How people differ in syllogistic reasoning

Sangeet Khemlani[1] and P.N. Johnson-Laird[2,3]
sangeet.khemlani@nrl.navy.mil, phil@princeton.edu
[1]Naval Research Laboratory, Washington, DC 20375 USA
[2]Princeton University, Princeton NJ 08540 USA
[3]New York University, New York, NY 10003, USA

Abstract

Psychologists have studied syllogistic inferences for more than a century, but no extant theory gives an adequate account of them. Reasoners appear to reason using different strategies. A complete account of syllogisms must therefore explain these strategies and the resulting differences from one individual to another in the patterns of conclusions that they draw. We propose a dual-process theory that solves these two problems. It is based on the manipulation of mental models, i.e., iconic simulations of possibilities. We also propose a new way in which to analyze individual differences, which depends on implementing a stochastic computer program. The program, *mReasoner*, generates an initial conclusion by building and scanning a mental model. It can vary four separate factors in the process: the *size* of a model, its *contents*, the propensity to consider *alternative models*, and the propensity to *revise its heuristic conclusions*. The former two parameters control intuitive processes and the latter two control deliberative processes. The theory accounts for individual differences in an early study on syllogisms (Johnson-Laird & Steedman, 1978). The computational model provides an algorithmic account of the different processes on which three subsets of participants relied (Simulation 1). It also simulates the performance of each individual participant in the study (Simulation 2). The theory and its implementation constitute the first robust account of individual differences in syllogistic reasoning.

Keywords: syllogisms, mental models, mReasoner, individual differences, deduction, counterexamples.

Introduction

It may be surprising to practitioners of the cognitive sciences that syllogistic reasoning is difficult to explain. After all, the first empirical study on syllogisms was carried out over a hundred years ago (Störring, 1908). Twelve separate theories now exist to explain reasoning by syllogism. And syllogisms are very simple inferences, such as:

(1) All of the architects are bankers.
Some of the bankers are not chefs.
What, if anything, follows?

The two premises each contain a single quantified term, such as, "all of the architects", and they can be in one of four separate *moods*, shown below (with their Scholastic abbreviations in parentheses):

All a are b. (*Aab*)	No a are b. (*Eab*)
Some a are b. (*Iab*)	Some a are not b. (*Oab*)

There are 64 possible pairs of syllogistic premises, depending on the moods of premises and the arrangement of the terms *a*, *b*, and *c* (i.e., the *figure* of a syllogism):

Figure 1	*Figure 2*	*Figure 3*	*Figure 4*
a – b	b – a	a – b	b – a
b – c	c – a	c – a	b – c

There are nine possible responses to (1), i.e., conclusions in four moods and two orders of end terms, *a* and *c*, and the response that no valid conclusion follows (hereafter, NVC). But, typically, reasoners do not consider all nine responses in their spontaneous conclusions; they generate just one or two. As a meta-analysis of six studies shows (Khemlani & Johnson-Laird, 2012), the most common response to (1) is that *Some of the architects are not chefs* – an error, since it is possible that all the artists are chefs, and so no definite conclusion follows validly.

In a typical study, responses to a syllogism vary from one individual to another. For instance, reasoners draw the erroneous conclusion to (1) about half of the time, but they also make a different error and conclude that *Some of the architects are chefs*. Only about a fifth of participants (university students) respond correctly that there is no valid conclusion, where a *valid* conclusion describes any conclusion that is true in all cases in which the premises are true (see Jeffrey, 1981). In logic, valid inferences can be drawn from any set of premises, including (1), from which it follows validly: *Possibly, all of the architects are chefs*. But, most experiments ask participant to draw definite conclusions, and it is rare for reasoners to infer spontaneous conclusions about possibilities (cf. Evans, Handley, Harper, & Johnson-Laird, 1999).

