Psychology Press
Taylor & Francis Group

Reconnecting interpretation to reasoning through individual differences

Keith Stenning
Edinburgh University, Edinburgh, Scotland, UK

Richard Cox
Sussex University, Falmer, UK

Computational theories of mind assume that participants interpret information and then reason from those interpretations. Research on interpretation in deductive reasoning has claimed to show that subjects' interpretation of single syllogistic premises in an "immediate inference" task is radically different from their interpretation of pairs of the same premises in syllogistic reasoning tasks (Newstead, 1989, 1995; Roberts, Newstead, & Griggs, 2001). Narrow appeal to particular Gricean implicatures in this work fails to bridge the gap. Grice's theory taken as a broad framework for credulous discourse processing in which participants construct speakers' "intended models" of discourses can reconcile these results, purchasing continuity of interpretation through variety of logical treatments. We present exploratory experimental data on immediate inference and subsequent syllogistic reasoning. Systematic patterns of interpretation driven by two factors (whether the subject's model of the discourse is credulous, and their degree of reliance on information packaging) are shown to transcend particular quantifier inferences and to drive systematic differences in subjects' subsequent syllogistic reasoning. We conclude that most participants do not understand deductive tasks as experimenters intend, and just as there is no single logical model of reasoning, so there is no reason to expect a single "fundamental human reasoning mechanism".

Computational theories of mind assume that subjects impose interpretations on information and then reason from those interpretations, possibly going through cycles of interpretation and reasoning. When subjects in the "immediate inference" task (Newstead, 1989) are presented with the information that *All A are B* and are asked about what follows, a substantial proportion of subjects respond that it follows that *All B are A*. Similarly, given *Some A are B* substantial numbers state that *Some A are not B* follows. And again, given *All A are B* substantial numbers state that *Some A are B* is false. Theories of discourse processing offer explanations about why these "implicatures" might be drawn (Grice, 1975). When the same subjects are introduced to syllogistic reasoning, they are presented with just one more premise from the same range of forms and are asked what now follows. One might expect subjects to adopt the same interpretations of the same English sentences in the second task as in the first, and one might expect them to apply the appropriate reasoning processes. Since the interpretations of the sentences would be the

Correspondence should be addressed to Keith Stenning, Human Communication Research Centre, Edinburgh University, 2, Buccleuch Place, Edinburgh EH8 9LW, Scotland, UK. Email: k.stenning@ed.ac.uk

DOI:10.1080/17470210500198759

same, one might expect the appropriate reasoning processes to be the same. Yet several studies have now shown that on assumptions prevalent in the psychology of reasoning, it is very hard to reconcile subjects' conclusions on the one-sentence (immediate inference) and two-sentence (syllogistic) tasks (Newstead, 1995; Roberts, Newstead, & Griggs, 2001). The conclusion of these studies is that the two-sentence task presents such calculative difficulties that subjects resort to quite different interpretations of the quantifiers.

We see this conclusion as pessimistic with regard to the relevance of these tasks to general reasoning and also as quite unnecessary on the evidence. Subjects can be seen as generally continuing the defeasible processes by which they normally interpret discourse, from the first task to the next. But seeing continuity between the one-sentence and two-sentence tasks within each subject requires alternative competence models for different subjects. Only by appreciating the individual difference between participants can we grasp the homogeneity of their interpretative procedures across the two tasks.

The proposal that subjects in these tasks should be modelled as doing defeasible reasoning has notably been made by Oaksford and Chater (see Oaksford & Chater, 2001, for a recent review) who have argued over the course of the last 15 years that the selection task, conditional reasoning, and syllogistic reasoning are understood by subjects as inductive tasks for which the appropriate "computational level" competence model is probability theory. Bonnefon (2004) relates mental models theory to a range of approaches using default logics to model human reasoning and provides useful references to earlier appeals to defeasible reasoning in the psychological literature. Although it is true that mental models theory makes periodic appeal to the defeasibility of reasoning, Stenning and van Lambalgen (2004, in press) show how failure to separate competence models for different construals of the tasks leads to little but confusion.

We share much with Oaksford and Chater (2001) and with Bonnefon (2004). We agree that the classical logical competence model is an insufficient basis for modelling most subjects in these tasks, and that their reasoning is often defeasible. We share their doubts that probability theory can be more than a theorists' tool as a computational level model and suspect that qualitative nonmonotonic logical models are both more insightful as competence models and nearer to the performance models that subjects actually use to reason with.

We also differ on a number of dimensions, first and foremost in believing that a range of competence models is needed for explaining participants' qualitatively different construals of reasoning tasks. Different logics are required for modelling different task construals. This richness of interpretation has to be acknowledged whether or not probabilistic information gain plays a role in subjects' reasoning.

In our discussion below of the source-founding model (Stenning & Yule, 1997) and its application to the present data, we return to other theories' stance toward individual differences. That discussion argues that construing mental models theory as a single "fundamental human reasoning mechanism", when it clearly has to be stretched to model all kinds of different reasoning tasks, has blunted curiosity about the qualitative differences between subjects' performances. It has also led to spurious arguments dismissing rule-based accounts of the suppression task. Stenning and van Lambalgen (in press) show that Byrne's (1989) claims that rules cannot be operating because they are suppressed in some circumstances turn out to be made with regard to logical rules inappropriate to the subjects' construal of the task. Logical modelling has the great virtue of requiring analysis of coherent construals of reasoning tasks, one model for each construal, a good starting point for cognitive analysis, which surely has to start from subjects' construal of the task.

Our interpretative approach was originally developed by showing that a single task (Wason's 1968 selection task) evokes multiple interpretational conflicts for participants. These conflicts arise because the overwhelmingly most likely initial interpretation of the rule in that task is nonclassical, nontruthfunctional, and nonmonotonic, and these properties conflict with the nature of the task—to test truth of a defeasible generalization by examination of cases (Stenning & van Lambalgen, 2004). Subjects' qualitatively different responses can be understood as attempts, variously successful, to resolve these conflicts. The effects of deontic variations of material and instructions in this task can be understood in terms of their different logical interpretations and different task demands. A formal default logic model of this understanding of conditionals is given by Stenning and van Lambalgen (in press) in the context of modelling Byrne's (1989) suppression task. Far from producing evidence against subjects' reasoning by logical rules, this task illustrates how subjects employ the same nonclassical default logic that underlies their problems in the selection task, to accommodate the conflicting statements presented in the suppression task. The particular logic proposed is also shown to be neurally implementable and extraordinarily efficient, properties that make it a good candidate for at least some of the functions of what have come to be known as "System 1" implicit reasoning processes (Evans, 2003; Stanovich, 1999). This logic also offers the possibility of a smooth articulation with System 2 processes modelled by classical logic.

These proposals present a modern logical view of the relation between task, materials, context, content, and logical form. On this view, to get an appropriate logical system in which to reason for some purpose, a wide range of logical parameters have to be set by clues from instructions, materials, context, and content in order to assign logical form, semantics, and the relevant concept of validity, which alone can fix what counts as "correct" reasoning.

Here, we assume the same broad defeasible logical framework in an empirical exploration of the range of defeasible understandings of the one- and two-sentence tasks (the immediate inference task and syllogisms). Tasks that appear to invoke disparate and unreconcilable behaviour can in fact be seen to invoke quite consistent interpretations, though different interpretations in different subjects. Appreciation of the variety of logics focuses attention on individual differences and makes modelling group data not generally justifiable. If subjects are aiming to do different tasks with different normative standards and different valid conclusions, then aggregate models of their "accuracy" of reasoning run the danger of being entirely misleading. Group models are only justified by unargued assumptions about the uniqueness of classical logical interpretation and the uniformity of the "fundamental human reasoning mechanism", or by generalized appeals to probability theory as a competence model.

The plan of the paper is as follows. The next section reviews the experimental work that has produced the evidence of disparities in interpretation between the immediate inference and syllogistic tasks, and it draws out from that work some of the assumptions that underlie its conclusions. The following section examines the relations between these experiments and logical and linguistic theories of discourse interpretation—Gricean approaches and default logics—arguing that although Grice (1975) has been widely appealed to in these papers, his theory has been torn from its context and only narrowly applied as a theory of (an aberrant classical) logical form rather than a theory of cooperative communication. The fourth section presents an experiment that collects and analyses immediate inference data in ways suggested by the logical framework, and it follows this by eliciting syllogistic reasoning from the same subjects. An exploratory statistical model of conclusion term order is then designed to reveal connections between patterns of interpretation and patterns of reasoning. A second experiment provides a replication and allows the analysis of conclusion term order to be reconnected to the analysis of reasoning accuracy. Finally, the General Discussion draws implications for empirical strategy in investigating interpretation and reasoning.

Experimental studies of interpretation's relation to reasoning

Newstead (1989) was among the first to study the relation between quantifier interpretation and reasoning in detail. This paper presents two tools for studying the interpretation of syllogistic quantifiers: the *immediate inference* task and the *diagrammatic* task. In the first, sentential, task, Newstead presented a premise to be considered true and a candidate conclusion sentence and asked subjects: "Is the conclusion sentence definitely true? Or else false?" (p. 85). The second tool for the measurement of interpretation was based on Euler diagrams. This task required subjects to consider the five possible relationships between two sets presented as circle diagrams and to indicate which of the four quantified statements were true with respect to them.

The study found, in both immediate inference and graphical tasks, some support for both conversion theory (e.g., Chapman & Chapman, 1959), in which subjects treat logically asymmetrical statements as reversible (e.g., *All A are B* as implying *All B are A*), and for Grice's (1975) theory of implicatures in which, as illustrated in the examples cited in the first paragraph, participants make extra inferences from the assumed cooperativeness and omniscience of the speaker. These theories are discussed in the next section. Politzer (1990) improved on methods and had more success at modelling immediate inference along Gricean lines but did not collect syllogistic data.

Newstead (1995) repeated the immediate-inference experiment of his earlier paper and ran the same participants on syllogisms. This paper focused on the question of whether Gricean interpretation as evidenced, for example, by the drawing of the implicature from *Some A are B* to *Some A are not B* is responsible for errors of syllogistic reasoning. An analysis of the then existing literature showed that the expected errors in syllogistic reasoning are quite rare. Four new experiments were presented. In Experiment 2, three "measures of Gricean interpretation" were computed: (a) the graphical criterion of choosing the set intersection or disjoint diagrams for *some or some . . . not* statements; (b) the frequency with which *some* was taken to imply *some . . . not*; and (c) the frequency with which *all* implied either the falsity or logical independence of *some*, and the frequency with which *no* implied either the falsity or logical independence of *some . . . not*.

These three alternative measures of Gricean interpretation did not even correlate with each other. Newstead (1995) concluded: "Surprisingly, since these are supposedly different measures of the same thing, the correlations were small, non-significant, and in one case actually negative. This suggests that the three measures may not be tapping the same thing" (p. 652). There was also little correlation between any of these three measures of Gricean interpretation and conclusions drawn in syllogistic reasoning, and the conclusion was drawn that subjects change their interpretation of the quantified statements when going from the one-sentence to the two-sentence tasks.

Roberts, Newstead, and Griggs (2001), perhaps because of the earlier difficulty in producing stable empirical indices of Gricean interpretation, adopted a top-down approach to the problem. The paper works through a total of 13 combinations of "logical", heuristic, Gricean, and "conversion" interpretations of the quantifiers and produces complete tables of which conclusions, if any, follow under which interpretations, for all 64 syllogism premise pairs. These models of interpretation are then fitted to some existing datasets (Dickstein, 1978; Johnson-Laird & Steedman, 1978, combined with Johnson-Laird & Bara, 1984) and one new dataset of the kinds of conclusion that subjects draw from each pair of syllogism premises.

