THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 2004, *57A* (1), 33–60

Individual differences in deductive reasoning

Stephen E. Newstead, Simon J. Handley, Clare Harley, Helen Wright, and Daniel Farrelly

University of Plymouth, Plymouth, UK

Three studies are reported, which examined individual differences in deductive reasoning as a function of intellectual ability and thinking style. Intellectual ability was a good predictor of logical performance on syllogisms, especially where there was a conflict between logic and believability. However, in the first two experiments there was no link between ability and performance on indicative selection tasks, in sharp contrast to previous research. This correlation did, however, return in the final study. Our data are consistent with the claim that the correlation with logical accuracy on abstract selection tasks is found primarily with participants of relatively high ability. At lower levels, pragmatically cued responses are given but those of slightly higher ability divorce the rule from the scenario and respond consistently (though incorrectly) across problems. Self-report questionnaires were generally poor predictors of performance, but a measure of the ability to generate alternative representations proved an excellent predictor. These results are consistent with a mental models approach to reasoning and also have implications for the debate about human rationality.

Until recently there were two quite distinct literatures in the psychology of reasoning. The first of these is in the psychometric tradition in which the aim has been to measure individual differences in thinking. The best known measures are those involving intelligence or ability, but measures have also been developed for other aspects of thinking such as cognitive style. The second tradition is that of cognitive psychology, where the aim is to discover the underlying cognitive processes involved in thinking. The assumption made by cognitive psychologists is that these processes are largely universal, and hence they have tended to downplay any individual differences.

All this has changed in the last few years. Researchers such as Stanovich and West (e.g., 2000) and Klaczynski and his colleagues (e.g., Klaczynski, Gordon, & Fauth, 1996) have pointed out that the two research traditions are complementary. In particular, individual differences have the potential to throw light on the existence of different ways of thinking, on the correct normative model, and on the rationality debate. These arguments have been well

Correspondence should be addressed to Stephen E. Newstead, Centre for Thinking and Language, Department of Psychology, University of Plymouth, Drake Circus, Plymouth, PL4 8AA, UK. Email: snewstead@plym.ac.uk

This research was supported by grant number R000222763 from the Economic and Social Research Council, UK. The authors would like to thank Keith Stanovich for his comments on an earlier version of this paper.

http://www.tandf.co.uk/journals/pp/02724987.html DOI:10.1080/02724980343000116

articulated by Stanovich and West (2000). The aim of the present research is to throw further light on these issues.

It seems widely agreed that there are two quite different types of thinking though there is less agreement about what they should be called. Stanovich and West (2000) list a whole range of dichotomies of thinking variously termed implicit versus explicit, associative versus rule based, experiential versus rational, and many more besides. Stanovich and West (2000) referred to the two systems neutrally as System 1 and System 2, and we use these terms here. System 1 is based on personal experience, is intuitive, and relies on associations, while System 2 is more deliberative, formal, and based on symbol manipulation. It is worth noting, however, that the distinction between the two different systems is not identically drawn by all authors (Newstead, 2000).

Individual differences research may be able to throw much needed light on these two systems. System 2 is closely related to (and may be the same thing as) the abilities measured by traditional intelligence tests. Hence one would expect to observe correlations between intellectual ability scores and reasoning performance that is based on formal rules (in general, logical reasoning). One might also expect that individual difference measures of intuitive thinking would correlate with System 1 thinking, which in many cases is related to reasoning biases. This would provide evidence both for the validity of the distinction between the two systems and for the situations in which they are activated.

Individual differences research can also provide evidence relevant to the rationality debate: the dispute as to whether humans are rational. People make many logical errors when faced with reasoning tasks, and the issue is whether these indicate that humans are inherently irrational or whether they simply show that otherwise rational performance is being disrupted by extraneous factors (misunderstanding the task, memory limitations, and so on). Those who argue in favour of humans being inherently rational often claim that deviations from rational performance can be attributed to the wrong normative model having been applied. For example, Oaksford and Chater (1994; Chater & Oaksford, 1999) have provided an alternative analysis of both Wason's selection task and of syllogistic reasoning in which the most commonly given response is in fact the rational one. These approaches are to some extent undermined by correlations between ability and the standard correct response. If more able people tend to give the standard correct response (but not that predicted by alternative rational analyses) then this suggests that the standard normative analysis is possibly the correct one.