The psychologist's task is to explain the robust patterns of inference across this small, restricted set of 64 problems. The difficulty is that reasoners approach the problems with different abilities and appear to develop different strategies. Perhaps as a result, none of the twelve theories surveyed in the meta-analysis provides an adequate account of overall performance. The variability in reasoners' responses was enough to convince some theorists that the only way to understand how people reason syllogistically is to examine their individual differences (Stenning & Cox, 2006). To address the deficit, we developed a new theory – one that explains reasoners' most common responses as well as their individual idiosyncrasies for all 64 syllogisms.

In what follows, we review recent investigations into individual differences in syllogistic reasoning, and then describe a computational theory of syllogisms. Next, we report two analyses of the results from an early experiment, based on the computational theory. One analysis accounts for the variation in performance among three subsets of participants; and the other simulates the performance of the individual participants.

Individual differences in syllogistic reasoning

Several proposals describe individual differences in syllogistic reasoning. For example, some reasoners appear to be more proficient at the task than others (Galotti, Baron, & Sabini, 1986; see also Bucciarelli & Johnson-Laird, 1999). Bara et al. (1995) measured separately several factors, such as the ability to understand quantified assertions, but found that only one correlated with syllogistic performance, namely, working memory capacity. It accounts for a small amount of the variance in accuracy. Likewise, as Galotti et al. argued, good reasoners appear to consider more alternatives than do poor reasoners.

Ford (1995) argued that participants' descriptions and diagrams, and their justifications of their inferences, suggest that they adopt two different sorts of mental representation: diagrams and verbal representations (see also Bacon, Handley, & Newstead, 2003). She based her argument on participants' spontaneous use of a verbal strategy that substitutes the terms of the premises to yield a conclusion, and argued that some reasoners use the substitution strategy more than others. But, as Johnson-Laird and Bara (1984) suggest, a substitution-based strategy may be compatible with one representation.

One prominent attempt to model individual differences is due to Stenning and Cox (2006). They gave participants an ancillary "immediate inference" task (Newstead & Griggs, 1983) to ascertain whether they interpret assertions in systematically different ways. They proposed that some reasoners were more hesitant to draw valid conclusions, whereas others were more rash in drawing invalid conclusions. These patterns correlated with their subsequent syllogistic reasoning. The results were compelling enough for the authors to claim that analysis of aggregate performance is fundamentally "unjustified and misleading" (p. 1477). What the authors did not do, however, was to present an account of the syllogistic reasoning of the individual participants in their study.

Bucciarelli and Johnson-Laird (1999) also argued that aggregate analyses of reasoning often fail to capture meaningful differences in inferential behavior. In their studies, they discovered that participants seldom have a fixed interpretation for each syllogistic premise, and that they differ in the strategies they adopt. For example, they differ in which premise they interpret first and how they go about searching for alternative representations. But, these authors also did not present an analysis of the individual participants in their study.

In sum, no existing theory provided a compelling account of the computations that underlie the differences among individuals. Part of the difficulty may be a methodological problem: how does one analyze and assess the reliability of an explanation of the variation in performance of a complex skill such as syllogisms? With the exception of Polk and Newell (1995), psychologists had not hitherto made much progress towards an appropriate methodology. The first step is to frame a theory that is able to explain individual differences. In the next section, we describe such a theory.

A computational theory of syllogisms

mReasoner (Khemlani & Johnson-Laird, 2013) is a unified computational implementation of mental model theory, which posits that reasoning depends on the construction and manipulation of *mental models*, i.e., iconic simulations of possibilities (Bucciarelli & Johnson-Laird, 1999; Johnson-Laird, 2006; Johnson-Laird, Khemlani, & Goodwin, 2015). The theory and its implementation are based on three fundamental principles:

- Mental models represent *possibilities*: a given assertion refers to a set of discrete possibilities that are observed or imagined (Johnson-Laird, 2006).
- The principle of *iconicity:* A mental models is *iconic* as far as possible in that its structure is isomorphic to the structure of what it represents (see Peirce, 1931-1958, Vol. 4). But, models can also include abstract symbols, e.g., the symbol for negation (Khemlani, Orenes, & Johnson-Laird, 2012).
- The principle of *dual processes:* reasoning, including syllogisms, is based on two interacting sets of processes: intuitions yield an initial conclusion by building and scanning a single model; and deliberations search for counterexamples to intuitive conclusions and, where possible, formulate alternative ones (Khemlani & Johnson-Laird, 2013; Khemlani et al., 2015).