When these thirteen interpretations are fitted to the three data sets, simple Gricean interpretations fit poorly to all three. Heuristics such as atmosphere and matching do poorly on the earlier datasets but rather well on the new one. The best fits are generally of conversion and "reversible" interpretations (the latter being a particular subcategory of conversion) either with or without elements of Gricean interpretation. These different models are rather poorly distinguished by the data. It is a bane of the study of the syllogism that remarkably simple heuristics get a high proportion of classically correct answers, and all models of representation and processing share a large core of predictions. Only by lumping all three datasets together is any significant separation of interpretative models possible.

The authors repeat Newstead's (1995) earlier conclusion that the extra complexity of syllogistic reasoning leads to the abandonment of the simple Gricean interpretations elicited by the immediate inference tasks, in favour of different interpretations for the syllogism task. They see this as a the result of an attempt by subjects to simplify the necessary calculations.

Seeing deductive tasks as discourse interpretation

This psychological literature reduces both Gricean interpretation and issues of the conversion or reversibility of statements' interpretations to exercises in modulating classical logical forms for sentences. But Grice's theory is not a theory for which there is some "classical logic of conversation", which can be translated from standard logic by the addition of conjuncts to classical logical forms. Grice's theory is about defeasible inferences drawn during the cooperative activity of producing/understanding a certain kind of discourse in which the hearer attempts to identify the speaker's intended model of their discourse, using general knowledge and assumptions about cooperativity and omniscience, as well as general maxims that cover this cooperative activity. We call this general discourse processing goal *credulous* reasoning— the goal is to construct (and believe) a model of what one is told. This technical term is not perjorative. It contrasts with the *sceptical* attitude to discourse adopted in classical logical proof where conclusions must be true in all models (not merely an intended one).

Roberts et al.'s (2001, p. 174) own calculations of what conclusions will follow from what interpretations illustrate the contrast between their narrow and our broad interpretations of Grice. Roberts et al. consider the syllogism *All B are A/Some C are B* and observes that with Gricean interpretations it may get encoded either with Set A identical to Set C, or with the two sets intersecting with nonintersecting subsets. They then argue that "with the outcome sets made explicit, a problem for anyone adopting these interpretations becomes apparent. Gricean interpretations affect not only the *encoding* of a problem, but also its *decoding* of the final outcomes so that conclusions can be generated. ... The assumption of the mutual exclusivity of *some* and *all* during encoding will result in a contradiction on decoding."

They then entertain several possible complex resolutions to this dilemma. But this argument supposes that the hearer treats the output models like models of classical logical conclusions true in all models of the premises, and that now, from the point of view of a speaker concluding from this construction, *Some A are B* is inconsistent with *All A are B*—not merely that it would be misleadingly uninformative to say the former when the latter is true in the intended model. In Grice's theory, the aim of processing determines the semantics of the representations, and in giving a competence model, one should not forget this and return to a classical interpretation in midstream. The credulous process that Grice described is defeasible and therefore nonmonotonic. Inferences that might be made at the end of Premise 1 may not be made after Premise 2. The construal of the task and the interpretation mechanism are perfectly homogeneous across tasks, though they have the effect that implicatures arising at one point may get cancelled at another.

Likewise, conversion and reversibility are given no theoretical basis. They are simply offered as observations from earlier experiments with no justification beyond the resulting simplification of reasoning. Yet there are well-developed logical frameworks for understanding issues of reversibility in the process of discourse comprehension. Closed-world reasoning is the general label for this reasoning, which can be treated technically within default logics. Far from being some arbitrary syntactic operation, or simply an attempt to find an easier problem, this reasoning is another example of Gricean credulous interpretation. It is not hard to see informally that default logical models for discourse can lead to conversion. The rough principle is: "Only add to the model what is necessary to understand what has been said." So, if a discourse begins "All A are B", then at this stage we get a model in which the only way that something can get to be B is because it is A, and in such a model it will be true that "All B are A". This is transparently a case of Gricean interpretation via the maxims of quantity and relevance—"say enough and not too much". It is true that the closed-world reasoning much modelled by default logics in AI was not discussed by Grice (1975). But Roberts et al. (2001) explicitly deny that there is any connection between Gricean interpretation and conversion/reversible interpretation (p. 184). This denial is perhaps a reflection of a general disinclination to treat deductive reasoning materials as discourses.

This close association of Gricean implicature and reversible interpretation points up the fact that credulous discourse processing encompasses a wide range of interpretational inferences. Another class relevant to syllogistic reasoning is anaphora resolution. Consider a subject disposed to construct a single intended model for the discourse: *Some A are B/Some B are C*. The speaker is likely to construct a model of the first sentence in which some As are Bs. When the second sentence arrives, what anaphoric relations is such a speaker likely to impose? The example immediately makes it clear that such syllogisms are anaphorically problematic, because the only shared term is used first as a predicate and then in a noun phrase. The second sentence's subject has the indefinite quantifier "some", and two indefinite phrases in a discourse are prone to be interpreted as introducing nonidentical referents. This line of reasoning might yield the response that "it's different Bs that are C than the ones that are A", indicating that the subject has constructed a model with some As that are B and some other Bs that are C, and so nothing may follow about the relation between A and C. On the other hand, the general pragmatic forces toward integration of the only information available pulls in the other direction toward the response "since there is a shared reference to B, and the speaker intends me to find some connection, then the Bs that are C must be intended to be the same Bs as the ones that are A". Such reasoning leads to an intended model in which some As are Cs. So even within a generally credulous "construct-the-intended-model" construal of the task, subjects may draw a variety of conclusions. Indeed, closed-world reasoning also requires modelling in logics with several subtly different degrees and kinds of closure of the world. One might object that it is unnatural to treat these premises as having anaphoric relations. The subjects would undoubtedly agree. However, if they have a credulous interpretation of the task, resolving these relations is a goal that is forced upon them, and a little Socratic tutoring does elicit exactly these kinds of commentaries.

The fact that Gricean "forward" implicatures such as that from *Some A are B* to *Some A are not B* are the result of the same credulous goals as are reversible closed-world inferences from *All A are B* to *All B are A* suggests an explanation of why the former show up strongly in the one-sentence immediate inference task, but are not well supported in the data from the two-sentence syllogistic task, whereas the latter are common in both tasks. For example, when *Some A are B* is followed by a second premise, say *All B are C*, the former's implicature *Some A are not B* is not relevant to the available candidates for syllogistic conclusions connecting A and C. However, when *All B are A* as a first premise is followed by *No C are B*, the former's reversible interpretation in which all B are A is relevant to possible conclusions about the relation between A and C. Reversibility, viewed as credulous closed-world reasoning, just yields more relevant connections between end terms in syllogisms than do simple forward implicatures.

The underdetermination even of credulous interpretations clearly presents a degrees-of-freedom problem for data analysis. Some have claimed that once interpretation is taken seriously, empirical constraint is impossible. But fortunately this shift toward seeing the task as discourse interpretation strongly suggests a change of experimental programme, which can provide considerably more empirical constraint. At a methodological level, Roberts et al. (2001, p. 177) make the assumption that there is a sufficiently dominant interpretation and reasoning process, shared among a large enough subgroup of subjects, persisting throughout their performance on 64 syllogisms, that fitting models to group data makes sense. Their eventual conclusion is that they have gone about as far as it is possible to go without resort to individual difference methodology, but even this concession appears to assume that any individual differences will be details to be hung on the common model. It might be convenient, but why should this be so?

Once the subjects are seen as attempting different tasks, the a priori attractiveness of fitting group models to data is diminished. For example, the authors note that whereas the best fitting group model for the older data sets is either reversible or Gricean reversible, the best fitting group model to their own new data is heuristic (matching). They suggest this is probably because of lower skill levels in the new subjects. So different populations of subjects emerge fitted to uniform but quite different group models of the task as one goes from one population of undergraduate student subjects to another. Is it not much more plausible that the groups contain different proportions of heterogeneous subjects, and that at least some of these heterogeneous kinds of subject are represented in all groups? At worst, it is perfectly possible for group data to produce models that do not fit any individual at all.

When it comes to syllogistic reasoning data, there is an issue about how to conduct exploratory analysis. We here investigate both accuracy of reasoning (on a classical logical criterion) and conclusion term ordering as measures of reasoning performance that might be expected to be sensitive to the interpretative differences just described. Global accuracy analysis allows comparison with earlier literature, but has several problems. As mentioned above, there is the conceptual problem that if participants have different understandings of the task, then accuracy should be assessed on appropriate competence models for their understanding. It is true that one of the problems with studying syllogisms is that many of the possible credulous models of syllogism understanding diverge from the classical sceptical model on only a few problems. There are good logical reasons. The syllogism is an unusual fragment of logic, which allows application of only slightly strategically modified processes of credulous discourse processing to achieve sceptical reasoning. This is because, as we shall see presently, it permits generation of classically valid conclusions through the identification of "critical individuals"—individuals fully specified for all three properties (see Stenning & Yule, 1997). So the conceptual problem may not be as severe here as in, say, the selection task. However, the fact that reasoning can be done in terms of the identification of critical individuals turns out to mean that conclusion term order provides an alternative measure of reasoning, which is more closely related to process and to accuracy than at first appears and has important methodological advantages for exploration. Exploratory statistical models of accuracy require the modelling of nine possible responses—eight conclusion types plus "no valid concluson". Conclusion term order is binary. In the end what we want is not models of global accuracy but models of reasoning process.

Conclusion term order is an obvious measure of syllogistic reasoning to relate to the reversibility of interpretation (or its refusal). When subjects decide to make an *ac* or a *ca* conclusion, they are applying information packaging to their own productions. We should expect that their structuring of their conclusions can tell us a great deal about the relation between their interpretations and their reasoning processes. Term-ordering in conclusions has been studied extensively since Johnson-Laird and Steedman (1978) popularized the conclusion construction (as opposed to evaluation) task, thus making this data available. Stenning and Yule's (1997) source-founding model is the widest coverage and most accurate model of conclusion term ordering available, and here we build on that model to analyse individual differences in our subjects' conclusion term ordering. The source-founding model conceives of the classical logical task as finding types of individual that must exist in all models of the premises. Equally it conceives of the credulous task as finding types of individual that constitute the speaker's intended model. The source-founding model proposes that subjects identify a source premise (one that entails the existence of the identified type of individual) and draw a conclusion by adding the end term of the other premise on to the end of this premise. So conclusion term order is decided by choice of source premise, and choice of source premise, along with some minor quantifier adjustments in a few cases, determines "accuracy".

Stenning and Yule (1997) used a version of the syllogistic task that provides data about the ordering of the middle term during reasoning. Instead of being asked to draw a conclusion that drops the middle term, subjects were asked to do a logically closely related task of describing, in terms of all three properties, a type of individual that must exist if the premises are true. So, for example, given the premises *All A are B* and *Some C are not B* the subject might conclude, for example: *a C that is not B and is not A*. Note that any ordering of the terms in such descriptions is equally logically valid. Subjects were instructed that if no critical type of individual was entailed, then to respond "no valid conclusion" (VC).