Hence from the perspective of both the rationality debate and the claim for the existence of two systems of thinking, individual differences in deductive reasoning are of considerable importance. Most previous research has focused on the correlations between intellectual ability and two reasoning tasks—the Wason selection task and syllogistic reasoning—and this research is now reviewed. The much smaller literature on intuition and deductive reasoning is then examined.

There is good evidence for the existence of reliable correlations between ability and logically correct performance on the standard, abstract Wason selection task (Dominowski & Dallob, 1991; Klaczynski, 2001; Stanovich & West, 1998a, 1998b; Valentine, 1975). Correlations between ability and correct responses to realistic selection tasks are more variable. It is well known that certain deontic forms of the selection task can facilitate performance. For example, if the rule is "If a person is drinking beer then they must be over 18", then correct logical choices are vastly increased (Griggs & Cox, 1982). It is not immediately obvious

whether one would expect a correlation between performance on these tasks and ability; the correct response is cued by context and experience, and it is possible that ability is unrelated to the extent to which such cues are utilized. Stanovich and West (1998a) found that intellectual ability was a worse predictor of performance on deontic tasks than it was on abstract tasks, but the effects were mostly in the right direction and in some instances achieved significance. Klaczynski (2001) found correlations between ability and performance on deontic tasks, and Dominowski and Dallob (1991) found that these were just as high as those with abstract tasks. Oaksford, Morris, Grainger, and Williams (1996) found that a task designed to suppress working memory (which is closely related to ability) impaired performance on deontic tasks. A reasonable conclusion seems to be that some individuals solve deontic problems using analytic processes, and others use contextual cues, and any correlations with ability stem from the former group (cf. Klaczynski, 2001; Stanovich & West, 2000).

Ability and logical performance also tend to correlate on syllogistic reasoning tasks (Stanovich & West, 1998a; Torrens, Thompson, & Cramer, 1999). In syllogistic reasoning, there is a well known bias, belief bias, whereby people ignore the logic and give the response that corresponds to their prior beliefs (e.g., Evans, Barston, & Pollard, 1983). Sá, West, and Stanovich (1999) found that ability correlated negatively with the tendency to respond in accordance with belief when logic and belief conflicted in syllogistic reasoning tasks. However, Torrens et al. (1999) found no correlation with believability effects.

In addition to shedding light on theories of human reasoning, these findings also give clues as to the nature of human intelligence. A key factor would appear to be the ability to decontextualize information (Stanovich, 1999). Those of higher ability are able to go beyond the content of the information and think in terms of abstract logical structures. They are able to avoid cues such as matching (i.e., choosing the cards named in the rule) on selection tasks and are less likely to be distracted by the believability of the information in syllogisms; they perform particularly well when conflict and belief suggest different responses. Those of lower ability tend to be dominated by cues derived from the content itself, for example the believability of the conclusions.

The relationship between intuition (and other thinking styles) and logical reasoning performance is much less well researched. One problem here is the need for an adequate measure of individual differences in intuition. Our search for such a measure led us to the Rational–Experiential Inventory (REI). The most recent version of this available to us (Pacini & Epstein, 1999) has two main scales: rationality (the tendency to engage in logical thinking) and experientiality (the tendency to rely on intuition and experience). Both of these have subscales corresponding to engagement and ability, though the subscales are more robust on rationality than on experientiality (Handley, Newstead, & Wright, 2000; Pacini & Epstein, 1999). The rational engagement measure is derived from earlier need for cognition scales while the experientiality scales are developed from earlier ones measuring faith in intuition. The term "rationality" is an unfortunate choice since it invites confusion with issues surrounding the wider debate about human rationality. For this reason, we henceforth refer to the REI rationality measure as cognitive motivation.

Of particular interest here are correlations between experientiality and heuristic responses on reasoning tasks. If experientiality is a measure of the tendency to rely on System 1 thinking, then one might expect it to be related to improved performance on deontic selection tasks and to the propensity to respond according to belief in syllogistic reasoning. Experientiality (and

its precursor, faith in intuition) has produced variable but generally positive correlations with biases in responding on a number of statistical reasoning tasks (Epstein, Pacini, Denes-Raj, & Heier, 1996; Klaczynski et al., 1996; Pacini & Epstein, 1999), but few if any studies have examined experientiality and logical reasoning.