The computational model makes syllogistic inferences by first constructing a small set of tokens that denote the entities referred to the premises. For example, mReasoner can build the following initial model for (1):

```
architect   banker
architect   banker
architect   banker   ¬chef
                      chef
```

where '¬' denotes the mental symbol for negation. The system then scans the model for an intuitive conclusion. It scans in the direction in which the model was built. In the model above, for instance, the system builds tokens for *architects* first, *bankers* second, and *chefs* third. Hence, the program draws an initial conclusion in the figure *a-c*, i.e., it concludes that *some of the architects are not chefs*. This conclusion matches the preponderance of conclusions that reasoners spontaneously generate. For other sorts of syllogisms, the system draws initial intuitive conclusions in the *c-a* figure, again depending on how the model was constructed.

Because the intuitive conclusion depends on just a single model, the system generates it quickly. But, as the example illustrates, the conclusion may be invalid. To correct the error, the program can call on a deliberative component to search for counterexamples to conclusions (Johnson-Laird, 2006). It operates by modifying the initial model using a finite set of search strategies (Bucciarelli & Johnson-Laird, 1999; Khemlani & Johnson-Laird, 2013). When the deliberative system is engaged, it can find a counterexample to the conclusion that *some of the architects are not chefs*:

```
architect   banker    chef
architect   banker    chef
architect   banker    chef
            banker   ¬chef
```

This model falsifies the intuitive conclusion, and so the program responds that no valid conclusion holds.

mReasoner's machinery for inference operates stochastically (see Khemlani et al., 2015). It builds models and searches for counterexamples based on four separate parameters:

1. The **λ parameter** controls the *size* of a mental model, i.e., the maximum number of entities it represents. It does so by basing the size on a sample drawn from a Poisson distribution of parameter λ. Hence, λ can be set to an approximation of a positive real number.
2. The **ε parameter** governs the model's contents, which are drawn from the most common set of possibilities corresponding to a particular assertion (the *canonical* set), or else the *complete* set of possibilities consistent with the assertion. For example, in the case of *All a are b*, reasoners tend to consider only one canonical possibility: *a's* that are *b's*. But the complete set of possibilities allows for *b's* that are not *a's*. The ε parameter sets the probability of drawing from the complete set. It ranges from [0, 1].
3. The **σ parameter** describes mReasoner's propensity to engage its deliberative component, i.e., its counterexample search mechanisms. It ranges from [0, 1].
4. The **ω parameter** is a nested parameter; it describes what happens when mReasoner finds a counterexample to its intuitive conclusion. When ω = 0, the system reports that no valid conclusion follows. When ω = 1, it *weakens* its initial conclusion and searches for counterexamples of the weakened conclusion, if possible. For example, it can transforms *All a are c* to a weaker claim, namely *Some a are c*. When ω is between 0 and 1, it is the probability of weakening the conclusion.

At present, mReasoner provides the closest fit to the results from the data presented in the meta-analysis on syllogistic reasoning (Khemlani & Johnson-Laird, 2012) compared to alternative theories. Figure 1 depicts the data from the meta-analysis across the 64 syllogisms, as well as mReasoner's closest-fitting simulation of the data, and shows a strong correlation between the two (r = .82). But, is the system capable of accounting for individual differences in syllogistic reasoning? In the following sections, we extend its analysis to different subsets of individuals (Simulation 1) and then to the idiosyncrasies of individuals' inferences (Simulation 2).