The source-founding model is shown to be considerably more accurate than mental models theory's predictions, and it explains the new data on three-term ordering generated by the novel task (cf. Stenning & Yule, 1997, Table 11). Intuitively, the source premise identifies the type of individual that the problem is about. In our example syllogism *All A are B. Some C are not B* the second premise is source as it entails the existence of the Cs that are not B on which the conclusion *Some C are not A* is founded. Stenning and Yule noticed that subjects overwhelmingly draw their conclusions by adding the end term of the other premise to the tail of the source premise (removing the middle-term). Of the possible three-term orders in their task, the model identifies one third that should be more common than the other two thirds of orders. In fact, between 70% and 90% of all responses fall in this predicted third of orders, for each of the four figures, in comparison with 60%, 44%, 72%, and 8.7% for mental models predictions. The improvement in fit is mostly due to the source-founding model's prediction of *bac* and *bca* term orders, which are observed to be the commonest of all. Mental models theory rules these orders out since it is organized around the principle that the middle term must be got into medial position to allow a "mental cancellation" operation. We therefore propose to incorporate our individual differences by extending the source-founding model.

Subjects tend to draw conclusions founded on the source premise, as identified by the classical competence model. However, the crux of theory as a processing model is obviously the heuristics that subjects use to identify the source premise. This point has been misunderstood. For example, Chater and Oaksford (1999, p. 237) assume that the theory is that subjects identify the source premise using the classical logical competence model, and so they conclude that the theory cannot be applied to invalid syllogisms, to conclusions drawn by implicature, or to quantifiers such as *few, most, many*. However, this is just to misunderstand the theory. Although the experimenter's initial evidence for source founding had to be based on identification of source premises estimated from a competence model, when it comes to constructing process models of participants' reasoning, a variety of heuristics suggest themselves on which a processing theory might be built. Some were proposed in the original paper and are refined here, and of course such heuristics can cover invalid syllogisms and inferences by implicature. It is quite natural to extend the central discourse processing idea that subjects attempt to anchor their reasoning on an established individual (or set) to quantifiers such as many and most. Stenning and Yule (1997) show that the heuristics *select unique existential premises as source and select unique positive premises as source*, applied in that order, approximate to classical logical competence and account for a large proportion of subjects' reasoning. The same heuristics approximate to credulous individual identification. So the source-founding model is a "shell" process model, which abstracts over different logics and representations and in which different strategies can be expressed by changing the heuristics for source premise identification. At this level of analysis, there are many parallels between the source-founding model and Chater and Oaksford's (1999) probability heuristics model (especially the attachment heuristic), however different is their general philosophy.

In summary, the credulous construction of single intended models for discourse is a family of interpretative processes, which have quite different goals from those of classical logical proof. Given the range of credulous discourse processing available, we here adopt a more appropriately descriptive approach to finding important synoptic patterns of interpretation across immediate inference task questions. These patterns are then used to predict subjects' syllogistic reasoning, starting from an exploratory statistical model of conclusion term order and using its findings to make and test predictions about subgroups of subjects' reasoning processes viewed through the source-founding model. Developing specific logical models is a task for later papers.

EXPERIMENT

The history of our experimentation is that an initial experiment yielded a much more complex regression model of the relations between individual differences in interpretation and subsequent syllogistic reasoning than had been anticipated. This meant that it was not possible to reserve a test subset of those data on which to test the model, so a second experiment of exactly the same form as the first was conducted on a new group of subjects drawn from the same student population. In what follows we refer to the first experiment's data as the "development dataset" and the second as the "test dataset". We report the results in parallel for the reader's convenience.

The experiment includes both an immediate inference task and the subsequent syllogistic task performed by the same subjects. We analyse and discuss the former before introducing the latter because the exploratory analysis of interpretations in immediate inference feeds into the modelling of the subsequent reasoning data.

IMMEDIATE INFERENCE

Stenning and Cox (1995) studied both sentential immediate inference and diagrammatic tasks and showed that with an appreciation of their different semantics, their measures could be brought into some correspondence. In this paper we focus on sentential measures of interpretation. We first collected a set of sentential immediate inference data on the syllogistic quantifiers, analogous to Newstead (1995, Exp. 2). We analyse these data initially from a descriptive standpoint. What are subjects' naive logical intuitions as gathered in the immediate inference task like? Are general patterns to be found?

Method

First-year undergraduate students who had not been exposed to formal logic teaching were given questionnaires about their interpretation of the quantifiers *all*, *no*, *some*, *some . . . not*.

Participants

Participants yielding the development dataset were 101 undergraduate psychology students at the University of Edinburgh. They were tested during a lecture on cognitive psychology. These students were drawn from a wide range of departments across the entire university with a predominance of social science faculty students. Few of these students had received any formal logical training at secondary school. None of the students had taken logic courses in the university. The second set of participants ($N = 62$) who yielded the test dataset were from the same introductory psychology class in the following year at the same point in their courses. As for the development dataset, the Participants' quantifier interpretations were tested during a lecture on cognitive psychology.

Materials and procedure

Immediate inference (II) questionnaire. The task used was similar to that described by Newstead (1995, Exp. 2), as described above. As in Newstead's study, the questionnaire consisted of four pages. At the top of each page one of the four standard quantified statements was displayed: All As are Bs; No As are Bs; Some As are Bs; and Some As are not Bs. These were the premise statements. Beneath these premise statements the four quantified statements were listed (All As are Bs, etc.) and the converses of these (All Bs are As, etc). These were the candidate conclusion statements. Alongside the eight response statements were response options "T" (true), "F" (false), and "CT" (can't tell; possibly true and possibly false). The order of the four stimulus statement pages was randomized across subjects.

Participants were instructed:

> This is a study of the way people draw conclusions from information. On each of the following pages there is a statement at the top of the page. An example is "All As are Bs". Assume that these statements at the top of the page are true and that there are both As and Bs.
>
> Below each statement is a line. Below the line are some more statements. For each of the statements below the line, decide whether you believe it is true, false, or one can't tell (because either is possible), given the truth of the sentence at the top of the page. Indicate your belief by circling ONE of either "T" (true), "F" (false), or "Can't tell".
>
> Examples:
>
> - If you believe that "No As are Bs" must be true given the true statement "All As are Bs" then circle T.
> - If you believe that "Some As are not Bs" must be false given the true statement "No As are Bs" then circle F.
> - If you believe that "No As are Bs" could be true or could be false given the true statement "No As are Bs", then circle CT.
>
> Again, please note that you should interpret "some" to mean "at least one and possibly all".

Participants were allowed as much time as they needed to complete the tasks (approximately 10 minutes).

Results

Table A1 (Appendix A) shows the proportion of "true", "false", and "can't tell" responses to each quantifier, along with the responses correct according to the logical model with the no-emptysets axiom. In Table A1, Newstead's (1989) results are shown if the results of the present study differ by more than .07 from those reported by Newstead (1989, Table 2, p. 86).

In Table A1, primed conclusion quantifiers (e.g., A′) represent the converse conditions (e.g., *ALL Bs are As*). The introduction of the "Can't tell" response option in the current study resulted in a marked lowering of conversion and Gricean errors of interpretation compared to the results of Newstead (1989). Of course, these absolute differences are not of great interest in themselves. Their real significance can only be appreciated through the analyses of subjects' concept of validity, which they enable (presented below).

The data shown in Table A1 (see Appendix A) and Table 1 are very similar to those of a previous study with a different sample (Stenning & Cox, 1995).

Table 1. *Development dataset: Numbers of subjects making each number of the two kinds of errors possible on* QAB:QBA*? questions*

	T/F for CT									
CT for T/F	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Tot*
0	2	5	5	5	8	5	2	5	8	45
1	2	0	1	1	3	3	0	2	0	12
2	0	1	2	1	0	1	0	0	1	6
3	0	2	1	1	1	0	2	0	0	7
4	0	0	0	0	0	1	0	0	1	2
5	0	0	2	2	2	0	0	0	0	6
6	0	2	3	1	3	0	0	0	0	9
7	2	1	1	0	0	0	0	0	0	4
8	7	3	0	0	0	0	0	0	0	10
Totals	13	14	15	11	17	10	4	7	10	101

Note: Can't tell (CT) for True or False (T/F) errors are judgements that the conclusion is logically independent when it is in fact dependent; T/F for CT errors are judgements that logically dependent conclusions are logically independent.

Participant profiles

Although grouped data provide a comparison to earlier work, and it is possible to examine piecemeal correlations between answers to different questions, our real interest is in finding patterns of interpretation characterizing a participant's interpretative scheme. There is a large space of possible patterns of responses to the immediate inference questions (3^{32}) and therefore a considerable discovery problem in finding useful descriptions. We approached this problem by exploring students' use of the CT response and their response to subject predicate inversion. Credulous interpretation should generally reduce CT responses, and closed-world varieties will generally increase acceptance of inverted conclusions.

Initial investigation revealed (Table 1) the expected substantial group of students who rarely respond CT and who overaccept inverted conclusions according to the classical logical measure. Substantial numbers of this group never responded CT to any question with transposed subject and predicate. Not only would these subjects respond, for example, T when given *Some A are B* and asked whether *Some B are A*, but they would also respond the same way when given *All A are B* and asked whether *All B are A*.

An early contrasting observation was that a substantial group of subjects would overuse the CT response (by the classical standard). That is, they responded CT whenever they were asked a question in which subject and predicate were transposed. Not only would these students respond CT when given, for example, *All A are B* and asked whether *All B are A*, but they would respond the same way when given *Some A are B* and asked whether *Some B are A*. This last pattern is particularly distinctive since it does not arise from either any obvious Gricean interpretation or reversible interpretation. Whereas "illicit conversion" has been discussed since Aristotle, this pattern of responding has remained unremarked in the literature. Table 1 shows the distributions of numbers of these two general patterns of response.

Here are two quite different dimensions of divergence of interpretation from the classical logical model, which need to be investigated in parallel. For convenience, we label a tendency to respond CT where T or F is correct as *hesitancy* and a tendency to respond either T or F where CT is correct as *rashness*. Note that these are empirical terms close to the data, which will turn out to have complex relations to theoretical concepts such as credulous and sceptical reasoning. Rashness and hesitancy can potentially be exhibited both when the conclusion sentence preserves subject/predicate (henceforth *in place*) and when it changes (henceforth *out of place*), potentially yielding four dimensions of classification. As a matter of observation, no participants had strong tendencies to be hesitant on in-place questions, thus reducing the space to three dimensions. These results therefore suggest a scheme for insightful abstraction over the sentential response data.

Setting thresholds on the number of CT responses required to qualify as hesitant and on the number of T or F responses required to qualify as rash can be done both within Q_AB questions and Q_BA questions. This transforms the space to three binary dimensions. Hierarchical log-linear modelling (e.g., Stevens, 2001) revealed that 3 second-order terms (rashness on in-place questions by rashness out of place; rashness out of place by hesitancy out of place; rashness in place by hesitancy out of place) made statistically significant contributions to a model of the data. The technique also permitted the cut-off points on each dimension used to categorize participants to be iteratively adjusted until residuals were minimized. The selected cut-off points were 0, ≥ 1 responses for rashness on in-place items; <5, ≥ 5 for rashness on out-of-place items; and <6, ≥ 6 for hesitancy on out-of-place items. The figures on the first line of Table 2 show the number of subjects assigned to the three binary dimensions. Of the participants, 23 are hesitant on out-of-place items, 70 are rash on in-place items, and 31 are rash on out-of-place items.