The cognitive motivation measure (Epstein et al.'s, 1996, rationality measure) is an indication of how much people enjoy engaging in logical tasks; it is arguably more a measure of motivation to engage in problem solving and other thinking tasks rather than a measure of ability to do so or of the type of thinking approach that is taken. There is reasonably good evidence that cognitive motivation correlates with logical performance. The correlations are in general lower than those with intelligence, but nevertheless significant correlations have been found with syllogistic reasoning (Torrens et al., 1999), and with the Wason selection task (Klaczynski, Fauth, & Swanger, 1998). There is evidence that cognitive motivation correlates negatively with biased responding on a variety of statistical tasks (Epstein et al., 1996) but little evidence either way for correlations with deductive reasoning.

To summarize, there is good evidence that ability correlates with logical performance on the abstract selection task and on syllogisms. This is taken as evidence that the correct normative model for performance on these tasks has been used, and that System 2 thinking is a key aspect of logical accuracy on these tasks. There is indirect evidence that the tendency to rely on System 1 processing is related both to performance on deontic selection tasks and to the tendency to respond according to belief on syllogisms. However, this evidence is little more than circumstantial at the moment. A major aim of the present series of studies was to investigate further the relationship between measures of ability and thinking style, and reasoning performance.

EXPERIMENT 1

Method

Participants

The 98 participants, 80 female and 18 male, were all students at the University of Plymouth who were either paid or received course credit for taking part.

Materials

The 40-item version of the REI (Pacini & Epstein, 1999) was used to measure thinking style. This is a self-report instrument designed to measure cognitive motivation (e.g., "I have a logical mind") and experientiality (e.g., "I believe in trusting my hunches"). Both cognitive motivation and experientiality have subscales corresponding to engagement and ability. Half of the items are positive in direction, and half are negative. The response format is a 5-point scale whose end-points are labelled "definitely not true of myself" and "definitely true of myself".

Part 1 of the AH5 intelligence test (Heim, 1968) was used to assess ability. This test is designed for adults, including students and research workers, and contains 36 items in the categories "directions", "verbal analogies", "numerical series", and "similar relationships". It is a widely used test of adult intelligence and was the test used by Valentine (1975) in her study of individual differences on the selection task.

Four selection tasks were used, two nondeontic and two deontic. The nondeontic problems were the original letter–number version ("If a card has an A on its letter side then it has 8 on its number side") and a destination version ("If Glasgow is on one side of the ticket, then train is on the other side of the ticket"). The instructions asked participants to indicate which of the four cards (e.g., A, K, 8, and 5 in the first example) needed to be turned over to test whether the rule was true or false. For ease of reference, we henceforth refer to these nondeontic problems as indicative selection tasks. The two deontic tasks were an anglicized version of the Sears problem in which Sears became Debenhams ("Any sale over £30 must be approved by the manager, Mr Jones") and the drinking age problem ("If a person is drinking beer then the person must be over 18 years of age"). The instructions asked participants to choose the cards they thought needed to be turned over in order to test whether the rule was being violated. The four cards they had to choose from were in each case instances of the *p*, *not-p*, *q,* and *not-q* cards.

Procedure

Participants were tested in groups of between 4 and 20. The REI was presented first for all participants, followed by the timed AH5 (which took 20 min), followed by the four selection tasks in the order letter–number, destination, Sears, and drinking age. Demographic information was collected on the students' sex and age.

Results

This study provided interesting information on the psychometric properties of the REI, which has been reported elsewhere (Handley et al., 2000). Suffice it to say that the REI has good internal reliability (all αs over .7), that the cognitive motivation and experientiality factors clearly emerged but did not intercorrelate, and that the ability and engagement subscales were clearly evident on the cognitive motivation measure but not on experientiality.