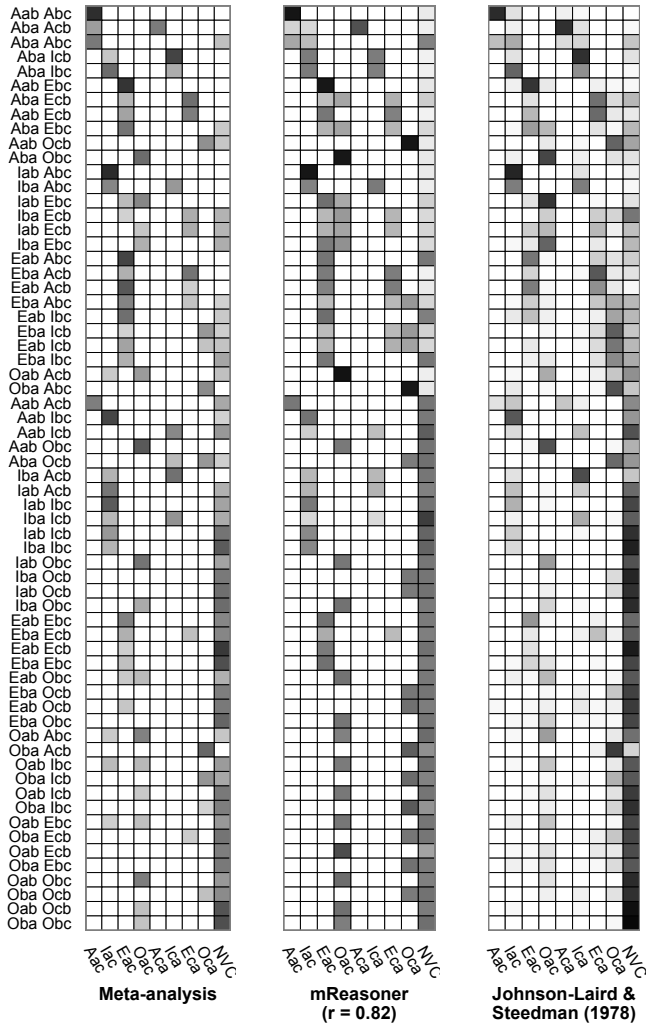

Figure 1. The percentages of responses to 64 syllogisms: a) in the meta-analysis in Khemlani and Johnson-Laird (2012); b) in mReasoner's best fit for the meta-analysis data (r = .82); and c) in the results of an experiment (Johnson-Laird & Steedman, 1978). *Aac* = All of the A are C, *Iac* = Some of the A are C, *Eac* = None of the A is a C, *Oac* = Some of the A are not C, and NVC = no valid conclusion. Each of the 64 pairs of premises occurs in a row, and each of the possible responses occurs in a column. The upper 27 rows denote syllogisms with a valid conclusion and the lower 37 denote NVC syllogisms. The grey scale in each cell indicates the proportion of corresponding conclusions (black = 100% and white = 16% or below). Hence, nearly 100% of participants in Johnson-Laird and Steedman (1978) responded that no valid conclusion (NVC) follows from the bottom-most syllogism, Oba Obc.

Simulation 1

We sought to analyze the data from an early experiment on syllogistic reasoning (Johnson-Laird & Steedman, 1978; see Figure 1) in order to model the participants' most frequent strategies. The participants were students at Columbia University (tested under the aegis of Janellen Huttenlocher), and they performed better than in any other study of all 64 problems. As Figure 1 shows, they made NVC responses more often than in the meta-analysis as a whole (45% vs. 30%, Wilcoxon test, z = 6.19, p < .001). Hence, the sample was biased towards higher performing, more deliberative reasoners.

We carried out an exploratory cluster analysis (Hartigan & Wong, 1979) to discover similarities in participants' patterns of reasoning. The data from the study were pooled over the 64 syllogisms and subjected to the *partitioning around medoids* (PAM) clustering algorithm, which is used to estimate the optimal number of distinct clusters in a given dataset (Kaufman & Rousseeuw, 1990). The analysis estimated that the optimal number of clusters in the data was 3. We used this estimate to constrain a hierarchical cluster analysis on the full range of participants' responses separated by the 64 syllogisms (see Hartigan,