A total of 92% of participants fall into four groups. A total of 23 participants are neither rash nor hesitant on either in-place or out-of-place questions, in general complying with the

Table 2. *Numbers of subjects classified by hesitancy on out-of-place questions, rashness on in-place questions and rashness on out-of-place questions*

	Not hesitant on OP				*Hesitant on OP*			
	Not rash on IP		*Rash on IP*		*Not rash on IP*		*Rash on IP*	
	Not rash on OP	*Rash on OP*	*Not rash on OP*	*Rash on OP*	*Not rash on OP*	*Rash on OP*	*Not rash on OP*	*Rash on OP*
Developmental dataset	23	2	24	29	6	0	17	0
Test dataset	12	3	4	17	13	2	5	6
Syll.	8	0	7	15	4	0	6	0

Note: OP = out-of-place questions. IP = in-place questions. Syll. = developmental subsample of subjects that completed the syllogistic reasoning task.

classical competence model. A total of 24 participants are just rash on in-place questions. However, both of the other substantial groups make two kinds of error. The largest group (29) are rash on both kinds of question. The fourth group (17) consists of participants who are both rash on in-place questions and hesitant on out-of-place questions.

Several generalizations can be made about associations between these three dimensions. Participants who are hesitant on out-of-place items are not rash on out-of-place items, and vice versa. (Note that it is logically possible to be both rash and hesitant on out-of-place items since the two sets of eight items contributing to each dimension are disjoint.) Participants who are rash on out-of-place items are rash on in-place items but not vice versa. Participants who are hesitant on out-of-place items are just as likely if not more so to be rash on in-place items as are participants who are not hesitant.

Test dataset: Interpretation

The assignment of participants to rashness/hesitancy categories in the test dataset was made in the same way as for the previous data and is displayed in Table 2.

These data show a generally similar distribution of patterns of interpretation to that in the development dataset.

Discussion

Typically, the data used in the literature to examine interpretation and to explain reasoning patterns have consisted of responses to particular inferences. The field has concentrated on specific errors—especially on errors of commission (e.g., illicit conversion, or Gricean implicatures), but has not noticed errors of omission (e.g., failing to conclude from *Some A are B* that *Some B are A*). Our analysis of the immediate inference data shows that there are strong response tendencies, which generalize across particular logical inferences and are most strongly driven by subject/predicate relations between premise and conclusion, and by whether conclusions are logically dependent or independent. What is more, these two factors interact strongly but in complex ways in determining responses. To take one example, one half of which has greatly occupied the literature, participants who commit the fallacy of illicit conversion of *all* (discussed above), tend not to omit to validly convert *some*. Participants who fail to validly convert *some* do not tend to commit the fallacy of converting *all*. With conclusions that preserve subject/predicate structures, omission errors are hardly ever made, while many commission errors are made by participants who never make them in assessing subject/predicate inverted conclusions.

Many of these findings can be understood in the context of our discussion of the discourse interpretations available. Rashness on in-place questions corresponds to the commission of the paradigm Gricean implicatures: for example, concluding *Some A are not B* from *Some A are B*, or vice versa, and similarly with claiming that *All A are B* is false given *Some A are B* or that *No A are B* is false when given *Some A are not B*. These are transfers of a credulous model of the task. A large proportion of the participants exhibit this tendency (78/101). This is consonant with earlier findings.

What rashness on out-of-place questions signifies is an intriguing question−another part of the pattern so far ignored. As the discussion above indicated, at least some of these errors could be generated by credulous closed-world reasoning. There is no reason in Grice's theory why change in subject/predicate assignment in the conclusion should affect the drawing of implicatures. Or, to turn the question around—why should participants who are rash in place but not out of place refuse credulous conclusions when subject/predicate structures are altered?

One possibility for explaining why participants do draw both in-place and out of place implicatures is a representational explanation. Participants who are rash regardless of subject/predicate structure may be pursuing a representational strategy that removes subject/predicate information (for example, a graphical approach such as Euler's circles), combined with applying a credulous model of the task. Participants who are rash on in-place but not on out-of-place questions might then be interpreted as having a credulous model, but being guided by the way the speaker has designed the discourse structure, they do not draw out-of-place implicatures. The two rashnesses (in-place and out-of-place) are significantly, if weakly, positively correlated (.40, $p < .001$). There are almost no participants (2 of 101) who are rash on out-of-place questions but not on in-place questions, and this is as this explanation might predict. The subject/predicate structure of out-of-place questions might block implicatures in a credulous model, but it is hard to see why someone applying a credulous model to out-of-place items would not apply it to the in-place items.

What are we to make of hesitancy, the striking novel empirical observation here? Hesitant participants draw too few inferences on out-of-place problems (by a classical logic benchmark), not too many. On the face of it they seem unlikely candidates for credulous approaches. However, about a quarter of them simultaneously behave rashly on in-place questions. Hesitancy on in-place questions barely occurs in the data. Here we merely suggest some speculative interpretations and bear in mind that it is possible that this is a heterogeneous group.

In natural language, structures such as subject and predicate, placement of negation, and indeed ordering of premises are strongly related to credulous approaches to communication. Even when they do not affect the truth conditions of sentences, they focus the credulous hearer onto the speaker's intended model. In contrast, classical logic, a formalism developed for sceptical adversarial proof, actually gets rid of subject/predicate in favour of function argument organization. One sees this focusing through subject/predicate structure in operation in figural effects in syllogistic reasoning. The rare counterfigural syllogisms—those whose only valid conclusions reverse the subject/predicate status of the premise terms—are solved by few participants. For example, for the syllogism *No A are B*/*All B are C* in our data presented below, about 75% of subjects draw invalid ac conclusions, and very few find the valid conclusion *Some C are not A*. Simply reversing the subject/predicate structure of the first premise, which puts the term required to be subject of the conclusion into subject position in its premise, leads 45% of participants to get the correct ca term order, even though there is still a competing candidate subject term in the other premise.

Linguists refer to the focusing effects of structures where they do not change truth conditions as *information packaging* (e.g., Vallduvi, 1992). So one interpretation of hesitancy is that hesitant participants with credulous models of the discourse rely heavily on information packaging to guide their reconstruction of the speaker's intended model. The information-packaging structures particularly relevant to the syllogism are subject/predicate, term order, and the placement of existential and positive premises. Subjects with credulous models of the discourse who are attempting to construct the speaker's intended model around a particular type of individual on the basis of these packagings will prefer conclusions whose subject terms occur as participants in positive existential first premises. The detailed rationale for these latter effects is discussed below in the context of a model of conclusion term ordering.

If hesitant participants with credulous models of the discourse rely on such information packaging to guide them, then they will avoid applying closed-world reasoning to universal premises (which automatically changes subject/predicate organization). They will be particularly influenced by premise order and the placement of negatives. There will be many credulous inferences that they will not draw, even though they are adopting a generally credulous attitude to the discourse—their "blind faith" in the speaker's credulous information packaging will avert many fallacies, but it will also lead to refusal of classically valid inferences.

We offer this tentative interpretation acknowledging that not all hesitant participants are rash in place, so they may be a heterogeneous group. Furthermore, a deeper theoretical analysis must take on the differences between things and properties, terms and predicates, in participants' representations and processing. Whether participants think in terms of processing individuals and their properties, or in terms of sets, is an issue closely related to information packaging.

Back at the data, hesitancy on out-of-place questions is significantly negatively correlated with rashness on out-of-place questions ($r = -.50$, $p < .001$), presumably because rashness on out-of-place questions indicates an indifference to information packaging rather than because it indicates a credulous model of communication. These results provide evidence that we can usefully distinguish contrasts between credulous/adversarial models of communication on the one hand, and of the uses that participants may make of information packaging on the other.

In summary, merely exploring the data for large-scale patterns of interpretation has revealed several striking but previously unnoticed patterns, which transcend particular quantifiers. We are no longer constrained to looking for correspondences between single implicatures in interpretation tasks and reasoning tasks, any of which may be disrupted by many extraneous factors. Instead we can look for systematic influences of credulous models and information packaging on the task of reasoning.

THE SYLLOGISTIC REASONING TASK

Method

Participants

A few days after the immediate inference task, a subset ($N = 40$) of the participants did a syllogistic reasoning task and were paid for their participation. The numbers of these participants classified by their interpretation data, in each of the rashness and hesitancy classifications, is shown in Table 2.

These 40 subjects were then given the full set of 64 syllogisms. We refer to these data as the development dataset. Given each pair of premises, participants were asked whether there was any conclusion of the form *quantifier ac or quantifier ca* (where the possible quantifiers are *all, some, no, some . . . not*) that must be true whenever the premises are true. If not, they were instructed to respond no valid conclusion—NVC.

The second set of 62 participants who yielded the test dataset of interpretation data shown in Table 2 also did an identical syllogistic reasoning task to yield the reasoning test dataset. This dataset is modelled in Table 4 (see later).

Materials and procedure

A categorial syllogism consists of two premises, which relate three terms (a, b, and c), one of which (the *middle* term, b) occurs in both premises, while the other two (the *end* terms, a and c) each occur in only one premise—a is the end term in the first premise, and c is the end term in the second premise.

There are four *moods* or premise types, distinguished by the quantifiers "all", "some", "none", and "*some . . . not*". The quantifiers *all* and *none are universal*. The quantifiers *some* and *some . . . not are existential*. There are four possible arrangements of terms in the two premises, known as *figures*. We also make use of the term *diagonal figures* to refer to the first pair of figures (ab/bc and ba/cb) and symmetric figures to refer to the second pair (ab/cb and ba/bc). Since each premise can be in one of four moods, and each premise pair can have one of four figures, there are 4 × 4 × 4 = 64 different syllogisms. Of these, 27 have at least some valid conclusion under the assumption that the sets are nonempty; 37 have no such conclusion.

All problems were presented with the abstract terms a, b, and c. Participants were allowed to work through the problems in their own time and took between 40 minutes and an hour. The order of problems was randomized for each subject.

Results

Reasoning accuracy (development dataset)

Participants' conclusions were scored for accuracy as defined in classical logic with the noemptysets assumption—that is, on a VC problem, they were scored 1 if their conclusion was a valid conclusion, otherwise 0; on NVC problems they were scored 1 for responding "NVC", otherwise 0.

Three separate analyses of variance (ANOVAs) were conducted, with score as the dependent variable: (a) validity (whether the problem had a valid conclusion or not, VAL) by rashness in place (RI) by hesitancy out of place (HO); (b) validity by rashness out of place (RO) by hesitancy out of place; and (c) validity by rashness in place by rashness out of place as the dichotomized independent variables.

In each case, validity had a significant main effect on accuracy: Valid syllogisms are solved more accurately by all participants than syllogisms without valid conclusions: RI by RO by VAL analysis, main effect of VAL, $F(1, 36) = 20.33$, $MSE = 0.03$, $p < .001$; RI by HO by VAL analysis, main effect of VAL, $F(1, 36) = 12.46$, $MSE = 0.03$, $p = .001$); RO by HO by VAL analysis, main effect of VAL, $F(1,36) = 7.42$, $MSE = 0.03$, $p = .01$.

Rash in place and rash out of place each have significant main effects on conclusion accuracy: Either kind of rashness is associated with lower reasoning accuracy: RI by RO by VAL analysis, main effect of RI, $F(1, 36) = 7.48$, $MSE = 0.06$, $p = .01$; RI by HO by VAL analysis, main effect of RI, $F(1, 36) = 8.48$, $MSE = 0.06$, $p < .01$; RO by HO by VAL analysis, main effect of RO, $F(1, 36) = 6.97$, $MSE = 0.06$, $p = .012$.