The responses given to the selection tasks showed the usual facilitation of deontic over indicative problems, with 64% correct (i.e., *p* + *not-q* responses) on deontic tasks and 8% correct on indicative tasks. The responses given are presented, broken down by ability level, in Table 1. In this and subsequent tables the data are presented for the top quartile, the middle two quartiles combined, and the bottom quartile. This was done in order to enable differences between extreme groups to be detected, in that most of our analyses compared the top and bottom quartiles.

The pattern was not what we had expected. On the indicative selection tasks, those of higher ability had been expected to be more likely to give the logically correct response, but there was no difference between the quartile groups. For all groups, the most popular response was *p* + *q*, and there was little discernible difference in the frequency of the various response patterns. However, on the deontic tasks there did seem to be a difference between ability groups in choice of the correct response. On both deontic problems it was the highest ability group who chose the *p* + *not-q* response most frequently. A comparison of the top and bottom quartiles revealed a significant difference on the Debenhams problem, $\chi^2 = 9.21$, $p < .01$, but not on the drinking age problem $\chi^2 = 1.63$, $p > .1$.

An unpredicted difference in ability emerged between those who responded consistently to the indicative tasks and those who responded inconsistently. We compared the AH5 scores of those participants who selected the same card combination for both the letter–number and the destination tasks with the scores of participants who selected different combinations. An independent groups two-tailed *t* test confirmed that the consistent reasoners ($n = 81$) had

TABLE 1
Cards chosen on each selection task as a function of ability in Experiment 1

		p + *not q*		*p* + *q*		*p*		*All*		*Other*	
Task		*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*
Letter–number	Top	2	9	11	52	1	5	4	19	3	14
	Middle	6	11	28	52	3	6	8	15	9	17
	Bottom	1	4	10	44	3	13	3	13	6	26
Destination	Top	2	10	13	62	1	5	3	14	2	10
	Middle	4	8	32	60	2	4	7	13	8	15
	Bottom	1	4	10	43	2	9	4	17	6	26
Debenhams	Top	16	76	–	–	–	–	1	5	4	19
	Middle	25	47	5	9	1	2	8	15	14	26
	Bottom	7	30	4	17	3	13	1	4	8	35
Drinking age	Top	18	86	1	5	–	–	–	–	2	10
	Middle	42	79	1	2	–	–	4	8	6	11
	Bottom	16	70	1	4	–	–	1	4	5	22

Note: Top = top quartile on AH5 ($n = 21$); Middle = middle two quartiles ($n = 54$); Bottom = bottom quartile ($n = 23$).

significantly higher AH5 scores (11.48) than the inconsistent reasoners ($n = 17$, AH5 score = 10.21), $t(96) = 1.99$, $p < .05$. Taken together with the previous findings, this would thus appear to show that the most able participants tended to choose consistent combinations (usually $p + q$) on the indicative tasks and the p + *not q* cards on the deontic tasks.

The measures derived from the REI, experientiality and cognitive motivation, produced little of interest in this (or indeed subsequent) experiments. For ease of presentation, composite tables are presented, which combine the results from all the studies that used these measures (see Tables 2 and 3). The only significant effect to come out of the present study involved experientiality. On indicative tasks, those in the top quartile in experientiality produced fewer correct (i.e., p + *not-q*) responses (2% vs. 15%), while on deontic tasks high experientiality participants produced more correct responses (72% vs. 55%). This interaction was significant in an analysis of variance, $F(1, 41) = 5.34$, $MSE = 0.293$, $p < .05$.

Discussion

The findings with respect to ability are surprising since, as our review highlighted earlier, the usual pattern observed is one in which logical performance on indicative tasks is associated with higher ability. In contrast, the relationship between ability and performance on deontic versions of the task is generally much weaker or absent. In the present study there was no effect at all on indicative tasks but a significant effect on one of the deontic tasks.