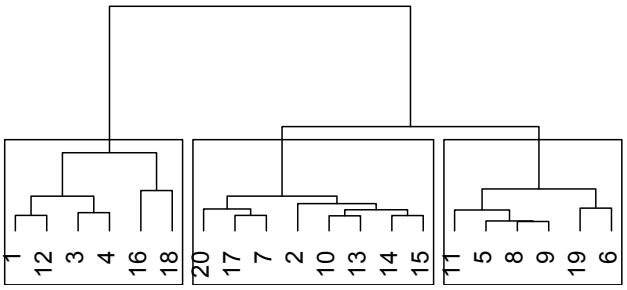

Figure 2. Dendrogram of a hierarchical cluster analysis performed on the 20 participants' propensity to yield 9 syllogistic reasoning responses pooling across 64 syllogisms from the data in Johnson-Laird and Steedman (1978). Each leaf in the tree reports a participant's unique identifying number. The analysis yielded three clusters of performance.

1975). Figure 2 shows how the cluster analysis grouped the 20 participants.

The three clusters suggest that there were systematic differences between subsets of individuals. To characterize them, we used mReasoner to simulate the three subsets by choosing appropriate parameter settings. If systematic differences exist, the simulations that best fit the data for each subset should differ, and the parameter settings of the best-fitting simulations should characterize the procedures on which the subsets relied.

Method and procedure

To simulate the three subsets of participants' performance derived from the cluster analysis, mReasoner generated simulated datasets for every possible combination of quantized settings of its four parameters. For each unique parameter setting, the system generated a dataset in which it carried out 64 syllogisms 100 times. The parameter settings were quantized to span their ranges as follows:

λ (size): *2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0*
ε (canonicality): *.0, .2, .4, .6, .8, 1.0*
σ (counterexample search): *.0, .2, .4, .6, .8, 1.0*
ω (weakening conclusions): *.0, .2, .4, .6, .8, 1.0*

Hence, the system generated 7 x 6 x 6 x 6 = 1512 separate simulated datasets. A grid search located the best fitting parameter setting for each of the three subsets of strategies.

Results and discussion

Figure 3 shows the data aggregated across the three subsets yielded by the cluster analysis, along with mReasoner's best-fitting simulations of those data, and Table 1 reports the parameter settings of the simulations. The parameter settings predict characteristics of the participants that comprise each subset.

As Figure 3 shows, reasoners' responses in Subset 1 appeared to vary more than in any other subset. The