The only significant interaction is between validity, rashness in place, and rashness out of place: RI by RO by VAL analysis, interaction of RI by RO by VAL, $F(1, 36) = 6.35$, $MSE = 0.03$, $p = .016$.

This interaction, shown in Table 3, is of the form that participants perform roughly equally well on problems with valid conclusions, but only those who are neither rash in place nor rash out of place are comparably accurate on problems without valid conclusions. All other effects failed to reach significance.

Discussion of syllogism accuracy results

The main effect of validity accords with all the data in the literature: Problems without valid conclusions are more difficult than problems with valid conclusions. Rashness of either kind is associated with lower accuracy across all problems,

Table 3. *Development dataset: The interaction of validity of problem with subjects' rashness on out-of-place problems and rashness on in-place problems in determining mean reasoning accuracy*

	Valid conclusion problems		*No valid conclusion problems*	
	Not rash IP	*Rash IP*	*Not rash IP*	*Rash IP*
Not rash OP	.63	.55	.60	.23
Rash OP	.56	.54	.35	.32

Note: OP = out of place. IP = in place.

a result consistent with rash participants having credulous understandings of the task. But rashness of either kind is not singly associated with more inaccuracy on NVC problems. It seems that a simple association between credulous models of the task and failure on NVC problems does not hold. This is related to Newstead's (1995) finding that although Gricean interpretation is common, it is not necessarily associated with drawing too many syllogistic inferences. However, we observed above that credulous approaches cover a wide range of specific logics and kinds of inference, and that disturbance of information packaging might save participants from invalid conclusions even though those conclusions follow from credulous implicatures. Some construals of the task might also reject deductively invalid Gricean conclusions for unrelated reasons (see, for example, Oaksford & Chater's, 1994, rational choice model).

There is an association between participants having both kinds of rashness and their reasoning accuracy on NVC problems. Essentially all rash out-of-place participants are also rash in place (though not vice versa). Many rash in−place participants are also hesitant on out-of-place items, or are simply neither rash nor hesitant on these items. If the former participants are being saved from overinference by their sensitivity to information packaging, then this might explain why rashness in place alone is insufficient to cause overinference on NVC problems, and why participants who are rash on both in-place and out-of-place questions do overinfer on NVC problems.

We return to the relations between reasoning accuracy and information packaging after we have developed a model of conclusion term ordering, which will allow a process-oriented analysis of reasoning accuracy.

Conclusion term ordering (development dataset)

Participants can draw conclusions in either of two term orders: *ac* or *ca*. This choice reflects participants' active information packaging of conclusions as influenced by problem structure. We now develop a statistical model for the effects of problem structure and individual differences in interpretation on the term orders of conclusions that subjects draw. The statistical framework is logistic regression: Structural problem variables and individual difference variables contribute to an equation predicting the probability of the reasoner drawing an *ac* conclusion, as opposed to a *ca* conclusion.

The term ordering data from both VC and NVC problems were modelled. Classically invalid conclusions generally show similar patterns of term ordering to those for valid conclusions. Participants' NVC responses, lacking term order, were discarded. Since doubly rash participants draw more conclusions from NVC problems, we included validity as an independent variable to check whether any effects are mediated through validity.

We seek a model that will identify the factors determining participants' choice of end-term order in drawing a conclusion. Participants could seize on any structural asymmetry in the problem and, on this basis, arrive at either premise order. Most simply, participants might choose to order their conclusion's terms in the order of their occurrence in the premises, leading always to *ac* conclusions. Or they might seize on structural asymmetry such as the placement of a unique quantifier, or the grammatical status of a term (subject or predicate) to order their conclusion's terms. For example, if just one of the premises contains a subject end term, this structure defines an asymmetry that could be the basis for putting the subject end term as subject of the conclusion (or for that matter as predicate). If both or neither premise contains such an end term, then this source of asymmetry cannot operate, and similarly for any other structural feature of problems.

The structural factors that we investigated were as follows: the sequence of premises, which is reflected by the *intercept* term of the regression model; the *grammar* of a problem encodes end term grammatical information (in Figure 1, AB BC, grammar is scored +1; in Figure 2, BA CB, grammar is scored −1; in Figures 3 and 4, grammar is scored 0); the presence of a unique *existential* premise (either *some* or *some . . . not*; this was encoded +1 if it was in Premise 1, and −1 if it was in Premise 2; and 0 if there were either two identical existential premises or none); similarly for the presence of a unique *all* premise, a unique *no* premise, for a unique *some . . . not* premise, and a unique *negative* premise.

Our aim was to reveal any patterns of individual differences in quantifier interpretation that predicted reasoning behaviour, specifically conclusion term order. The individual differences of interpretation investigated were as defined by the quantifier interpretation data described above: *hesitancy out of place* on subject/predicate reversed questions; *rashness in place* on subject/predicate preserved questions; and *rashness out of place* on subject/predicate reversed questions. *Hesitancy-out-of-place* scores ranged from 0 to 8 (mean = 2.30). *Rashness in-place* scores ranged from 0 to 4 (mean = 1.68); *rashness out-of-place* scores ranged from 0 to 8 (mean = 3.90).

Stepwise methods in regression analysis provide a useful tool for exploratory research in new areas (Hosmer & Lemeshow, 1989), where the focus is upon initially descriptive model building. The SPSS backwards elimination algorithm was used, since, compared to forward entry methods, backward elimination is less likely to exclude predictor variables that are relevant but that are involved in suppressor effects (Menard, 1995). A logistic regression model with independent variables was selected from the range described, predicting as dependent variable the proportion of *ac* conclusions on each of the 64 syllogism problems. All possible two- and three-way interactions between the structural variables were investigated. All two-and three-way interactions between *hesitancy*, *rashness in place* and *rashness out of place*, and structural problem variables and pairs of them were investigated.

Two abstract classifications of quantifiers were explored: positive versus negative; universal versus existential. The best models resulted from separating the variables for the universal quantifiers *all* and *no*, and the negative quantifiers *no* and *some . . . not*, but having a joint variable for the existentials (*some* and *some . . . not*) plus a distinguishing variable *some . . . not*. This arrangement has theeffect that for syllogisms with both *some* and *some . . . not* the variables *existential* and *some . . . not* both have the value 1. Note that such problems are NVC problems.

The logistic regression model of term conclusion order was developed on the development dataset and was tested on the test dataset. All variables included contribute significantly to fit, and adding any of the other variables fails to improve fit significantly, model $\chi^2(27) = 360.51, p < .0001$. The model correctly predicts the term order of 68% of conclusions of the dataset on which it was developed; the model's$\chi^2(27) = 620.37$, $p < .001$. The same model correctly classified marginally more of the test dataset (71% compared to 68% in Development dataset). The classification of the two datasets by this model is shown in Table 4, and the model's parameters in Table B1 (Appendix B1. Examination of the residuals in Tables C1 and C2 (Appendix C) suggests that the largest discrepancies in fit to the structure of problems are in predicting conclusions for problems in Figures 3 and 4 with *some . . . not* as Premise 1.

Before moving to an analysis of the model, we begin with comparison of development and test dataset fits. The effects of grammar, *all* and

Table 4. *Development and test dataset fits: The logistic regression model's classification of conclusion term order (*ac *and* ca*)*

		Development dataset			*Test dataset*		
		Predicted			*Predicted*		
		ac	*ca*	*% Correct*	*ac*	*ca*	*% Correct*
Observed	*ac*	729	257	75.66	1,274	323	79.8
	ca	318	511	63.21	474	677	58.8
	Overall			68.32			71.0

existential on term ordering are all significant and similarly signed as in the previous data. The interaction between existential and *some . . . not* is insignificant in the new data. The interaction between *all* and validity is similar but only marginally significant in these data. The main effect of hesitancy on out-of-place questions is roughly halved in size but is still significant. Its interaction with grammar is similar and significant. Its interaction with grammar and *none* is similar and still significant. The interaction between rashness in place and *some . . . not* is similar and significant, but its interaction with *some . . . not* and grammar is insignificant in these data. The interaction between rashness on out-of-place questions and *none* is similar and significant in the new data. On the whole, the model of the old data fits the new data rather well. The interactions with *some . . . not* are less well supported than those with *no*. The remainder of our discussion focuses on the fit to the development dataset.

The model shows that several structural and individual difference variables, and interactions between them, contribute significantly to the determination of conclusion term order. The positive constant in the equation (coefficient = 0.2098) reflects the overall bias towards *ac* conclusions seen in Table 4 and observed in all published studies. The end term from the first premise tends to occur as the subject of the conclusion. The end term from the second premise occurs as the predicate of the conclusion. Although the intercept constant marginally fails to reach significance, we see that specific subgroups of subjects exhibit more strongly this tendency to place terms in the order they occur in premises.

Of the structural variables, grammar makes the greatest contribution to fit. Figure 1 problems produce *ac* conclusions, and Figure 2 *ca* conclusions. This effect can be summarized by saying that where end terms are of different grammatical category (i.e., in Figures 1 and 2), they tend to preserve those categories in conclusions. This effect again accords with all other studies. Note that this effect preserves one particular aspect of a problem's information packaging in conclusions.

No systematic analysis of quantifiers' effects on conclusion term order has ever been conducted before. The overall main effects of the quantifiers' positions on conclusion term order are summarized in Table 5. We observed above that Stenning and Yule's (1997) source-founding model of conclusion term order assumes that participants use heuristics to determine the source premise on which to found conclusions. The most important heuristic is that any unique existential premise is the source premise. The regression model shows that after grammar, a variable built into the foundations of source-founding model, the most powerful structural determinant of term order is the existential variable: Unique existential quantifiers identify source premises, and their end terms become the subjects of conclusions. This effect works powerfully when the existential is in either premise, but more strongly when it is in the second where it operates against the congruence of conclusion terms with premise order captured by the intercept.

Because of the way the quantifier variables are defined, the interaction between existential and *some . . . not* applies only in problems with *some . . . not* in Premise 1 and *some* in Premise 2. In these problems, the negative coefficient means that there is an added preference to take the positive quantifier as source (against premise order), as the source-founding heuristics suggest.

The source-founding model's heuristics of preferring existential to universal and positive to negative premises as source combine to make *no* premises the least preferred sources. Our results show that *no*, alone among the quantifiers, does not have any main effect of placing its end term in subject position of the conclusion. All the other quantifiers' Premise 1 net coefficients are positive, and all the Premise 2 net coefficients

Table 5. *Development dataset: Summary of contributions to Z of the quantifiers' premise position, in the logistic regression model*

	All	*Some*	*None*	*Some . . . not*
Premise 1	0.5012	0.7112	—	0.0228
Premise 2	−0.9762	−1.1777	—	−1.0511

are negative (see Table 5). This means that each quantifier except *no* tends to put its end term into *subject* position, though to different extents.

The sizes of the coefficients for each quantifier in the two premises vary. The negative Premise 2 coefficients are absolutely larger than the positive Premise 1 coefficients, so unique quantifiers have the overall impact of producing a tendency toward *ca* conclusions, except when the second quantifier is *no*, and especially when the first quantifier is *no* or *some . . . not*. Note that it is a consequence of the way that the quantifier variables are defined that they do not affect problems with repeated quantifiers, and so this result means that these problems with repeated quantifiers have a relatively greater tendency for *ac* conclusions. When there is no quantifier asymmetry, premise order is used to break symmetry.