There are a number of possible explanations for this failure to replicate. One of these is that the participants in this study were of lower overall ability than those used in previous studies. Experiment 1 used the AH5 test, which was also used by Valentine (1975) in her study. In

TABLE 2
Cards chosen on each selection task as a function of experientiality in Experiments 1, 2, and 3 combined

		p + *not q*		*p* + *q*		*p*		*All*		*Other*	
Task		*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*
Letter-number	Top	7	9	46	62	7	9	5	7	9	12
	Middle	15	9	87	51	25	15	9	5	33	20
	Bottom	7	9	39	51	16	21	6	8	9	12
Destination	Top	9	12	48	65	3	4	5	7	9	12
	Middle	14	9	86	51	29	17	10	6	29	17
	Bottom	7	9	40	52	15	19	5	6	10	13
Debenhams	Top	37	50	11	15	5	7	3	4	18	24
	Middle	66	39	29	17	34	20	6	4	33	20
	Bottom	33	43	15	19	8	10	2	3	19	25
Drinking age	Top	55	74	3	4	4	5	3	4	9	12
	Middle	108	64	11	7	23	14	1	1	25	15
	Bottom	51	67	11	14	5	6	1	1	9	12

Note: Top = top quartile on REI experientiality scale (n = 74); Middle = middle two quartiles (n = 168); Bottom = bottom quartile (n = 77).

TABLE 3
Cards chosen on each selection task as a function of cognitive motivation in Experiments 1, 2, and 3 combined

		p + *not q*		*p* + *q*		*p*		*All*		*Other*	
Task		*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*
Letter-number	Top	8	11	42	57	9	12	6	8	9	12
	Middle	12	7	93	53	33	19	12	7	23	13
	Bottom	9	12	38	51	6	8	2	3	19	26
Destination	Top	10	14	43	58	11	15	4	5	6	8
	Middle	13	8	94	55	27	16	13	8	25	15
	Bottom	7	10	37	51	9	12	3	4	17	23
Debenhams	Top	31	42	15	20	14	19	3	4	11	15
	Middle	70	41	30	17	26	15	6	3	40	23
	Bottom	35	48	10	14	7	10	2	3	19	26
Drinking age	Top	50	68	7	9	8	11	2	3	7	9
	Middle	112	65	13	8	19	11	3	2	25	15
	Bottom	52	71	5	7	5	7	–	–	11	15

Note: Top = top quartile on REI cognitive motivation scale (n = 75); Middle = middle two quartiles (n = 172); Bottom = bottom quartile (n = 73).

Valentine's study, over 40% of the participants were in categories A and B of the norms (i.e., the highest levels); not one of the participants used in Experiment 1 achieved these levels of ability. Indeed, relative to the norms for university students on the test ($n = 879$, AH5 = 14.5, $SD = 4.8$), our participants were of significantly lower ability ($n = 98$, AH5 = 10.8, $SD = 3.21$), $t(975) = 5.37$, $p < .001$. Hence the ability level of the participants may be an important clue to the discrepancy in our findings, and might suggest that the correlation between indicative tasks and ability is found primarily with higher ability people.

Another possible explanation for the present results is that the measures we used were unreliable. The scales on the REI emerged as highly reliable but the reliability of other measures such as selection task performance was not tested. In order to obtain further information on both replicability and reliability, we decided to bring back as many of the 98 participants as we could to repeat and extend the testing we had carried out. The REI and the same selection tasks were used, and we also included a further measure of ability: the AH4 Parts 1 and 2 (Heim, 1967). This is a general test of adult intelligence consisting of 65 items in two parts, with the first part designed to assess verbal and numerical ability, and the second part visuo-spatial ability.

We were able to re-investigate 44 participants, 36 female and 8 male, from the original 98. The REI had good test–retest reliability (.78 or greater on the main scales). The AH4 correlated fairly well (.56) with the AH5 used in Experiment 1, and the correlation between the AH5 and Part 1 of the AH4, which it most closely resembles, was .61. The test–retest scores on accuracy (i.e., number correct) on the indicative and deontic selection tasks were .62 and .34, respectively, with an overall correlation of .38. These correlations are all significant ($p < .01$) but modest, suggesting that there is a certain amount of random variation in performance on selection tasks, which may in turn provide part of the explanation for the conflicting findings in previous studies of individual differences using this task. These reliabilities are a little lower than those found by Klaczynski et al. (1998), though they used more tasks. However, it is important to note that the results of both Experiment 1 and the follow up are completely typical: The card choices are the same as those usually observed, and the usual facilitation with deontic tasks is present.

