Journal of Experimental Psychology: General
1986, Vol. 115, No. 1, 16-25

0096-3445/86/$00.75

Individual Differences in Syllogistic Reasoning: Deduction Rules or Mental Models?

Kathleen M. Galotti, Jonathan Baron, and John P. Sabini
University of Pennsylvania

Two studies examined the correlates of reasoning ability on a syllogistic reasoning task in subjects who lacked formal background in logic. The main issue addressed was the extent to which reasoning proficiency arises from the consideration of multiple possible set relations (mental models) as opposed to explicit or implicit reliance on deduction rules. Evidence for the use of both models and rules was obtained. Although "good" and "poor" reasoners differed even when time constraints were imposed (consistent with the supposition of a better set of rules among good reasoners), good reasoners showed more improvement and chose to take longer amounts of time when time constraints were removed, suggesting that they considered more alternatives than did the poor reasoners. A comparison between these two groups and a third group of subjects, graduate students who had studied logic, reveals striking differences in both accuracy and speed.

A central problem in the investigation of thinking is to describe how people reason deductively. A standard task to investigate this ability is comprised of categorical syllogisms (e.g., those with premises of the form, "All A are B," "Some C are B"). This task has long been of interest to experimental psychologists (Wilkins, 1928; Woodworth & Sells, 1935), and tests of syllogistic reasoning were included on early intelligence tests (Guilford, 1959; Thurstone, 1938). The topic of reasoning with categorical syllogisms has recently stirred renewed interest (e.g., Begg & Denny, 1969; Dickstein, 1975, 1976, 1978, 1981a, 1981b; Erickson, 1974, 1978; Fisher, 1981; Frase, 1968; Guyote & Sternberg, 1981; Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Steedman, 1978; Revlin & Leirer, 1978; Revlis, 1975a, 1975b; Roberge, 1970; Sternberg & Turner, 1981).

The general finding is that although some syllogisms are solved easily and quickly, the average untutored reasoner makes many errors, despite a variety of attempts by investigators to word the premises carefully and to explain the task thoroughly. Our work seeks to describe the general characteristics that an explanation of both correct and incorrect performance should have. Our concern is primarily with the nature of the reasoning task, and we approach this question, in part, by asking about the source of individual differences.

This article is based on a dissertation submitted by Kathleen Galotti to the Department of Psychology, University of Pennsylvania. Support was provided by the National Institute of Mental Health Grant MH 37241 to Jon Baron and by a grant from the Department of Psychology, University of Pennsylvania to Kathleen Galotti.

We thank Warren Stewart for help in running the studies, W. Francis Ganong and Rochel Gelman for profitable discussions, and John Bare and Lloyd Komatsu for comments on earlier drafts. We also thank Philip Johnson-Laird and Martin Braine for their detailed and thoughtful reviews.

Correspondence concerning this article should be addressed to Kathleen Galotti, who is now at the Department of Psychology, Carleton College, Northfield, Minnesota 55057.

The Task

A categorical syllogism consists of two premises and a conclusion, each of which describes the relationship between two sets of things. The first premise relates one term, A, to a second, B; the second premise relates B to a third term, C; and the conclusion states a relationship, if one exists, between A and C. Premises and conclusions are of four types: (a) universal affirmative (all A are B); (b) universal negative (no A are B); (c) particular affirmative (some A are B); and (d) particular negative (some A are not B). It is important to note that *some* functions existentially, meaning "at least one and possibly all." In everyday English usage, "Some A are B" is usually taken to imply that some A are not B, but in logic, *some* is intended narrowly.

In these studies, we consider only syllogisms of the form A–B; B–C.[1] Each of the two premises can be of four types, as described above, yielding 16 different combinations of premises. Only six of these have valid conclusions that definitely relate the A to the C term or the C to the A term. In the task itself, the conclusion can be supplied by the subject or by the experimenter in true/false or multiple-choice format.

Models of Syllogistic Reasoning

Early work on syllogistic reasoning (Begg & Denny, 1969; Ceraso & Provitera, 1971; Chapman & Chapman, 1959; Woodworth & Sells, 1935) focused almost exclusively on errors, giving little description of the processes used in deduction. Three different sources of error were proposed. The first, contained in the *atmosphere hypothesis* (Begg & Denny, 1969; Woodworth & Sells, 1935), proposed that all conclusions were based on global impressions created by surface characteristics of the premises.

[1] Recent work by Johnson-Laird (1983; Johnson-Laird & Bara, 1984) highlighted the importance of the effects of figure on performance. According to that model, the figure we have used is easiest. We discuss the implications of this later.

For example, if one or both of the premises contained the word *some*, the conclusion drawn would also contain the word *some*. An alternative account of errors, called the *conversion hypothesis* (Chapman & Chapman, 1959), held that errors occurred for two reasons: (a) Upon reading the premises, subjects automatically inferred the converses and reasoned from both the original premises and their converses; and (b) subjects reasoned in accordance with what they thought to be probable, rather than with what they thought was strictly necessary (see also Dickstein, 1976, and Henle, 1962).

Both the atmosphere and conversion hypotheses focus on the reasons for errors. The atmosphere account gives little description of the processes used in a deduction. The conversion account speaks somewhat more to process issues, but it also fails to provide a detailed description of the mental steps involved in a deduction.

More recent work has provided such descriptions, ranging from accounts that postulate general mechanisms of reasoning to ones that provide syllogism-specific accounts. In the studies below, we set out to test a general class of reasoning models rather than to evaluate any particular model. In order to describe this class, we first describe a classification of existing models.