The only significant effect of whether a problem has valid conclusions is an interaction with *all*, whose effect is diminished in NVC problems. It is reassuring that the existence of valid conclusions plays little role in determining conclusion term order, indicating that similar processes determine term order whether participants are drawing conclusions validly or in error. Again the evidence in the literature is that term order effects in NVC problems are similar to those in VC problems.

To summarize the structural effects, when the grammatical categories of the two end terms are different (i.e., in diagonal Figures 1 and 2), terms strongly tend to preserve their category in conclusions (Figure 1 = *ac*; Figure 2 = *ca*). Quantifiers except *no* tend to make their premise's end terms into subjects, and more so when in Premise 2. These tendencies generally work against the overall tendency just noted to put the terms in premise order. *Some . . . not* acts mainly when in Premise 2.

On top of these structural effects across all subjects, the individual differences between subjects differentially affect conclusion term ordering. The only main effect of any individual difference variable is that hesitancy increases the general tendency toward *ac* responding throughout. Hesitancy interacts with grammar. Although hesitant participants have a greater tendency to conclude *ac*, preserving the premise ordering of terms, they are also more likely to override premise order by grammar in Figure 2 and draw a *ca* conclusion.

Although *no* is alone among the quantifiers in having no main effect, it interacts with both hesitancy and rashness out of place in three-way interactions with grammar. Without going into the full detail, these interactions are consistent with the view that hesitant participants are more sensitive, and rash out-of-place participants less sensitive, to the logical influence of *no* on conclusion term order noted above.

The source-founding model and individual differences

This exploratory regression model is rather complex. It clearly demonstrates that there are systematic effects of patterns of quantifier interpretation on participants' term ordering when they reason. These patterns are grossly comprehensible in terms of differential sensitivities to information packaging. However, the complexity of the model makes it hard to digest, and we want to relate conclusion term order to reasoning processes. Stenning and Yule's (1997) source-founding model aids interpretation. The main goal of this modelling is to understand the nature of the interactions between the negative quantifiers and the individual differences in quantifier interpretation.

The source-founding model was described earlier. Here we supplement the model to account for the interactions between interpretation patterns (particularly hesitancy and rashness out of place) and the quantifiers (particularly the negative quantifiers). In this model, the main locus of operation of effects is on the choice of source premise. Hesitant participants' preference for maintaining premise order in their conclusions constitutes a preference for choosing the first premise as source, thereby processing the discourse sequentially. Rash out-of-place participants are less sequential. To further understand the interactions between participants' patterns of interpretation with negative quantifiers we first need to examine some logical generalizations about *no* and then to re-represent the data in a way that brings out strategies for source premise choice.

The relation of *no* to classically valid conclusion term ordering is logically as well as empirically unique. For the range of problems where logic determines the subject/predicate structure of valid conclusions (i.e., problems with valid conclusions in only one term order), we can ask how their quantifiers influence that structure logically. It turns out that in all such problems with *no*, the end term of the *no* premise always can become the predicate of the conclusion. In other words, the end term of a *no* premise is never obligatorily the subject of the only valid conclusion. No other quantifier has this simple logical relationship to conclusion structure. There is a tendency for end terms of *all* premises to become subjects of valid conclusions, but there are exceptions. *Some* and *some . . . not* premises each contribute roughly equal numbers of subject and predicate terms to uniquely ordered valid conclusions. As far as we know this generalization has not been noted before. This logical generalization is reflected in Stenning and Yule's (1997) source-founding model of reasoning. The two heuristics (pick existential over universals and then positives over negatives) determine that *no* premises are never the source identified by the algorithm. They may be alternative sources but they are never the source picked by these heuristics.

In order to help understand the interactions between interpretation pattern and term order in reasoning, we need a representation of the data that emphasizes the influence of premise ordering on source identification. Accordingly, we can re-represent the problems as pairs that are related by premise reordering (e.g., *All A are B/All B are C* and *All B are A/All C are B* are such a pair). The 27 problems with valid conclusions are composed of 13 such pairs and one singleton, which is symmetrical about this ordering (*All B are A. All B are C*—reordering these premises merely leads to reassigning the end terms).

The source identification heuristics of the model provide a criterion for defining *canonical* and *noncanonical* orderings of premises for each of these 13 problem pairs. We call the problem with the source premise first (the one identified by the heuristics above) *canonical* and the other member of the problem pair *noncanonical.* There is one *all/all* problem pair that has to be ordered by its grammar (*All AB. All BC* is the canonical problem of this pair). This definition of canonical problems according to the source premise identifying heuristics means that *no* premises are always second premises in canonical problems.

Having defined canonicality in terms of the source premise identifying heuristics and observed the peculiar logical properties of *no*, we now apply canonicality to understanding the relation between term order and individual differences in processes of drawing conclusions. In canonical problems, participants read a source-founding premise first, and if they have strong tendencies to sequentially construct their representations, then we can expect that they will find canonical problems easier. If, on the other hand, they are rather indifferent to the surface order of arrival, canonicality of problem should have less impact. As we have already seen, hesitant participants are in general much more susceptible to premise order in choosing conclusion term order.

Hesitancy interacts with *no* and with *no* and *grammar* together. Taking the three-way interaction first, because the regression model's significant term for the hesitancy by grammar by *no* interaction is with *no* in Premise 2, it affects only canonical problems. Overall, hesitant participants draw even more premise ordered conclusions from canonical *no* problems than they do from canonical problems in general. When premise order and the properties of *no* line up with the source-selecting heuristics, the two have an intensifying effect on hesitant participants' conclusion term ordering.

Next, we consider the interaction between rashness out of place and *no*. Because the regression model's significant terms for the rashness out of place by *no* interaction are with *no* in Premise 1, they affect only noncanonical problems. Because the coefficient's sign is positive it increases the number of *ac* conclusions, which are here conclusions with terms ordered noncanonically. So rash out-of-place participants here tend to treat the *no* premise like the other quantifiers.

Finally, canonicality can help us to understand the interactions between rash in place, *some . . . not*, and the three-way interaction of both with grammar. The significant terms of both two- and three-way interactions are with *some . . . not* in Premise 2, and so these effects are on noncanonical problems (save for a single exception that we return to later). Even though in Figure 2 problems, the *some . . . not* end term is subject, rash-in-place participants still prefer the other quantifier as source, thus showing an unusual indifference here to grammatical structure.

So with all these interactions between individual differences in interpretation and negative quantifiers in determining term ordering in reasoning, hesitant participants tend to be more influenced by premise order and by *no* in their choice of source premise. Rash participants are less affected by premise order and the negativeness of quantifiers than are other participants.

Canonicality can also help us to understand the part that interpretation differences play in determining reasoning accuracy as mediated by conclusion term ordering. However, for such analysis we need to pool the data from the development and test datasets.

With the larger dataset it is possible to pursue the question of whether interpretation patterns' effects on reasoning accuracy can be shown to be mediated through their effects on conclusion term order. Experiment 1 showed that interpretation patterns had gross effects on reasoning accuracy, but the concept of canonicality and the source-founding model can help us to explore relations between interpretation and reasoning accuracy in a much more articulated way. If the source-founding model is correct, the different weights given by different subject groups to the factors determining source premise should influence reasoning accuracy as well as conclusion term order. Certain information packagings will hide certain conclusions from certain subgroups of participants. Canonicality provides a way of analysing the effect of premise order on reasoning by controlling all the other factors influencing choice of source premise. Generally, if any group finds canonical problems easier than their noncanonical counterparts, that means that premise order is playing some instrumental role in those participants' reasoning, because these pairs of problems differ only by their premise order.

From the statistical model of conclusion term ordering, we know that hesitant participants are more influenced in their conclusion term order by premise order than are rash participants. Similarly, from the interpretation task, we know that rash-out-of-place participants are particularly indifferent to term order in drawing implicatures. Are these participants duly more, or less, affected in their term ordering and reasoning accuracy by canonicality of problem? An especially informative place to look for interactions between premise order and the source premise identifying heuristics is the exceptional problems for which the crude version of the heuristics construct invalid conclusions.

There is just one problem pair (out of 13) for which the heuristic of adding the end term of the non source premise (as identified by the source premise identifying heuristics) to the end of the source premise, rules out drawing a valid conclusion. This is the Figure 4 pair: *Some B are not A*/*All B are C* (canonical); and *All B are A*/*Some B are not C* (noncanonical). The conclusion-drawing heuristic (attach the end term of the non source premise to the end of the source premise and remove the middle term) applied to these problems yields the conclusion *Some A are not C* for the canonical first problem and *Some C are not A* for noncanonical second problem, whereas the valid conclusions are the other way round. Note that this is because to get the valid conclusion the heuristic requires adjustment to the negations as it initially yields a negated predicate as subject term.

This pair of problems therefore provides an interesting test case for interactions between interpretation patterns and reasoning accuracy, as mediated by conclusion term ordering. We can usefully compare this problem pair with two other pairs. The Figure 3 problem that has the same quantifiers has valid conclusions, which are found by the source premise identifying and conclusion-drawing heuristics: *Some A are not B*/*All C are B* (canonical) and *All A are B*/*Some C are not B* (noncanonical). The simple heuristics work on these problems because the end terms are both subjects, just as they fail to work in the Figure 4 example because the end terms are both predicates. The second insightful comparison is with the Figure 1/2 pair where canonicality is resolved only by grammar: *All AB*/*All BC* (canonical) and *All BA*/*All CB* (noncanonical). Since grammar is a powerful resolver of conclusion term order for all interpretation groups, and it does not interact with *all*, this problem pair provides a control. The obvious interpretation dimension to explore is hesitant versus. rash out of place since these are negatively correlated and identify two almost disjoint groups.

Canonicality interacting with the source premise identifying heuristics makes rather precise predictions about hesitant and rash out-of-place participants' reasoning accuracy for these three pairs of problems. The *all*/*all* problem pair should show a reasoning accuracy advantage for the canonical problem over the noncanonical problem for both interpretation groups, because grammar overrides premise order in determining choice of source.

Hesitant participants for whom premise order has a strong influence on choice of source should show a strong canonicality advantage for the *some . . . not*/*all* problem pair in Figure 3. Rash out-of-place subjects, who are not much influenced by premise order, should show little canonicality effect on this pair of problems. But for the *some . . . not*/*all* problem pair in Figure 4, hesitant participants should show a reverse canonicality effect because determining source by premise order gives the right conclusion term order in the anticanonical problem and the wrong one in the canonical problem. Again, rash out-of-place participants should show little canonicality effect here.

The data from Experiments 1 and 2 were pooled. This yielded 17 hesitant participants, 43 rash out-of-place participants, and 42 participants who were neither (Table 2).

One participant was excluded from the latter group due to missing data.

An ANOVA was conducted with the within-group factors canonicality (two levels), problem pair (three levels), and group factor (hesitant vs. rash out of place). The dependent variable was reasoning accuracy adjusted by subtracting the participant's score on the problem from the mean score on that problem of the group of participants who were neither hesitant nor rash out of place. This was done to remove some of the effects of absolute difficulty of problem. The results showed that there was no main effect of canonicality of problem and no main effect of subject group. There was a significant interaction between canonicality and problem pair, $F(z) = 3.68$, $p = .028$, and between subject group, canonicality, and problem pair, $F(z) = 3.59$, $p = .031$. The means for the three-way interaction appear in Table 6.