It would thus appear that the measures we used in Experiment 1 are reasonably reliable, with the possible exception of the selection tasks themselves. What is more, the results from this replication study precisely mirrored those of the main experiment. There was the same facilitation of deontic over indicative tasks, and once again those of higher overall ability did not give the correct *p + not-q* response on the indicative task: Not one person in the top quartile for ability gave this response on either indicative problem. Those of higher ability did, however, tend to give the correct response on the deontic tasks. Those in the top quartile gave 64% correct responses on the Debenhams problem and 91% correct on the drinking age problem, compared to 18% and 63% by the bottom quartile. This difference was significant on the Debenhams problem, $\chi^2 = 4.7$, $p < .05$, but not on the drinking age problem, $\chi^2 = 2.35$, $p > .1$.

Those participants who gave consistent responses to the letter–number and the destination tasks in both Experiment 1 and this replication study were significantly higher in ability ($n = 35$, z score = 0.27) than the inconsistent reasoners ($n = 9$, z score = –1.11), $t(42) = 2.18$, $p < .05$. (The z scores were used in this analysis in order to enable the scores from the AH5 and AH4 to be combined into a composite ability score.)

We have thus fully replicated the unusual effects obtained in Experiment 1 with a subset of the participants. However, we also have evidence that the participants we used may have been of slightly lower ability than would have been expected from the norms and we will argue that this might provide an explanation for the results. In addition the number of participants was small by the standards of individual differences research, and further evidence is clearly required. Hence it was decided to carry out a further study using more participants and additional tasks. As indicated in the introduction, syllogisms have been frequently used in previous research, and hence it was decided to investigate syllogistic reasoning as well as selection tasks.

EXPERIMENT 2

Method

Participants

Participants were 152 students and staff, 86 female and 66 male, recruited from the University of Plymouth, who received either course credit or payment or both for taking part.

Materials

The AH4 and REI were used, together with a further individual difference measure, the Thinking Dispositions Composite (TDC), which has been shown by Stanovich and West (1998b) to correlate with reasoning performance. The subscales used in this vary from one study to another; we chose to use the version used by Stanovich and West (1998b, Experiment 1), which has subscales of actively open-minded thinking, absolutism, dogmatism, counterfactual thinking, and belief in the paranormal (we chose to omit the social desirability questions). These measures may be indirectly related to intuition; actively open minded thinking, for example, may be the opposite of intuitive thinking.

Eight selection tasks were used, four indicative (letter–number and destination as used in Experiment 1, and the menu and grades problems used by Stanovich & West, 1998a), and four deontic (Debenhams, drinking age, charity, and cigarettes). These latter two were designed especially for this study; the charity problem involves a worker in a charity shop where the rule is that any sale under £10 must be paid for in cash, while the cigarette problem involves a shopkeeper who needs to ensure that purchases of cigarettes are made only by children aged 16 years or more.

In addition, 16 syllogistic reasoning problems were used. These were in two forms, which can be represented abstractly as:

(1)	(2)
No A are B	Some A are B
Some C are B	No C are B
Therefore, some C are not A	Therefore, some C are not C

These two forms have been widely used in studies of belief bias (see, e.g., Evans, Barston, & Pollard, 1983; Newstead, Pollard, Evans, & Allen, 1992). Realistic material was used, of which some led to definitionally believable (or unbelievable) conclusions, such as "Some mammals are not dogs", and some led to empirically unbelievable (or believable) conclusions, such as "Some vitamin tablets are not nutritional". Within each form, four syllogisms employed definitional material, and four employed empirical material. Within each set of four, one syllogism led to a valid believable conclusion, one to a

valid unbelievable conclusion, one to an invalid believable conclusion and one to an invalid unbelievable conclusion. This is an example of a valid, believable, Form 1 syllogism using definitional material:

No dogs are unhappy
Some mammals are unhappy
Therefore, some mammals are not dogs

The instructions were taken from the standard instructions used by Evans, Newstead, Allen, and Pollard (1994). They indicated that the information given should be assumed to be true and stressed that a conclusion should be accepted only if it followed logically from this information.

Procedure

Participants were run in groups of two or more. The order of presentation was: AH4; first four selection tasks; TDC; REI; syllogisms; remaining four selection tasks. The four selection tasks within each group were chosen at random with the requirement that there were always two indicative and two deontic tasks.