A Conceptual Framework: Models Versus Rules

The two classes of existing models fall into what we call *models* accounts and *rules* accounts. The distinction between the two hinges on whether or not premises are assumed to be mapped onto representations of individual tokens or of set relationships. Models that include processes that generate such representations fall into the models class. In other words, in models accounts, one or more representations are generated that are consistent with a particular interpretation of the premises. The first representation generated is used to form a tentative conclusion (i.e., a preliminary idea of what the valid conclusion is). Additional representations, if generated, are used to evaluate the validity of the tentative conclusion.

Processes that operate solely on the form of the premises, that do not generate intermediate representations on the way to formulating a conclusion are in the rules class. That is, rules are assumed to take the place of consideration of particular interpretations of premises.

To clarify this distinction, first consider how a prototypical models account would work. The reasoner would first need to encode and mentally represent each premise, then to combine such encodings into one or more combined representations. A verbal statement consistent with all of these possibilities (a tentative conclusion) is made. The validity of this conclusion can be checked by exhaustively generating all possible models, looking for a falsifying one (i.e., one consistent with the premises but not consistent with the tentative conclusion).

What apparatus would such a model make use of? First, there must be a mechanism to encode premises and generate models. Because both single premises and pairs of premises often allow more than one interpretation, there must be some set of principles that govern the order in which possibilities are generated and considered and the number of total possibilities that get generated. There must also be some representation of the possibilities, or models, be they of three individuals (an A, a B, and a C), of three sets (all the As, all the Bs and all the Cs), or of some combination. There must be some mechanism that abstracts a verbal statement of the conclusion from a set of representations of alternative interpretations.

How would such a model account for individual differences? Individuals could in theory differ in (a) the mere possession of any of the mechanisms described above, (b) the efficacy or thoroughness with which the mechanisms operate, (c) the representations that are used, or (d) some combination of any of the above three alternatives. The most testable of these alternatives is the one that holds that good reasoners generate and test more models than do poor reasoners. This account in particular is the one we set out to test.

Now consider a prototypical rules account. Reasoning in this case proceeds without a mapping of premises into examples (of either a configuration of individuals or of sets). Instead, the process looks more like reliance on abstract templates that operate on the form of premises to yield conclusions. Rules might be triggered by quantifiers, negation, and/or the figure of the problem (e.g., "If both premises contain the word *some*, then there is no valid conclusion").

It is crucial to distinguish between deduction rules, described immediately above, and other rule-governed behavior. Certainly in the models class, there must be some regularity to the order of generation of possibilities, and one might wish to describe such regularities as rule governed. With deduction rules, we have something quite different in mind. Specifically, deduction rules do not generate intermediate examples. Although two or more deduction rules can be chained together in the course of reasoning, they never generate examples of possible states allowed by premises. Thus no tentative conclusions (only final conclusions) are generated.

How do deduction-rules models account for individual differences? People could differ either in the mere possession of rules, the "goodness" of rules possessed, the efficacy with which rules are applied, or some combination of these three.

Existing models can be classified with respect to this framework, although some are more typical instances. The work of Erickson (1974, 1978), Fisher (1981), Johnson-Laird (1982, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Steedman, 1978), Revlin & Leirer (1978), Revlis (1975a, 1975b), and Sternberg (Guyote & Sternberg, 1981; Sternberg & Turner, 1981) seems to fall into the models class. It is true that each of these authors proposes different representations and different mechanisms; still, all share the idea that premise representations generate possible states of affairs consistent with the premises. To reason validly, a reasoner must on any of the above accounts generate enough possibilities (models) either to (a) arrive at a contradiction to a tentatively held conclusion or (b) examine a sufficient number of possibilities to be sure that no contradiction exists. Errors in reasoning arise (either wholly or in part, depending on the model) from a failure to generate enough relevant models. (For a comprehensive overview of many of these models, see Johnson-Laird & Bara, 1984).

Proposals that come under our rules category include those of Braine (1978; Braine & Rumain, 1983) and others who deal mainly with propositional reasoning but adopt a similar approach (Osherson, 1974, 1975; Rips, 1983).

The studies discussed below address the following questions:

(a) Do reasoners reason primarily by generating multiple models or by relying on rules? (b) Which type of account—models or rules—offers a better explanation of individual differences? In particular, is good reasoning to be explained by a tendency to generate more models? (c) Assuming that individual differences are to be explained by a models account, are poor reasoners less able, or merely less willing to generate multiple models?

Experiment 1: Necessity Versus Possibility

In the first experiment, we looked to see whether differences between good and poor reasoners occurred early or late in the process of reasoning. Note that a rules approach (as well as some models approaches) holds that group differences would emerge in the construction of the initial conclusion. Other models accounts, on the other hand, hold that group differences would arise solely or largely because good reasoners generate multiple models, in particular, ones that falsify the tentative conclusion. Group differences on these accounts ought to show up after tentative conclusions are formulated.

We asked subjects on some trials to give only what we called a possible conclusion, that is, a relation between the A and the C terms that is consistent with (allowed by) the premises but that did not always have to hold. On these trials, we hoped subjects would respond with the first conclusion they considered. On the remaining trials, we asked subjects to give a logically necessary conclusion—one that always had to be true given the premises. In a models-account framework, these necessary trials would require the generation of multiple models to check tentative conclusions.

Models accounts that explain individual differences in terms of the number of models generated would predict that good and poor reasoners who were equally accurate when asked for a relation that "could be true" (a possible conclusion) would differ on problems that asked for a conclusion that "must be true" (a necessary one). Good reasoners would thus honor the could-be-true/must-be-true distinction and would generate additional models when the latter type of answer was required (as evidenced by both higher accuracy and longer latencies). In contrast, poor reasoners should be relatively insensitive to the distinction and in both cases should tend to answer with their tentative conclusion. Poor reasoners' latencies should tend not to differ as a function of conclusion type asked for by the problem.

