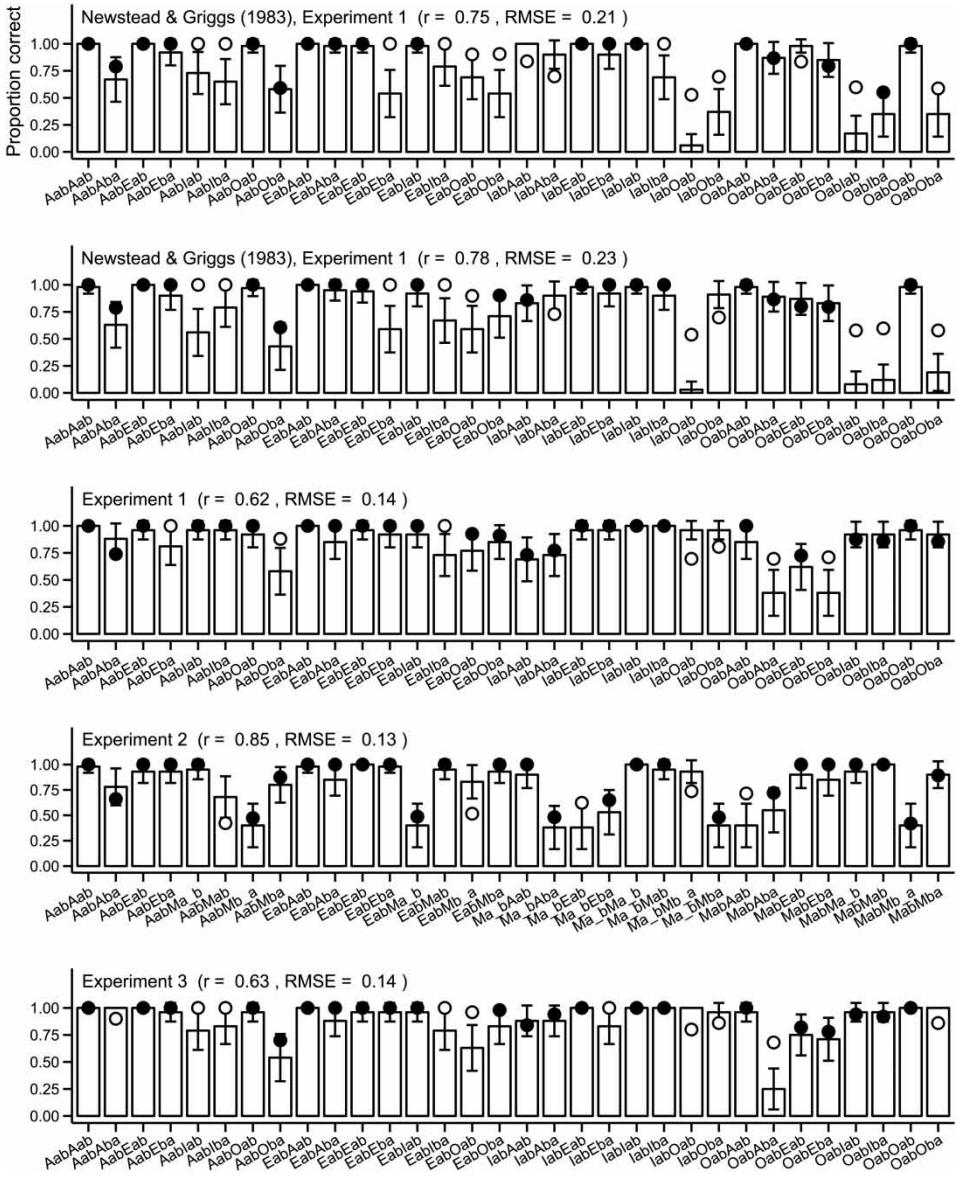

Figure B1. *Observed (histograms and error bars) and predicted (circles) proportions of correct response for 32 immediate inferences across the five data sets under investigation. Error bars reflect 95% confidence intervals. Black circles indicate when the predictions fell within the confidence interval of the observed proportion of correct responses, while white circles indicate deviations from the predictions and the observation. Premises and conclusions are stated using scholastic abbreviations, as follows: Aab = All of the A are B; Iab = Some of the A are B; Mab = Most of the A are B; Eab = None of the A is a B; Oab = Some of the A are not B; Ma_b = Most of the A are not B. Each inference is abbreviated as a premise concatenated with its putative conclusion, e.g., AabAba denotes the inference: All of the A are B. Does it follow that all of the B are A?*