Results

The mean AH4 score for our sample in this experiment was 93.14. As with the AH5 in Experiment 1, this mean was significantly below the norms for a sample of university students (n = 726, AH4 = 96.36), $t(876) = 2.39$, $p < .01$.

Reliability of measures. The α values on the REI were at a similar level to those reported in Experiment 1, and the factors of cognitive motivation and experientiality clearly emerged in the factor analysis. Of more interest are the findings with the TDC since little has been published on the reliability of this instrument. Stanovich and West (1998b) calculated the TDC score by subtracting the scores on three of the subscales (dogmatism, absolutism, and paranormal belief) from the scores on the other subscales; we ensured that all scores were positive by reverse scoring these three subscales. The coefficients for most of the subscales were somewhat low: For actively open minded thinking, $\alpha = .44$, for absolutism $\alpha = .58$, and for dogmatism $\alpha = .49$. Only the paranormal belief subscale yielded an acceptable α of .78 (no coefficient could be calculated for counterfactual thinking since this measure involved only two items). However, the overall Cronbach's α coefficient for the complete scale was acceptable at $\alpha = .70$.

Factor analysis of the TDC revealed two factors, on one of which belief in the paranormal and some of the absolutism and dogmatism items loaded (see Table 4). The second factor was primarily one on which belief in the paranormal items loaded negatively. Whilst overall the scale is relatively coherent, the reliabilities of the subscales are on the low side. By way of contrast, the REI gave an overall α of .81, and none of the four subscales produced α values lower than .73.

Selection tasks. Responses given on the eight selection tasks are presented as a function of ability level in Table 5. There was the usual massive facilitation on deontic tasks (53% correct vs. 5% correct), but it is also worthy of note that there were differences within the deontic tasks. The Debenhams and charity problems were modest facilitators (36% and 35 % correct, respectively), the drinking age was better than these (67%), and the best facilitator of all was

TABLE 4
Factor analysis of TDC

	Factor	
	1	*2*
AOT1	–.19	–.26
AOT2	.07	.12
AOT3	–.38	.05
AOT4	–.32	–.04
AOT5	–.15	.07
AOT6	–.47	–.10
AOT7	–.22	–.23
AOT8	–.16	–.34
AOT9	–.11	.10
AOT10	–.09	–.11
AB1	.40	.14
AB2	.21	.11
AB3	.38	.30
AB4	.13	.18
AB5	.37	.22
AB6	.11	.08
AB7	.44	.39
AB8	.20	.20
AB9	.48	.04
DOG1	.49	.45
DOG2	.43	.25
DOG3	.04	.08
CFT1	–.22	–.04
CFT2	–.15	–.14
PB1	.53	–.38
PB2	.53	–.42
PB3	.29	–.42
PB4	.50	–.40
PB5	.35	–.43
PB6	.44	–.43
Proportion of variance explained[a]	11	7

Note: AOT = actively open minded thinking; AB = absolutism; DOG = dogmatism; CFT = counterfactual thinking; PB = paranormal belief.
[a]In percentages.

TABLE 5
Cards chosen on each selection task as a function of ability in Experiment 2

		p + *not q*		*p* + *q*		*p*		*All*		*Other*	
Task		*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*
Letter-number	Top	1	3	21	57	11	30	1	3	3	8
	Middle	4	5	44	56	18	23	2	3	10	13
	Bottom	–	–	31	84	2	5	1	3	3	8
Destination	Top	2	5	24	65	7	19	–	–	4	11
	Middle	4	5	41	53	19	24	2	3	12	15
	Bottom	3	8	27	73	5	14	1	3	1	3
Menu	Top	2	5	23	62	6	16	2	5	4	11
	Middle	4	5	38	49	22	35	1	1	8	10
	Bottom	1	3	24	65	5	14	2	5	5	14
Grades	Top	1	3	23	62	6	16	–	–	7	19
	Middle	5	6	41	53	16	21	4	5	12	15
	Bottom	2	5	23	62	5	14	2	5	5	14
Debenhams	Top	18	49	11	30	5	14	–	–	3	8
	Middle	28	36	15	19	16	21	–	–	19	24
	Bottom	8	22	12	32	8	22	–	–	9	24
Drinking age	Top	29	78	2	5	5	14	–	–	1	3
	Middle	53	68	5	6	12	15	–	–	8	10
	Bottom	20	54	10	27	2	5	–	–	5	14
Cigarettes	Top	31	84	1	3	5	14	–	–	–	–
	Middle	55	71	5	6	12	15	2	3	4	5
	Bottom	26	70	3	8	4	11	1	3	3	8
Charity	Top	14	38	9	24	3	8	2	5	9	24
	Middle	25	32	13	17	11	14	2	3	27	35
	Bottom	14	38	10	27	1	3	1	3	1	30