Finally, models accounts predict group differences in both accuracy and in latency (good reasoners taking longer) for those syllogisms that fail to yield a logically necessary conclusion (because these always require generation of two or more models on any account within this class) relative to those syllogisms that do have a logically necessary conclusion. Of the 16 syllogisms presented to subjects, 6 yielded a necessary conclusion and 10 did not.

Rules accounts as a class make no specific predictions about errors or latencies. Further, this class makes no specific predictions about individual differences (although specific models in this class would of course make specific predictions). However, the lack of a difference in either measure, as a function of the type of conclusion asked for (necessary or possible), would be consistent with these accounts.

Table 1
Mean Proportion Error by Group, Conclusion Type, and Problem Type: Experiment 1

	Conclusion type			
	Necessary		Possible	
Group	SF	NF	SF	NF
Good	.36	.13	.10	.09
Poor	.43	.48	.11	.08

Note. NF = nothing follows. SF = something follows.

Method

Subjects. Twenty-four undergraduates at the University of Pennsylvania participated. They were recruited as follows: A pretest, consisting of eight syllogisms (of the form A–B, C–B), was handed out in lecture courses in psychology with enrollments of 100 or more students. Approximately 250 were returned. We selected for further study students who (a) had no formal training in logic, (b) had scores in the top third (6–8 correct) or the bottom third (0–2 correct) of the sample, and (c) indicated their willingness to participate. These subjects, henceforth designated *good* and *poor* reasoners, were paid $3.50 for a 1-hour session. Data from four poor reasoners and one good reasoner were lost due to computer or experimental error. These subjects were replaced from the same pool, leaving 12 subjects in each group.

Materials. Subjects received two pages of written instructions and a booklet in which to make notes while solving the syllogisms. Syllogisms were presented on a Commodore PET microcomputer that recorded reaction times. Syllogisms were all of the form A–B, B–C. Each problem referred to sets of toy blocks. The A term described the color of a set, the B term the markings of a set (striped, checked, etc.), and the C term described the material out of which the blocks were made.

Procedure. The experimenter explained the task, highlighting the distinction between conclusions that could be true (i.e., that were allowed by the premises) and ones that must be true (i.e., that were logically necessary consequences of the premises). Conclusions were described as statements relating the A to the C term or the C to the A term, and examples were provided. The meanings of various premises were reviewed, with particular emphasis on the quantifier *some*, which functions existentially. The experimenter also demonstrated the use of the PET keyboard.

After four practice trials, subjects worked alone and at their own pace. They were presented with each of the 16 syllogisms from the first figure twice, once being asked for a conclusion that could be true, and once being asked for one that must be true. The two repetitions were separated by 16 other trials. Half of the syllogisms were presented such that a possible (could be true) answer was asked for first. Aside from these requirements, order of presentation was random and counterbalanced across subjects.

Results and Discussion

Unless otherwise noted, the analyses to be reported were three-way mixed analyses of variance (ANOVAs), with group (good vs. poor reasoners), conclusion type (necessary vs. possible), and problem type (ones that had a logically necessary conclusion, or "something follows" [SF] problems vs. ones that did not, or "nothing follows" [NF] problems), as factors, with repeated measures on the last two.

Results from such an ANOVA on percentage error are presented in Table 1. As predicted by a models account, the groups did

Table 2
Antilog of Mean Log RT by Group, Conclusion Type, and Problem Type: Experiment 1

	Conclusion type			
	Necessary		Possible	
Group	NF	SF	NF	SF
Good	24.13	40.40	27.82	32.83
Poor	28.53	41.54	35.08	35.37

Note. NF = nothing follows. SF = something follows.

not differ in accuracy when giving possible conclusions (of either the NF or the SF type) but did differ when giving necessary conclusions. This interaction, between group and conclusion type, was significant, $F(1, 22) = 7.79$; $p < .02$, and the group difference on the necessary conclusion problems is confirmed by a Tukey test ($p < .01$).

Within necessary problems, the groups differed largely on NF problems, a finding consistent with models accounts, because NF problems require the generation of at least two models in order to find a contradiction. This result is demonstrated in the significant three-way interaction for Group × Conclusion Type × Problem Type, overall $F(1, 22) = 7.66$, $p < .02$. A Tukey test confirms the interaction between groups and problem type (NF vs. SF) within the necessary condition at the .01 level.[2]

An alternative explanation of these results is that the groups differed only on necessary problems because possible problems were too easy and were subject to ceiling effects. Essentially, this argument holds that the ceiling effects make measurement in the possible conclusion condition unreliable. To examine this, we calculated reliability coefficients (Cronbach's alpha) for the two conditions separately. They were .81 for the necessary condition and .69 for the possible condition. These values are not highly discrepant, and further, .69 is not particularly low. Hence, this alternative explanation is rendered less plausible.

We also analyzed log RT data but found no significant group differences in latency. Table 2 presents antilogs of mean reaction time (RT), presented by group, problem type, and conclusion type. Note that good reasoners did not spend reliably more time than poor reasoners. Another puzzling aspect of these data is that possible conclusions seem to have taken as long as necessary conclusions. One possible explanation is that subjects initially ignored the could-be-true/must-be-true distinction. In fact, about 25% of the time subjects responded with a logically necessary conclusion when merely asked for a possible one.(The two groups did not differ in this tendency.) This might imply in turn that the task of giving a possible conclusion did not get at the early stages of deduction as we had initially assumed.