The interaction is of the form that when grammar identifies source premise, both hesitant and rash participants are equal to their nonhesitant, nonrash peers for the canonically ordered problem, but both suffer roughly equally when the premise order is anticanonical. When grammar does not identify source, rash participants suffer relative to their nonrash, nonhesitant peers regardless of the premise order. However now hesitant participants show a sensitivity to whether premise order defines source accurately. In the standard problem where the source identified by the model's heuristics can be used to make the simple construction of the correct conclusion, hesitant participants actually outperform their nonhesitant, nonrash peers on the canonically ordered member of the pair, but underperform them on the noncanonical member. In the exceptional problem where the source identified by the heuristic cannot be so used, they underperform their peers on the canonically ordered problem but outperform them on the noncanonical member of the pair. These results are consistent with the idea that hesitant participants tend to identify Premise 1 as source, whereas rash out-of-place participants are less affected by premise order and more by grammar and the quantifier attributes that drive the heuristics. Note that whereas there

Table 6. *Combined development and test datasets: Canonicality advantage on reasoning accuracy scores*

	Subject group			
	Hesitant		*Rash-out-of-place*	
Problem pair	*M*	*S.E.*	*M*	*S.E.*
All AB. All BC. So, All AC	−.02	.09	−.09	.06
All BA. All CB. So, All CA	−.23	.12	−.22	.08
Some A not B. All CB. So, Some A not C	+.11	.12	−.16	.08
All AB. Some C not B. So, Some C not A	−.16	.12	−.19	.08
Some B not A. All BC. So, Some C not A	−.22	.12	−.10	.08
All BA. Some B not C. So, Some A not C	+.12	.12	−.17	.08

Note: Adjusted for nonhesitant non-rash-out-of-place mean scores, for three canonical/noncanonical problem pairs and for hesitant and rash-out-of-place subject groups. The top member of each problem pair is canonical. A positive value means the participants are more accurate than non-hesitant, non-rash-out-of-place subjects: A negative score means they are less accurate.

are no global differences in accuracy, there are large and opposite effects on groups of problems for different subject groups.

Summary of experimental results

The regression models' fits to both sets of data show that groups of participants classified by their interpretation of quantifiers exhibit radically different patterns of term ordering. Even if we take some of the most powerful effects known in the syllogistic reasoning literature (such as the effect of Figure 1 vs. Figure 2 on term order) the model allows us to find subgroups of participants and subgroups of problems that systematically fail to show these effects, or even show reversals of them. Group data, in fact, are highly misleading here.

We treated Grice's (1975) theory as a broad framework for a range of credulous reasoning processes and therefore adopted an exploratory approach to yield two dimensions for classifying participants' interpretations. We looked for effects of patterns of interpretations on patterns of reasoning. Highly structured patterns of interpretation do have systematic effects on participants' reasoning. The novel observations of individual differences in interpretation affecting processes of reasoning can be accommodated smoothly into the source-founding model. Indeed, without the model it would be hard to give an overall picture of how interpretation affects reasoning.

An exploratory approach throws up new empirical phenomena to explain. The hesitant are an important group of hitherto unnoticed participants who underinfer and who are saved from classical fallacies by their sensitivities to information packaging. Rash out-of-place participants play counterpoint to them. The distinctive logical and psychological properties of *no* have also not been remarked on before. Exploratory results are messy and complicated. We could present more examples of subgroups of participants behaving oppositely on subgroups of problems but space limitation forbids. Such messy explorations are necessary if premature dismissals of theories are to be avoided.

GENERAL DISCUSSION

Where do these findings leave our interpretational approach to human reasoning more generally? Participants have a variety of credulous interpretations of these tasks which they carry over from reasoning in the one-sentence to the two-sentence task. The source-founding model is a useful abstraction over many details of interpretation and representation, which allows differences in strategies to be explored. Rashness and hesitancy are likewise coarse concepts close to the data. We have not offered logical models here, but rather evidence that a variety is necessary, and that they should cross tasks (see Stenning & van Lambalgen, in press, for the detailed development and neural implementation of such a default logical model). We agree with Roberts, Newstead, and Griggs that Gricean interpretations play important roles (several roles: Grice (1975) does not provide a single interpretation or processing model) in the immediate inference task but have provided evidence that that these roles continue into syllogistic reasoning. We agree with Chater and Oaksford (1999) that the reasoning processes of most participants do not correspond to deductions in classical logic, but we would point out that their resort to classical probability theory assumes that classical logic *is* the underlying logic of all subjects. However, we disagree with both groups of researchers in that we believe that group models are unjustified and misleading.

The family of default logical models of these processes that we envisage are, however, closely related to Oaksford and Chater's (2001) computational-level probabilistic models. Indeed, they might be seen as offering qualitative models of reasoning about likelihoods through their databases of defeasible conditionals. If this can be substantiated, then it would bring additional benefits, because these logics offer plausible process models. Our approach does reveal how close credulous models are to the classical sceptical one in the domain of the syllogism, and the source-founding model particularly offers a way of expressing this continuity through strategic variations.

Mental models theorists have recently claimed that their theory encompasses defeasible reasoning, and, as mentioned in the Introduction, Bonnefon (2004) has shown that mental models theory lies somewhere between sceptical and classical reasoning. It is obvious that if mental models theory can model many different logical systems (at least classical, modal, probability, and nonmonotonic reasoning have been claimed), the system is in theory capable of modelling individual differences. Furthermore, it is clear on purely logical grounds that if allowed access to the full panoply of set theory in its metalanguage (as it seems to be), just about any reasoning can be described as manipulations on models. However, then the interesting systems can be captured in proof theories. Our complaint is with the effect that claiming a "single fundamental human reasoning mechanism" has on the investigation of individual variety. From the earliest papers (e.g., Johnson-Laird & Steedman, 1978), mental models theory described many subject' reasoning as "failure to search for alternative models in which the candidate conclusion is invalid". But such patterns just are defeasible reasoning to what the subject takes to be "intended models". Is the subject failing to reason classically or succeeding in reasoning defeasibly? When competence theories are entwined with performance theories, the question does not seem to arise.

How is one to view these individual differences between subject' interpretation and reasoning patterns? Are they traits that persist as the same subject travels across all contexts? Or are they more subtle differences in which contexts evoke which kinds of reasoning from subject? We favour the latter view, believing that most participants have access to most reasoning systems in some context or other. Subjects perhaps differ mainly in which situations evoke which systems, and especially in their strategic flexibility in choosing appropriate systems in the abstracted tasks of the reasoning laboratory and in formal education. Taking subjects' own interpretations seriously in evaluating their reasoning does not entail that all interpretations be treated as equally appropriate or justified. A beginning has been made on studying issues of reasoning styles by investigating individual differences in diagrammatic reasoning and learning (Cox, 1999; Stenning, 2002; Stenning, Cox, & Oberlander, 1995). To find out whether this approach is fruitful will require an empirical focus on what is common in the same subject's reasoning as she or he faces different tasks in different contexts.

Original manuscript received 14 June 2004
Accepted revision received 24 May 2005
First published online 25 October 2005

REFERENCES

Bonnefon, J.-F. (2004). Reinstatement, floating conclusions, and the credulity of Mental Models Theory. *Cognitive Science*, *28*(4), 621–631.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*, 61– 83.

Chapman, K. J., & Chapman, J. P. (1959). The atmosphere effect reexamined. *Journal of Experimental Psychology*, *58*, 220–256.

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

Cox, R. (1999). Representation construction, externalised cognition and individual differences. *Learning and Instruction*, *9*, 343–363.

Dickstein, L. S. (1975). Effects of instructions and premise order on errors in syllogistic reasoning. *Journal of Experimental Psychology: Human Learning and Memory*, *104*, 376–384.

Evans, J. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics Vol. 3. Speech acts*. London: Academic Press.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: John Wiley & Sons.

Johnson-Laird, P. N., & Bara, B. (1984). Syllogistic inference. *Cognition*, *16*, 1–82.

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

Menard, S. (1995). *Applied logistic regression analysis.* Thousand Oaks, CA: Sage Publications.

Newstead, S. E. (1989). Interpretation errors in syllogistic reasoning. *Journal of Memory and Language*, *28*, 78–91.

Newstead, S. E. (1989). Interpretation errors in syllogistic reasoning. *Journal of Memory and Language*, *28*, 78–91.

Newstead, S. E. (1995). Gricean implicatures and syllogistic reasoning. *Journal of Memory and Language*, *34*, 644–664.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Politzer, G. (1990). Immediate deduction between quantified sentences. In K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & G. Erdos (Eds.), Lines of thinking: *Vol. 1. Representation, reasoning, analogy and decision making*. John Wiley & Sons London.

Roberts, M. J., Newstead, S. E., & Griggs, R. A. (2001). Quantifier interpretation and syllogistic reasoning. *Thinking and Reasoning*, *7*(2), 173–204.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Stenning, K. (2002). *Seeing reason: Language and image in learning to think*. Oxford, UK: Oxford University Press.

Stenning, K., & Cox, R. (1995). Attitudes to logical independence: Traits in quantifier interpretation. In J. D. Moore & J. Fain Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 742–747). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Stenning, K., Cox, R., & Oberlander, J. (1995). Contrasting the cognitive effects of graphical and sentential logic teaching: Reasoning, representation and individual differences. *Language and Cognitive Processes*, *10*(3/4), 333–354.

Stenning, K., & van Lambalgen, M. (2004). A little logic goes a long way: Basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science*, *28*, 481–529.

Stenning, K., & van Lambalgen, M. (in press). Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science*.

Stenning, K., & Yule, P. (1997). Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology*, *34*, 109–159.

Stevens, J. P. (2001). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Vallduv́, E. (1992). *The informational component*. New York: Garland.

Wason, P. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

APPENDIX A

Response accuracy—immediate inference task

Table A1. *Development dataset: Proportion of subjects responding "True", "False", and "Can't tell" to immediate inference questions, along with correct responses and missing data (no response)*

		Premise							
		ALL		*NO*		*SOME*		*SOME NOT*	
Conclusion		*Prop.*	*Newst.*	*Prop.*	*Newst.*	*Prop.*	*Newst.*	*Prop.*	*Newst.*
TRUE	All	0.99		0.01		0.09		0.06	
	All′	0.33	0.57	0.02		0.07		0.03	
	No	0.00		0.94		0.02		0.07	
	No′	0.03		0.52	0.80	0.01		0.05	
	Some	0.83		0.02		0.97		0.35	0.83
	Some′	0.67	0.87	0.03		0.69	0.87	0.27	0.77
	Some. . .not	0.03		0.76		0.45	0.93	0.95	
	Some. . .not′	0.15	0.47	0.34	0.77	0.34	0.83	0.50	0.90
FALSE	All	0.01		0.94		0.34		0.87	
	All′	0.15		0.64		0.26		0.46	
	No	0.99		0.01		0.96		0.49	
	No′	0.68		0.06		0.67		0.34	
	Some	0.13		0.93		0.01		0.07	
	Some′	0.07		0.54		0.01		0.05	
	Some. . .not	0.92		0.14		0.05		0.00	
	Some. . .not′	0.34		0.10		0.07		0.03	
CAN'T TELL	All	0.00		0.00		0.55		0.06	
	All′	0.52		0.27		0.64		0.47	
	No	0.01		0.00		0.00		0.43	
	No′	0.29		0.36		0.30		0.56	
	Some	0.03		0.00		0.00		0.54	
	Some′	0.26		0.37		0.28		0.64	
	Some. . .not	0.05		0.04		0.48		0.03	
	Some. . .not′	0.51		0.32		0.57		0.42	
MISSING DATA	All	0.00		0.05		0.02		0.01	
	All′	0.00		0.07		0.03		0.04	
	No	0.00		0.05		0.02		0.01	
	No′	0.01		0.06		0.02		0.05	
	Some	0.00		0.05		0.02		0.04	
	Some′	0.00		0.06		0.02		0.04	
	Some. . .not	0.00		0.06		0.02		0.02	
	Some. . .not′			0.06		0.02		0.05	
CORRECT	All	T		F		CT		F	
	All′	CT		F		CT		CT	
	No	F		T		F		CT	
	No′	F		T		F		CT	
	Some	T		F		T		CT	
	Some′	T		F		T		CT	
	Some. . .not	F		T		CT		T	
	Some. . .not′	CT		T		CT		CT	

Note: Primed conclusion quantifiers (e.g., A′) represent the converse conditions (e.g., *ALL Bs are As*). Newst. = Newstead's (1989) results if the results of the present study differ by more than .07 from those reported by Newstead (1989, Table 2, p. 86).