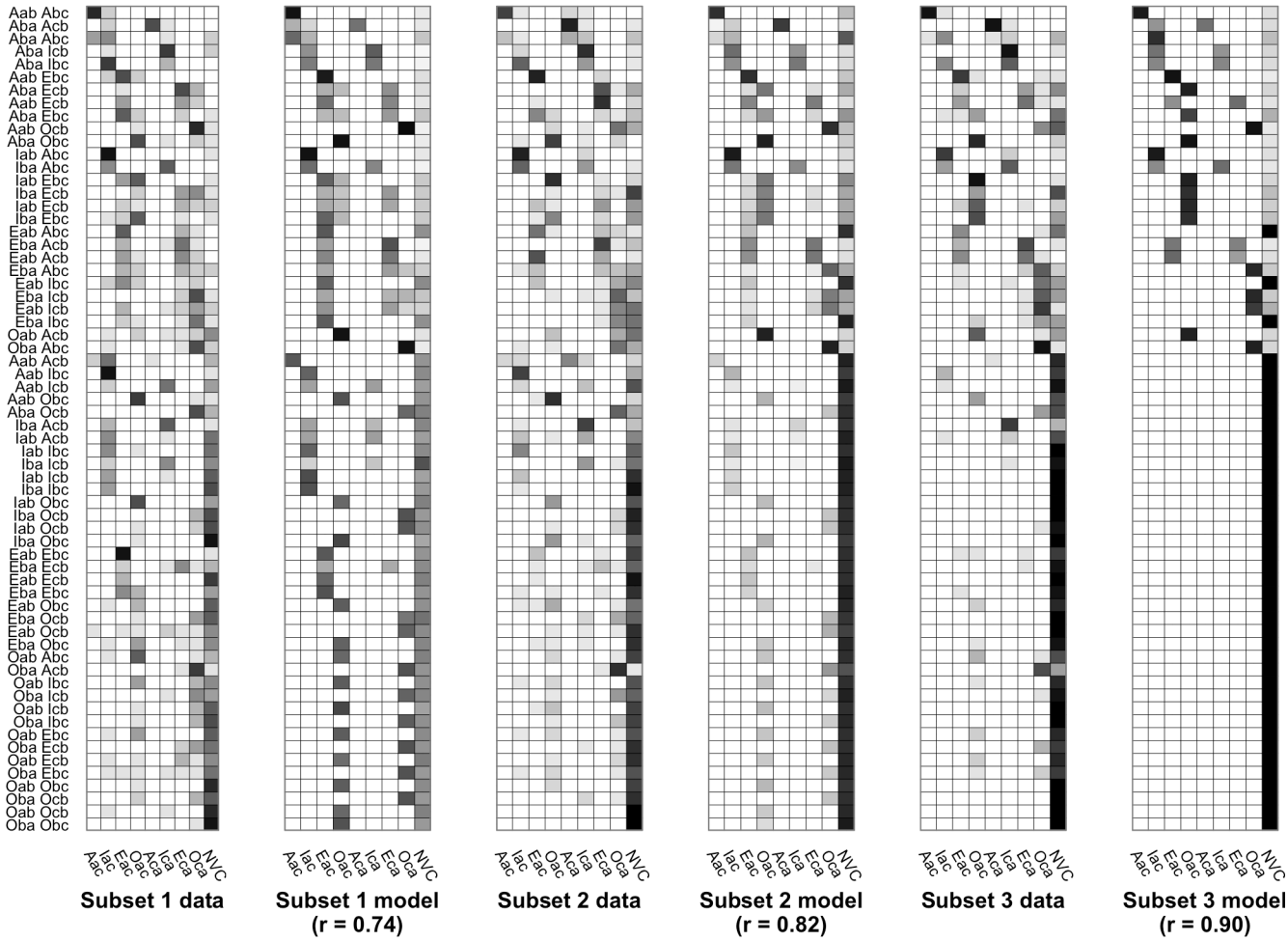

Figure 3. The percentages of responses to the syllogisms for three subsets of participants in Johnson-Laird and Steedman (1978), the corresponding mReasoner simulations, and their correlations with the data (see the text for an interpretation).

	Parameter settings and fit					
Subset	λ	ε	σ	ω	*r*	*Performance*
1	2.0	0.0	0.4	0.6	.74	Intuitive
2	3.0	0.6	0.8	0.6	.82	Intermediate
3	2.0	0.0	1.0	0.8	.90	Deliberative

Table 1. The parameter settings of mReasoner's best-fitting simulations for each of the three subsets of participants, the Pearson correlation of the simulated data with the actual data from the corresponding subset, and an assessment of each subset's performance based on an interpretation of the parameter settings.

parameter settings for Subset 1 explained this variability: reasoners appeared to build relatively small models, since the optimal parameter setting of λ, the size of the model, was 2.0: a model representing only 2 individuals. Since ε = 0, it meant that participants stuck to canonical possibilities, and were unable to explore the problem space fully. The relatively low value of the σ parameter (0.4) reinforces this interpretation; a value higher than 0.5 was likely to lead to a search for counterexamples that could have corrected errors. We accordingly characterize the participants in Subset 1 as *intuitive* reasoners. The term is not meant to impugn their intelligence, but rather reflects difficulty to consider alternative possibilities.

Subset 2 has an optimal setting for λ of 3.0, so the participants built larger initial models. They have a relatively high value of ε (0.6), and so they often considered non-canonical possibilities. Likewise, their propensity to search for counterexamples was high (σ = 0.8), which explains why they accurately inferred that the invalid problems had no valid conclusion more often. Hence, these individuals appear to be of *intermediate* ability.