The evidence that differences in generating multiple models is at least partially responsible for group differences is mixed. First, the groups differed in accuracy only for necessary conclusions, a central prediction of models accounts. The lack of a group difference on possible problems suggests that the problem goes beyond differences in comprehension of the task. Second, the group difference occurred in particular on NF problems, exactly the ones where failure to generate additional models would lead to error. Rules accounts have a harder time accounting for these findings, unless they postulate that good reasoners' competence lies only in a specific set of rules that all lead to an NF conclusion. The lack of a group difference in reaction time does not support a models account unless one wants to argue that good reasoners are faster at generating representations.

To uncover more about the way subjects were approaching the task, we recalled most of our subjects (10 from each group) approximately 3 weeks after the initial session. We asked them this time to think aloud as they worked. Although this technique has the difficulty that the verbal reports might distort or otherwise fail to capture subjects' thinking, it can serve as a useful ancillary source of evidence.

Eight of the original syllogisms (with changed content) were used. As before, each syllogism appeared twice (although this time the replications had different content). On one repetition, the subject was asked for an answer that must be true; on the other, an answer that could be true. Instructions to subjects regarding thinking aloud were adopted from Perkins (1981).

We found some evidence for the existence of some deduction rules, apparently spontaneously discovered by some of our subjects, as the following two examples illustrate:

> *Subject MK*
> [Reads] "Some orange books are not philosophy books. Some philosophy books are not long. It must be that" . . . nothing follows, again. Oh, I say that just out of, I thought about it a lot when I was doing the computer thing [presumably he refers to the initial session] and I realized that, when there's a *some* and a *some*, nothing ever follows."
>
> *Subject JuS*
> [Reads] "Some purple books are history books. Some history books are thick. It must be that." Well, both of these just have *some* in them. Which means that, nothing follows from it.

These excerpts provide direct evidence of the existence of deduction rules, in particular ones that lead to a NF conclusion. Good reasoners ($M = .90$) were more likely to announce such rules than were poor reasoners ($M = .10$), $t(18) = 2.26$, $p < .025$, one-tailed. Almost all of these announcements were of the "two somes" rule, presented above.

Was good reasoners' superior performance to be attributed to the possession of a few rules, such as the two-somes rule or the two-negatives rule, mentioned earlier? To examine this question we reanalyzed some of the data from the initial session. We considered in this analysis only the 10 NF problems. We subdivided these into three groups: (a) problems where the two-somes rule applies (four SOMES problems); (b) those where the two-negatives rule applies (four NEG problems, one of them in the first group

[2] To guard against the possibility that the interaction is the result of performance on a few individual problems, we tested across problems as well. The mean proportion correct for poor reasoners was subtracted from the mean proportion correct for good reasoners for each individual problem. *T* tests were run to compare these differences for NF and SF problems. This comparison was reliable in the necessary condition, $t(14) = 2.68$, $p < .01$, one tailed, but not in the possible condition, $t(14) = -1.03$, $p > .10$, one-tailed. The three-way interaction was tested by subtracting the mean group difference for the possible condition from the mean group difference in the necessary condition for each problem. A *t* test for NF versus SF problems was significant, $t(14) = 3.19$, $p < .005$, one-tailed.

also); and (c) those where neither rule applies (three OTHER problems). No group difference as a function of this distinction was found in either accuracy or reaction time.

Our protocol data also suggest another difference between good and poor reasoners: a differential tendency to misinterpret the quantifier *some* to mean *not all.* Poor reasoners evidenced greater misunderstanding of the premises, indicated by statements of the form, "Well, if some blue books are not sociology books, some others are." Poor reasoners made an average of 1.2 such statements, good reasoners, .50, $t(18) = 2.97$, $p < .005$, one-tailed. This finding indicates a problem in the method, one we took pains to correct in the following experiment. Therefore, in the second experiment, all subjects were given training designed to teach the specific meaning of each premise, until a specific criterion was met.

Experiment 2: Initial Versus Final Answers

In this study we took a different approach to examining the early stages of reasoning. Subjects were asked to state their initial impression ("gut reaction") of the correct conclusion to a syllogism. Three features of the procedure encouraged subjects to provide a true initial impression: (a) they had to respond within 20 s, which was half the time subjects typically spent on the same problems in Experiment 1; (b) subjects were awarded points for responding as quickly as possible; and (c) subjects were told ahead of time that after giving the initial impression, they would have the opportunity to give their best considered response to the problem, with unlimited time allowed.[3]

Method

Subjects. Sixteen good and 15 poor reasoners, chosen by the methods previously described, participated. They were paid $3.50 per hour; most took about 2 hours to complete the session. In addition, seven other students, all of whom had had some experience studying logic (graduate students in psychology and in computer science) participated. This group was designated the "expert" reasoners for expository purposes, although their expertise was not always extensive, and in any event was with logic in general rather than syllogisms in particular. Experts also differed from other subjects in terms of age, years of education, and probably ability, so expert–novice comparisons must be interpreted with care.

Materials. Subjects worked at a PET computer, which presented all stimuli and recorded responses and reaction times. Subjects were shown how to operate the keyboard at the start of the session; after this they proceeded at their own pace.

Procedure. In the first part of the experiment subjects had called to their attention the meaning of the quantifiers (*all, some, no*) that are used in categorical syllogisms. Definitions of single premises were presented, followed by a set of questions for the subject to answer to assess comprehension. If a subject missed any of the questions, the example was later repeated. Subjects repeated examples until all questions had been correctly answered.

Following this phase, subjects were given 10 practice trials in which they learned a numeric code to be used in answering problems. Each of nine possible responses to a syllogism was displayed, and the subject was given practice finding the number corresponding to each alternative. Next, the subject was given practice at responding under a deadline. A trial began when the subject indicated readiness by pressing the space bar. A syllogism appeared at the top of the screen, followed by the nine possible responses, each preceded by its numeric code. At the top right-hand corner of the screen, the digits 20 appeared, but changed after 1 second to 19, then to 18 after another second, etc. Subjects had to respond before this clock reached 0. Points were awarded for each second left on the clock. (Points had no external value, but appeared to be motivating.) If the subject failed to respond in time, the message, "Time's up!" appeared on the screen, five points were subtracted from the running total of points, and subjects were not allowed to respond.