APPENDIX B

Logistic regression models of *ac* conclusion probability for development and test datasets

Table B1. *The logistic regression model of subjects' probability of* ac *conclusion: Development and test data*

	Model development data						*Model test data*					
Variable	*B*	*SE*	*Wald*	*df*	*Sig*	*Exp(B)*	*B*	*SE*	*Wald*	*df*	*Sig*	*Exp(B)*
Grammar			86.0162	2	0.0000	0.1810			175.705	2	0.000	
Figure 1	1.0830	0.1849	34.3038	1	0.0000	0.1136	0.958	0.164	34.237	1	0.000	2.607
Figure 2	−0.8903	0.1694	27.6283	1	0.0000	−0.1012	−1.450	0.147	97.416	1	0.000	0.235
All			15.1603	2	0.0005	0.0668			23.574	2	0.000	
All in Premise 1	0.5012	0.2365	4.4914	1	0.0341	0.0316	0.392	0.189	4.310	1	0.038	1.480
All in Premise 2	−0.9762	0.2509	15.1436	1	0.0001	−0.0725	−0.978	0.202	23.542	1	0.000	0.376
Existential			21.6887	2	0.0000	0.0841			50.409	2	0.000	
Existential in Premise 1	0.7112	0.2628	7.3246	1	0.0068	0.0461	0.493	0.211	5.463	1	0.019	1.638
Existential in Premise 2	−1.1777	0.2558	21.1979	1	0.0000	−0.0876	−1.460	0.206	50.167	1	0.000	0.232
Existential (i.e., Some) * Some_not			8.3517	2	0.0154	0.0417			1.350	2	0.509	
Some_not in Premise 1	−0.6884	0.2426	8.0502	1	0.0045	−0.0492	−0.078	0.214	0.133	1	0.716	0.925
Some_not in Premise 2	0.1266	0.2530	0.2505	1	0.6167	0.0000	0.240	0.221	1.183	1	0.277	1.271
All * Validity			4.8619	2	0.0880	0.0186			5.483	2	0.064	
All in Premise 1 by Invalid	−0.1021	0.2922	0.1220	1	0.7268	0.0000	0.300	0.226	1.765	1	0.184	1.350
All in Premise 2 by Invalid	0.5971	0.2796	4.5587	1	0.0328	0.0320	0.477	0.235	4.133	1	0.042	1.611
Hesitant-out-o-place	0.1718	0.0347	24.5091	1	0.0000	0.0948	0.069	0.029	5.817	1	0.016	1.072
Hesitant-out-o-place * Grammar			23.7547	2	0.0000	0.0888			15.222	2	0.000	
Figure 1	0.0082	0.0713	0.0131	1	0.9089	0.0000	0.140	0.070	3.955	1	0.047	1.150
Figure 2	−0.2900	0.0620	21.9138	1	0.0000	−0.0892	−0.147	0.052	8.099	1	0.004	0.863
Hesitant-out-o-place * None			9.1877	2	0.0101	0.0455			5.295	2	0.071	

None in Premise 1	−0.0026	0.0604	0.0018	1	0.9657	0.0000	0.008	0.052	0.022	1	0.881	1.008
None in Premise 2	−0.1815	0.0599	9.1742	1	0.0025	−0.0535	−0.118	0.052	5.138	1	0.023	0.888
Hesitant-out-o-place * None * Grammar			13.2385	4	0.0102	0.0458			9.983	4	0.041	
None in Premise 1 in Figure 1	−0.1508	0.0984	2.3503	1	0.1253	−0.0118	−0.056	0.111	0.252	1	0.616	0.946
None in Premise 1 in Figure 2	0.1230	0.0986	1.5564	1	0.2122	0.0000	0.177	0.088	4.025	1	0.045	1.193
None in Premise 2 in Figure 1	0.1738	0.1456	1.4245	1	0.2327	0.0000	0.024	0.119	0.042	1	0.838	1.024
None in Premise 2 in Figure 2	0.2622	0.0948	7.6449	1	0.0057	0.0475	0.225	0.083	7.256	1	0.007	1.252
Rash-out-o-place * None			14.8556	2	0.0006	0.0659			5.043	2	0.080	
None in Premise 1	0.1706	0.0443	14.8257	1	0.0001	0.0716	0.062	0.032	3.754	1	0.053	1.064
None in Premise 2	−0.0666	0.0437	2.3264	1	0.1272	−0.0114	−0.058	0.032	3.352	1	0.067	0.943
Rash-in-place * Some_not			24.8569	2	0.0000	0.0913			10.590	2	0.005	
Some_not in Premise 1	0.0296	0.0895	0.1090	1	0.7413	0.0000	−0.136	0.061	4.899	1	0.027	0.873
Some_not in Premise 2	−0.4546	0.0920	24.4149	1	0.0000	−0.0946	−0.165	0.064	6.737	1	0.009	0.848
Rash-in-place * Some_not * Grammar			25.1133	4	0.0000	0.0827			3.424	4	0.490	
Some_not in Premise 1 in Figure 1	0.0573	0.1683	0.1158	1	0.7336	0.0000	−0.043	0.100	0.183	1	0.668	0.958
Some_not in Premise 1 in Figure 2	0.0086	0.1403	0.0038	1	0.9510	0.0000	0.133	0.098	1.813	1	0.178	1.142
Some_not in Premise 2 in Figure 1	0.3779	0.1519	6.1889	1	0.0129	0.0409	0.015	0.101	0.021	1	0.884	1.015
Some_not in Premise 2 in Figure 2	0.6830	0.1386	24.2807	1	0.0000	0.0944	0.093	0.093	1.004	1	0.316	1.098
Constant	0.2098	0.1250	2.8180	1	0.0932		0.813	0.112	52.821	1	0.000	2.254

Note: Boldface indicates a level of a variable with significant effect in the development dataset.

APPENDIX C

Conclusion term order data by problem

Table C1. *Development dataset: Mean proportion* ac *conclusions, residuals of model's predictions, and number of observations, by syllogism*

	Premise 1											
	All			*Some*			*No*			*Some not*		
Premise 2	*Mean*	*Resid*	*N*	*Mean*	*Resid*	*N*	*Mean*	*Resid*	*N*	*Mean*	*Resid*	*N*
	Figure 1 ABBC											
All	0.90	0.05	39	0.84	0.03	38	0.69	−0.11	35	0.81	0.06	36
Some	0.74	−0.01	31	0.67	−0.18	24	0.63	−0.13	32	0.84	0.05	25
No	0.86	0.02	37	0.97	0.11	29	0.79	−0.06	14	0.80	−0.01	25
Some_not	0.75	0.09	32	0.79	0.01	24	0.75	0.07	16	0.76	−0.08	17
	Figure 2 BACB											
All	0.27	−0.10	37	0.30	−0.02	30	0.24	−0.06	38	0.24	0.00	33
Some	0.21	−0.04	39	0.31	−0.05	26	0.35	0.11	34	0.21	−0.07	24
No	0.40	0.04	35	0.45	0.05	29	0.38	0.02	13	0.27	−0.04	22
Some_not	0.25	0.08	32	0.35	0.07	23	0.25	0.07	24	0.33	−0.03	21
	Figure 3 ABCB											
All	0.79	0.16	28	0.75	0.17	32	0.63	0.08	40	0.65	0.17	37
Some	0.36	−0.12	28	0.60	−0.03	20	0.50	0.01	34	0.67	0.14	24
No	0.58	−0.04	36	0.65	−0.01	31	0.77	0.13	13	0.48	−0.10	21
Some_not	0.23	−0.15	31	0.27	−0.25	26	0.38	−0.01	21	0.47	−0.15	17
	Figure 4 BABC											
All	0.70	0.08	37	0.47	−0.10	38	0.54	0.00	35	0.29	−0.19	38
Some	0.53	0.05	36	0.68	0.05	19	0.41	−0.07	32	0.64	0.11	25
No	0.65	0.03	31	0.65	−0.01	31	0.67	0.04	15	0.41	−0.16	22
Some_not	0.54	0.02	37	0.48	−0.04	25	0.46	0.08	24	0.76	0.15	17

Table C2. *Test dataset: Mean proportion* ac *conclusions, residuals of model's predictions and number of observations, by syllogism*

	Premise 1											
	All			*Some*			*No*			*Some not*		
Premise 2	*Mean*	*Resid*	*N*	*Mean*	*Resid*	*N*	*Mean*	*Resid*	*N*	*Mean*	*Resid*	*N*
	Figure 1 ABBC											
All	0.93	0.05	61	0.85	0.03	62	0.80	0.00	51	0.84	0.03	57
Some	0.88	0.05	51	0.86	−0.02	37	0.63	−0.08	41	0.76	−0.07	34
No	0.89	0.01	61	0.90	0.03	52	0.73	−0.15	26	0.88	0.05	41
Some_not	0.84	0.05	58	0.82	−0.04	34	0.64	−0.07	25	0.77	−0.11	31
	Figure 2 BACB											
All	0.26	−0.07	62	0.21	−0.08	52	0.32	0.03	61	0.19	−0.02	51
Some	0.15	0.00	60	0.30	−0.02	36	0.27	0.07	49	0.15	−0.09	38
No	0.35	−0.06	47	0.47	0.10	49	0.36	0.03	22	0.41	0.12	31
Some_not	0.28	0.05	54	0.25	−0.01	29	0.17	−0.04	29	0.34	−0.03	30
	Figure 3 ABCB											
All	0.85	0.19	48	0.75	0.09	53	0.66	0.06	57	0.67	0.20	51
Some	0.42	−0.18	53	0.71	0.05	33	0.29	−0.19	51	0.64	0.05	28
No	0.53	−0.20	58	0.72	0.00	51	0.65	−0.05	17	0.56	−0.06	31
Some_not	0.34	−0.11	53	0.58	−0.07	32	0.36	−0.06	26	0.65	−0.04	28
	Figure 4 BABC											
All	0.92	0.22	52	0.53	−0.06	58	0.55	−0.02	51	0.25	−0.26	54
Some	0.69	0.24	59	0.76	0.05	24	0.40	−0.05	47	0.51	−0.09	35
No	0.70	−0.01	44	0.70	−0.03	50	0.92	0.22	13	0.62	−0.01	29
Some_not	0.75	0.06	57	0.69	0.03	29	0.59	0.17	29	0.84	0.14	25