Subset 3 has two parameter settings that were almost mirror images of those for Subset 1. They reflect a tendency to consider small, canonical models at the outset (λ = 2.0, ε = 0.0), but they always searched for counterexamples (σ = 1.0), and when they found one, they were likely to consider alternative weaker conclusions (ω = 0.8). We refer to the participants in this subset as *deliberative* reasoners.

The divergent sets of strategies yield systematic, behavioral predictions between the subsets of individuals, particularly between Subsets 2 and 3. For example, because the participants in Subset 2 consider more possibilities at the outset, they should provide correct responses faster than the members of Subset 2.

These simulations provide a computational explanation for why some reasoners are better than others. But, within these subsets, individuals are also likely to differ, and so our second simulation examined individual results.

Simulation 2

Simulation 2 used the same procedure and modeling technique as Simulation 1, but instead of comparing mReasoner's simulated datasets to reasoning performance at the subset level, we compared the predictions of the datasets to performance against the each of the 20 participants' data.

Method

For each of 1512 unique settings of the four parameters, the system generated a simulated dataset in which it carried out 64 syllogisms 100 times. An automated analysis discovered the parameter settings of the best-fitting simulations for each of the 20 participants.

Results and discussion

Table 2 provides the parameter settings and fit statistics of the best fitting simulations. The mean correlation between the best-fitting simulated datasets and participants' performance was .70, and ranged from .54 to .87. The mean of the worst-fitting simulated datasets and participants data was .25 (Wilcoxon test, $z = 3.9$, $p < .0001$) and the significant difference between the two sets of correlations suggests that the model's settings have drastic and qualitative effects on the patterns of inference it is capable of simulating. The correlations were lower than those for the analyses of subsets, and they reflect the reduction in power between subset and individual results. Reasoners' responses are much less systematic than when they are aggregated; nevertheless, all but four of the best-fitting simulations achieved a correlation of .60 or higher. Hence, while the variation in participants' responses was high, the computational model was capable of accounting for a significant proportion (about 50%) of their variance.

The optimal parameter values obtained from the analysis provide insight into participants' behaviors. For instance, if there exists significant concordance between the individual parameter values, then it would be because the participants are behaving similarly, i.e., no significant subsets or

	Parameter settings and fit statistics					
Participant	λ	ε	σ	ω	r_{BEST}	r_{WORST}
1	4.0	0.2	0.6	0.6	.69	.38
2	3.0	0.6	0.8	0.6	.69	.28
3	2.5	0.4	0.6	0.8	.59	.37
4	4.0	0.6	0.6	1.0	.57	.33
5	2.0	0.0	1.0	0.8	.86	.10
6	2.0	0.0	1.0	0.6	.81	.13
7	3.0	0.6	0.8	0.6	.63	.25
8	2.5	1.0	0.8	0.4	.82	.20
9	2.5	0.6	0.8	0.8	.80	.20
10	2.5	0.6	0.8	0.8	.72	.22
11	2.0	0.0	0.8	1.0	.80	.16
12	2.5	0.0	0.6	0.8	.65	.34
13	4.5	0.4	0.8	0.4	.74	.30
14	3.5	0.8	0.8	0.2	.75	.26
15	2.5	0.0	0.8	0.4	.65	.23
16	2.5	0.4	0.4	0.4	.58	.36
17	4.5	0.8	0.6	0.6	.65	.30
18	2.0	0.0	0.4	0.0	.54	.35
19	2.5	0.0	1.0	0.4	.87	.07
20	2.0	0.0	0.8	0.6	.62	.15

Table 2. The parameter settings of mReasoner's best-fitting simulations for each of the participants in the analysis, and the Pearson correlations of the best- and worst-fitting simulation with the data from the corresponding participant. The highlighted rows denote settings in which the correlation between mReasoner's best-fitting simulation and the data was < .60.

individual differences exist. In fact, the data show no significant concordance (Kendall's $W = .17$, $p = .89$) and is consistent with the cluster analysis reported earlier.

In sum, mReasoner successfully simulates the patterns of inference produced by individual reasoners.