The experiment proper was quite similar to this last set of practice trials. Sixteen syllogisms (those used in Experiment 1) were presented in a random order, counterbalanced across subjects. A syllogism was presented with the nine alternatives below, and the clock ran from 20 s to 0 s in the upper right-hand corner of the screen. Before the clock ran out, subjects were to give their initial impression of the answer, as quickly as possible. Immediately after the subject responded, the same problem reappeared, again with the nine alternatives but with no clock. This time, the subject was asked to give a final answer to the syllogism and was encouraged to take as much time as needed.

Results

Because the central questions being addressed involve group differences, we opted to analyze the data using one-way ANOVAs. For each measure taken, we first analyzed the overall performance of the three groups and then examined performance on NF problems only. To assess whether group differences obtained selectively on NF problems, we performed the ANOVA on the measure of mean performance on NF minus mean performance on SF problems. (When this was significant, we tested across problems as well.) Throughout some of the tables presented below, group means of performance on SF problems are included for completeness, although no analyses were carried out on these data due to their redundancy.

Premise training. The three groups of subjects required a different number of presentations of premise examples in the initial phase of the experiment. (Recall that until a subject answered all questions about an example correctly, the example was repeated.) The mean number of presentations, by group, was 5.29 for expert, 5.15 for good, and 8.60 for poor reasoners (minimum = 4), $F(2, 35) = 3.95$, $p < .05$. Poor reasoners differ from the other groups at the .05 level by a Tukey test; no other differences are reliable.

Accuracy. We first examine errors made in the initial impression condition. Table 3 presents the results of these analyses, and shows that differences in error rates among the three groups were apparent even in this tentative conclusions. Group differences obtain especially on NF problems as the significant interaction between group and problem type shows. Tukey tests showed that all group differences are reliable. However, as Table 3 shows, there were possible scaling differences in the mean performance on the two types of problems (NF vs. SF) between good and poor reasoners.[4] Therefore, to test the interaction properly, we

[3] Johnson-Laird & Bara (1984) have independently carried out a similar procedure, although not to examine individual differences.

[4] *Scale* refers to the function relating a measure of ability to the ability itself. We assume that measures are monotonically related to ability. Scaling problems arise when two measures are suspected to be differentially sensitive to ability in the range of interest. If one test is more sensitive to ability differences, then group differences that are larger on one test could be wholly due to the superior discriminating power of the first test.

Table 3
Analyses of Error Data: Experiment 2

Analysis	Group means: Expert	Good	Poor	Overall F[a]	Tukey tests
		Initial-impression condition (proportion error)[b]			
Overall	.32	.56	.74	20.898*	p–g, p–e, g–e .01
NF problems	.24	.64	.89	22.431*	p–g, p–e, g–e .01
SF problems	.45	.42	.50	—	—
NF − SF	−.21	.23	.63	8.248*	p–g, p–e, g–e .01
		Final-answer condition (proportion error)[c]			
Overall	.13	.36	.65	32.208*	p–g, p–e, g–e .01
NF problems	.04	.40	.80	31.369*	p–g, p–e, g–e .01
SF problems	.29	.29	.40	—	—
NF − SF	−.24	.11	.40	16.196*	p–e, g–e .01, p–g .05

Note. NF = nothing follows; SF = something follows. p = poor; g = good; e = expert.
[a] $df = 2, 35$.
[b] jacknife p–g = −.32; $t(30) = -1.905$, $p < .10$.
[c] jacknife p–g = −.49, $t(30) = -3.411$, $p < .01$.
* $p < .01$.

examined the correlation between group membership and the difference in z scores of the percentage error on NF versus SF problems. To test this correlation for significance we used a jacknife method (see Footnote 4). The correlation was only marginally significant ($p < .10$).

In the final-answer condition the three groups also differ in overall error rates and in errors on NF problems alone (see Table 3). The real test of interest is again a test of the interaction between group membership and the difference in performance on NF versus SF problems. Again, to rule out scaling difficulties, we computed jacknife correlations on z score differences for good and poor reasoners only. This correlation was significant ($r = -.49$), $t(30) = -3.41$, $p < .01$.[5]

In a related analysis, we tabulated the number of times subjects changed their answers between giving an initial impression and a final answer and found that the total number of changes was not reliably predicted by group membership. Expert, good, and poor reasoners made totals of 3.71, 6.19, and 6.27 changes, respectively, $F(2, 35) = 2.31$, $p > .05$. Changes can be classified into two types: those that correct a previously incorrect answer (call these correcting changes) and those that do not. The proportion of changes that are correcting for expert, good, and poor reasoners is .73, .60, and .31, $F(2, 35) = 8.87$, $p < .01$; poor reasoners differ significantly from the other groups by Tukey tests. The mean *number* of correcting changes, by groups, is 3.1, 3.5, and 1.8 for expert, good, and poor reasoners, $F(2, 35) = 4.16$, $p < .05$. Good and poor reasoners differ at the .05 level; expert and poor reasoners, at the .10 level by Tukey tests.

Reaction time. The measure used was again log latency. Analyses of these data in the initial-impression condition are presented in Table 4. The mean time taken by experts was significantly shorter than that taken by the other two groups. This finding holds in particular for NF problems, relative to SF problems.[6]

In the final-answer condition, only good and expert reasoners differed significantly in log latency, and the difference occurred especially on NF problems.[7] Recall that in the final-answer condition subjects could take as much time as they wished. Of interest, then, is the comparison of time taken in this condition relative to the time taken when a deadline was imposed. Table 4 shows the mean antilog of the log ratio of time spent in the final-answer relative to the initial-impression condition. Good reasoners spent proportionately more time in generating a final answer than did poor reasoners. Experts did not differ from poor

However, if it is reasonable to assume that one measure is a linear transform of the other (to a first approximation under the null hypothesis) then one can convert all measures for the two tasks to z scores. Under the assumption of no real differences except for scale, z(Measure 1) = z(Measure 2), so z(Measure 1) − z(Measure 2) = 0, over all subjects. To test for a group difference on the two tasks with the factor of scale removed, one computes a correlation between group membership and the difference between z scores. To assess the standard error of the correlation, one employs the jacknife method (see Mosteller & Tukey, 1977), calculating N correlations, deleting one subject at a time for each calculation. The mean over the N correlations is the jacknife correlation, and the standard error allows a test of statistical significance. For more on when such procedures are warranted, see Baron & Treiman (1980).

[5] This interaction was tested across problems by three t tests, one for each pair of groups. For each problem, we subtracted the mean proportion error of one group from the mean proportion error of the other. All the one-tailed t tests for NF versus SF problems were reliable: for good versus poor reasoners, $t[14] = 5.02$, $p < .001$; for expert versus good reasoners, $t[14] = 3.48$, $p < .005$; and for expert versus poor reasoners, $t[14] = 7.45$, $p < .001$. This suggests that the interaction is not an artifact of performance on a few individual problems.

[6] One-tailed t tests across problems were: for expert versus good reasoners, $t[14] = -3.81$, $p < .005$; for expert versus good reasoners, $t[14] = -2.51$, $p < .025$.

[7] A one-tailed test across problems confirmed this finding for expert versus good reasoners, $t(14) = 1.79$, $p < .05$.

Table 4
Analyses of RT Data (Antilog of Mean Logs): Experiment 2

Analysis	Group means			Overall F[a]	Tukey tests
	Expert	Good	Poor		
			Initial impression condition (seconds)		
Overall	6.76	10.30	11.16	8.658†	p–e, g–e .01
NF problems	6.11	10.88	11.43	11.441†	p–e, g–e .01
SF problems	7.73	9.19	10.64	—	—
NF − SF	.79	1.18	1.07	9.500†	p–e, g–e .01
			Final answer condition (seconds)		
Overall	19.81	40.53	28.71	4.667**	g–e .01
NF problems	20.18	46.20	30.38	5.300†	g–e .01
SF problems	18.67	29.85	24.86	—	—
NF − SF	1.08	1.55	1.22	3.042*	—
			Final/initial		
Overall	2.93	3.94	2.57	3.676**	p–g .05
NF problems	3.31	4.24	2.66	3.636**	p–g .05
SF problems	2.42	3.25	2.34	—	—
NF − SF	1.37	1.31	1.14	0.840	—

Note. NF = nothing follows; SF = something follows. p = poor; g = good; e = expert.
[a] $df = 2, 35$.
* $p < .10$. ** $p < .05$. † $p < .01$.

reasoners. Although this finding holds for NF problems alone, the interaction between group and the difference between NF and SF problems is not reliable.

Experts' higher accuracy and shorter latencies on NF problems suggest the use of deduction rules. Did experts rely only on the SOMES and NEG rules, those most frequently discovered by novices and mentioned by logic textbooks? Or was experts' superiority a more general phenomenon? To answer these questions, we again restricted attention to the 10 NF problems. These were subdivided into three types as before, SOMES, NEG, and OTHER problems.

For the accuracy measure, a significant Group × Problem Type (SOMES, NEG, OTHER) × Answer (initial impression vs. final answer) interaction was obtained, $F(2, 70) = 2.65$, $p < .05$. The means for this interaction are presented in Table 5. Experts differed in all instances from the other groups at the .05 level by a Tukey test. Poor reasoners differed significantly from good reasoners at the .05 level in all cases except on OTHER problems in the initial impression condition. Poor reasoners showed no reliable improvement from the initial-impression to the final-answer condition. Good reasoners improved significantly only on NEG problems; experts showed significant improvement only on OTHER problems.

Table 5
Performance on Nothing-Follows Problems by Rules: Experiment 2 (Proportion Error)

Group	Condition					
	Initial impression			Final answer		
	SOMES	NEG	OTHER	SOMES	NEG	OTHER
Expert	.07	.21	.43	.00	.07	.05
Good	.56	.55	.81	.34	.23	.65
Poor	.85	.72	.98	.75	.68	.98

Note. SOMES = two-somes rule. NEG = two-negatives rule. OTHER = neither rule applies.

A similar analysis was carried out on log latency data. A significant interaction between group and problem type was obtained, $F(4, 70) = 5.53$, $p < .01$. Table 6 presents the relevant means, expressed as antilogs of mean logs. Experts are reliably faster than the other groups on SOMES problems (Tukey test, .05) and reliably faster than good reasoners on NEG problems. The three groups do not differ on OTHER problems.

These analyses are consistent with the following interpretation: Experts have good command of the SOMES and NEG rules, in particular the former, and apply them quickly and effectively. Good reasoners have some command of these rules, although not as much as do experts. Poor reasoners show little evidence of familiarity with either rule.

At the same time, both good and expert reasoners benefit from additional time—good reasoners benefitting most on NEG and SOMES problems, experts on OTHER problems. This finding suggests that experts' superiority does not reside merely in the possession of and facility with the two familiar rules.

Finally, for the benefit of readers who wish to test their own models of syllogistic reasoning, we present in Tables 7 and 8 accuracy and RT data, respectively, by group for each individual syllogism for both the initial-impression and final-answer conditions.