General discussion

The mReasoner program is a computational implementation of the mental model theory of reasoning. It includes accounts of various sorts of deduction including syllogistic reasoning. It models in a stochastic way both intuitive and deliberative reasoning. In the past, investigators have lacked a systematic way in which to capture individual differences (though cf. Polk & Newell, 1995). In our view, the appropriate methodology consists in five main steps:

1. Obtain detailed data on the performance of a set of participants in carrying out the relevant task, such as their accuracy in drawing conclusion from the 64 different sorts of premise.
2. Develop a theory of their performance, and implement a computer program with stochastic parameters for the key factors that should determine performance, such as whether or not a search is made for alternative models.
3. Carry out a cluster analysis on the participants' data in order to determine whether there are systematic differences among different subsets of participants in their responses to the task.
4. Simulation 1: carry out an automated search through quantized settings of the parameters in order to determine whether it is possible for the settings to account for the differences amongst the subsets that step 3 revealed.
5. Simulation 2: carry out a similar search in order to account for the data from the individual participants.

The mReasoner system provided a close match to aggregated data from six syllogistic reasoning studies. We therefore carried out the five-step procedure described above in an analysis of the experimental results from one study of 20 participants. The cluster analysis yielded three main subsets of participants, and Simulation 1 showed that the theory could characterize them: intuitive reasoners who maintain small mental models and tend not to search for alternative models; intermediate reasoners who build larger, more varied models but do not search for counterexamples; and deliberative reasoners, who actively engage in a search for counterexamples, and weaken their conclusions in the light of them. The analysis of participants' individual reasoning patterns likewise showed a close match between the simulation and their data.

How might the theory be improved? It fails to capture certain systematic patterns in participants' responses; the theory makes the wrong prediction for syllogisms such as, e.g., Oba Acb. Hence, a more accurate theory needs to explain these and some other aberrant results. Of course, it remains a tentative achievement to fit data from just one domain of reasoning. The present analyses show promise that the computational model we describe is flexible enough to account for individual differences in syllogisms; a true test of its power is one in which it can model differences across a broad swath of reasoning domains.

Acknowledgements

This research was supported by a grant from the Office of Naval Research awarded to SK and by NSF Grant No. SES 0844851 to PJL. We are grateful to Ruth Byrne, Monica Bucciarelli, Geoff Goodwin, Tony Harrison, Laura Hiatt, Robert Mackiewicz, Marco Ragni, and Greg Trafton. Special thanks goes to Thad Polk for having kept the raw data from Johnson-Laird and Steedman (1978) and providing us with them.

References

Bacon, A., Handley, S., & Newstead, S. (2003). Individual differences in strategies for syllogistic reasoning. *Thinking & Reasoning, 9,* 133-168.

Bara, B., Bucciarelli, M., & Johnson-Laird, P.N. (1995). The development of syllogistic reasoning. *American Journal of Psychology*, *108*, 157-193.

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

Evans, J.St. B. T., Handley, S.J., Harper, C.N.J., & Johnson-Laird, P.N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1495-1513.

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition, 54,* 1–71.

Galotti, K. M., Baron, J., & Sabini, J. P. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General, 115,* 16–25.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: Wiley.

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

Johnson-Laird, P. N., & Bara, B. G. (1984b). Syllogistic inference. *Cognition, 16,* 1–61.

Johnson-Laird, P. N., Khemlani, S., & Goodwin, G.P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, 19.

Johnson-Laird, P. N., & Steedman, M. J. (1978). The psychology of syllogisms. *Cognitive Psychology, 10,* 64–99.

Kaufman, L. & Rousseeuw, P.J. (1990). *Finding groups in data.* Wiley & Sons: New York, NY.

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, 1-20.

Khemlani, S., Lotstein, M., Trafton, J.G., & Johnson-Laird, P. N. (2015). Immediate inferences from quantified assertions. *Quarterly Journal of Experimental Psychology*, 68, 2073–2096.

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology, 24,* 541–559.

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

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

Stenning, K., & Cox, R. (2006). Reconnecting interpretation to reasoning through individual differences. *Quarterly Journal of Experimental Psychology, 59,* 1454–1483.

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