Discussion

The group differences in accuracy, unaccompanied by latency differences, in the initial-impression condition favor either a rules

Table 6
Performance on Nothing-Follows Problems by Rules: Experiment 2 (Antilog of Mean Log RT)

	SOMES		NEG		OTHER	
Group	Initial	Final	Initial	Final	Initial	Final
Expert	3.67	10.83	5.45	20.81	7.61	19.83
Good	9.78	36.03	11.33	44.81	10.04	30.24
Poor	9.92	19.92	11.46	30.68	10.81	20.03

Note. SOMES = two-somes rule. NEG = two-negatives rule. OTHER = neither rule applies.

or a models account that locates the difficulty in reasoning in the initial model construction. These differences could also be explained by claiming that the three groups differ in mental speed and that, contrary to instructions, better reasoners constructed multiple models before responding. We have no reason to make this assumption but no way to rule it out. The group difference in accuracy in the final-answer condition, taken together with the fact that poor reasoners spent proportionately less time generating final answers, implicates a models account as at least a partial account of reasoning proficiency.

Even when poor reasoners do spend time revising initial impressions, the "dividends," measured in terms of proportion of correcting changes, are significantly lower. This lack of effectiveness may provide a rational basis for poor reasoners' choice to spend less time at the revision task (see Baron, Badgio, & Gaskins, in press). Alternatively, poor reasoners may be less self-critical in their search for alternative models, that is, less prone to seek models that are inconsistent with their initial conclusion.

The data for expert reasoners is strikingly different from the other two novice groups. Experts' latencies are everywhere shorter, consistent with the notion that they rely on deduction rules. The analysis of performance on individual NF problems, however, argues against the idea that the sole source of experts' superiority is the possession and efficacy of application of the SOMES and the NEG rules.

General Discussion

We summarize the findings of the two studies within the framework of the questions posed initially. First, there is evidence that people make some use of deduction rules when solving syllogisms. The protocol data of Experiment 1 reveal that novice reasoners sometimes spontaneously articulate such rules after only brief practice at the task. The RT data of Experiment 2 give evidence that experts regularly rely on such rules. Experts are always faster than good reasoners and often faster than poor reasoners.

There is also evidence that novice groups at least do engage in the generation of multiple models. Good and poor reasoners differed most on NF problems, where failure to generate more than one model would lead to error. Good reasoners honored the necessary/possible distinction in Experiment 1, whereas poor reasoners tended not to. Good reasoners chose to spend more time and improved their accuracy more in the final-answer condition of Experiment 2.

The findings suggest that the tendency to generate multiple models explains some part of ability differences. At the same time, group differences also emerge in the formation of a tentative conclusion, as evidenced by the group difference in accuracy in the initial-impression condition of Experiment 2. This finding implicates either a rules account or a models account that locates ability differences in the mechanism that generates models. Such differences may be even more apparent in other syllogisms (using different figures, or ordering of terms, e.g., B–A, B–C). Johnson-Laird and Bara (1984), for example, suggest that the A–B, B–C order of premises is the easiest one to process. We suspect, therefore, that the group differences found here would be even more pronounced in other, more difficult syllogisms.

Regarding the performance of experts, a rules account will work very nicely. This may be unsurprising, because inference

Table 7
Proportion Correct by Group, Syllogism, and Condition: Experiment 2

		Expert		Good		Poor	
First premise	Second premise	I	F	I	F	I	F
All A are B	All B are C	.86	1.00	.94	.94	.73	.93
	Some B are C	.71	1.00	.19	.38	.07	.00
	No B are C	.86	1.00	1.00	1.00	.87	.93
	Some B are not C	.43	1.00	.13	.31	.00	.00
Some A are B	All B are C	.86	1.00	.88	.94	.73	1.00
	Some B are C	1.00	1.00	.38	.56	.07	.20
	No B are C	.71	1.00	.63	.94	.47	.67
	Some B are not C	.86	1.00	.38	.56	.13	.27
No A are B	All B are C	.00	.29	.00	.13	.00	.00
	Some B are C	.00	.00	.06	.31	.00	.07
	No B are C	.71	1.00	.56	.94	.07	.33
	Some B are not C	.86	1.00	.44	.81	.20	.53
Some A are not B	All B are C	.57	.86	.25	.38	.07	.07
	Some B are C	.86	1.00	.44	.75	.13	.20
	No B are C	.57	.71	.25	.56	.20	.07
	Some B are not C	1.00	1.00	.56	.75	.27	.33

Note. I = Initial-impression condition. F = Final-answer condition.

Table 8
Antilog of Mean Log RT (in Seconds) by Group, Syllogism, and Condition: Experiment 2

		Expert		Good		Poor	
First premise	Second premise	I	F	I	F	I	F
All A are B	All B are C	5.39	8.90	6.13	9.97	7.87	10.14
	Some B are C	6.48	19.40	10.01	31.74	10.61	17.72
	No B are C	7.66	13.49	8.56	17.50	10.63	19.43
	Some B are not C	7.19	16.45	8.89	26.24	11.28	19.65
Some A are B	All B are C	6.41	13.34	8.65	17.34	9.35	14.74
	Some B are C	2.61	8.66	8.68	28.20	8.24	16.07
	No B are C	10.00	16.30	10.69	26.41	11.76	24.97
	Some B are not C	4.14	10.23	8.72	30.08	8.89	19.30
No A are B	All B are C	6.31	28.00	9.21	41.42	9.38	29.31
	Some B are C	7.81	14.33	9.37	44.61	12.47	26.21
	No B are C	5.37	26.48	9.85	30.80	9.88	32.57
	Some B are not C	6.85	19.42	12.82	44.75	13.10	36.61
Some A are not B	All B are C	9.46	24.45	11.38	33.19	10.56	23.09
	Some B are C	4.08	10.77	10.58	44.94	12.67	25.75
	No B are C	5.86	25.28	11.42	66.21	12.72	37.71
	Some B are not C	4.07	14.43	11.43	44.18	10.46	19.69

Note. I = Initial-impression condition. F = Final-answer condition.

rules are taught in logic courses (that some of our experts had taken). Experts were in all cases faster than the other two groups, consistent with the view that they have developed or learned shortcut rules of inference. This is not a vacuous explanation; it might have turned out that learning such rules in a formal setting did not result in heavy reliance on the rules later on.

We note here that one good reasoner in Experiment 1 mentioned his own discovery of the SOMES rule during the course of his participation. It would be informative to chart the course of development of such rules or principles, to provide insight into the question of how reasoning ability develops naturally (as opposed to when tutored).

Further work is also needed to address the following questions: How do rules, once acquired, come to be effectively and efficiently used? When are rules used, and when are models used? When models are used, what governs the order of generation and the number of models generated? In what ways are the rules and/or models used in syllogistic reasoning applicable to other types of reasoning, both formal and everyday? Answers to such questions will address the original issue of what syllogistic reasoning reveals about thinking and intelligence in general.

References

Baron, J., Badgio, P., & Gaskins, I. W. (in press). Definition and improvement of cognitive style: A normative approach. In R. J. Sternberg (Ed.), *Advances in the study of human intelligence* (Vol. 3). Hillsdale, NJ: Erlbaum.

Baron, J., & Treiman, R. (1980). Some problems in the study of differences in cognitive processes. *Memory and Cognition, 8,* 313–321.

Begg, I., & Denny, J. P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology, 81,* 351–354.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85,* 1–21.

Braine, M. D. S., & Rumain, B. (1983). Logical reasoning. In J. Flavell & E. Markman (Eds.), *Handbook of child psychology: Vol. 3. Cognitive development* (4th ed.; pp. 263–340). New York: Wiley.

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

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect reexamined. *Journal of Experimental Psychology, 58,* 220–226.

Dickstein, L. S. (1975). Effects of instruction and premise order on errors in syllogistic reasoning. *Journal of Experimental Psychology: Human Learning and Memory, 1,* 376–384.

Dickstein, L. S. (1976). Differential difficulty of categorical syllogisms. *Bulletin of the Psychonomic Society, 8,* 330–332.

Dickstein, L. S. (1978). The effect of figure on syllogistic reasoning. *Memory and Cognition, 6,* 76–83.

Dickstein, L. S. (1981a). The meaning of conversion in syllogistic reasoning. *Bulletin of the Psychonomic Society, 18,* 135–138.

Dickstein, L. S. (1981b). Conversion and possibility in syllogistic reasoning. *Bulletin of the Psychonomic Society, 18,* 229–232.

Erickson, J. R. (1974). A set analysis theory of behavior in a formal syllogistic reasoning task. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 305–330). Potomac, MD: Erlbaum.

Erickson, J. R. (1978). Research on syllogistic reasoning. In R. Revlin & R. E. Mayer (Eds.), *Human reasoning* (pp. 39–50). Washington, DC: V. H. Winston.

Fisher, D. L. (1981). A three-factor model of syllogistic reasoning: The study of isolable stages. *Memory and Cognition, 9,* 496–514.

Frase, L. T. (1968). Associative factors in syllogistic reasoning. *Journal of Experimental Psychology, 76,* 407–412.

Guilford, J. P. (1959). Three faces of intellect. *American Psychologist, 14,* 469–479.

Guyote, M. J., & Sternberg, R. J. (1981). A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology, 13,* 461–525.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69,* 366–378.

Johnson-Laird, P. N. (1982). Ninth Bartlett Memorial Lecture. Thinking as a skill. *Quarterly Journal of Experimental Psychology, 34A,* 1–29.

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

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 17,* 1–61.

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

Mosteller, F., & Tukey, J. (1977). *Data analysis and regression.* Reading, MA: Addison-Wesley.

Osherson, D. N. (1974). *Logical abilities in children: Vol. 2. Logical inference: Underlying operations.* Potomac, MD: Erlbaum.

Osherson, D. (1975). Logic and models of logical thinking. In R. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 81–92). Hillsdale, NJ: Erlbaum.

Perkins, D. N. (1981). *The mind's best work.* Cambridge, MA: Harvard University Press.

Revlin, R., & Leirer, V. O. (1978). The effect of personal biases on syllogistic reasoning: Rational decisions from personalized representations. In R. Revlin & R. E. Mayer (Eds.) *Human reasoning* (pp. 51–81). Washington, DC: V. H. Winston.

Revlis, R. (1975a). Syllogistic reasoning: Logical decisions from a complex data base. In R. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 93–133). Hillsdale, NJ: Erlbaum.

Revlis, R. (1975b). Two models of syllogistic reasoning. *Journal of Verbal Learning and Verbal Behavior, 14,* 180–195.

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90,* 38–71.

Roberge, J. J. (1970). A reexamination of the interpretations of errors in formal syllogistic reasoning. *Psychonomic Science, 19,* 331–333.

Sternberg, R. J., & Turner, M. E. (1981). Components of syllogistic reasoning. *Acta Psychologica, 47,* 245–265.

Thurstone, L. L. (1938). *Psychometric Monograph* (No. 1 ed.). Chicago: University of Chicago Press.

Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology, 16,* 83.

Woodworth, R. J., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18,* 451–460.

Received March 25, 1985
Revision received July 10, 1985 ■