Note: Top = top quartile on AH4 (n = 37); Middle = middle two quartiles (n = 78); Bottom = bottom quartile (n = 37).

the cigarettes problem (74%). Sign tests (p = .01) revealed no difference between the Debenhams and charity problems, but these were both significantly worse facilitators than the drinking age and cigarettes problems, which did not differ from each other.

As can be seen in Table 5 there were no differences in the number of *p* + *not-q* answers given by the different ability groups on the indicative problems. However, there was a hint of a difference in the number of *p* responses. The difference between the top and bottom ability groups was not significant, $t(72) = 1.32$, $p > .1$, but there was a marginally significant difference between the top and middle groups combined and the bottom group, $t(150) = 1.95$, $p = .052$, two-tailed. The choice of the *p* card alone has been argued (e.g., by Klaczynski, 2001) to be an indicator of analytic ability on this task. No other ability-related differences emerged in

the data. On the deontic tasks the top ability group produced more correct responses than the bottom group, Mann–Whitney $U = 451$, $z = 2.63$, $p < .01$. No other significant effects emerged in the data.

We repeated the analysis carried out in Experiment 1 by looking at the ability scores of those who made the same selection pattern on all four of the indicative selection tasks compared to those who gave different responses. Once again the consistent responders were significantly higher in ability ($n = 75$, AH4 = 95.5) than the inconsistent responders ($n = 77$, AH4 = 90.9), $t(150) = 1.89$, $p < .05$.

There were no differences in the patterns of responses between those high and low in either experientiality or cognitive motivation. The only trend of possible interest was that those high in experientiality tended to give more $p + q$ responses on the indicative tasks than did the low group (66% vs. 51%) and to give fewer p responses (10% vs. 29%). However neither of these trends reached significance and, as can be seen in Table 2, no significant trends emerged when the data from different experiments were combined. There were no discernible trends on the data from the TDC measure.

Syllogisms. The typical pattern of results was found on the syllogisms—that is, a main effect of belief, $F(1, 151) = 79.76$, $MSE = 1.33$, $p < .001$, a main effect of logic, $F(1, 151) = 52.77$, $MSE = 1.22$, $p < .001$, and an interaction between belief and logic, $F(1, 151) = 5.86$, $MSE = 0.76$, $p < .05$ (see Table 6).

A logic index was calculated by subtracting the number of endorsements of invalid conclusions from the number of valid conclusions endorsed, and a belief index was calculated by subtracting endorsement of unbelievable conclusions from endorsement of believable ones. Intellectual ability correlated significantly ($p < .05$) with logic scores ($r = .17$) and with scores on conflict problems (i.e., valid unbelievables and invalid believables, $r = .18$). Cognitive motivation correlated significantly with performance on nonconflict problems ($r = .19$), and TDC correlated with logic scores ($r = .18$). There was a suggestion that consistent responders (i.e., those who gave the same response on all four indicative tasks) were less influenced by beliefs than inconsistent responders (belief index = 1.10 vs. 1.49) and also performed better on conflict problems (mean number correct = 4.39 vs. 3.99), but neither of these effects was significant, ts(150) = 1.07 and 1.43, respectively, ps > .1.

TABLE 6
Percentage of syllogistic conclusions endorsed as a function of belief and logic

	Believable		*Unbelievable*	
	Exp. 2	*Exp. 3*	*Exp. 2*	*Exp. 3*
Valid	82	76	79	70
Invalid	70	62	45	37

Discussion

With respect to the selection task, the present results broadly confirm the findings of Experiment 1. Those of higher ability did not do any better on the indicative selection tasks but they did perform better on the deontic tasks. We have also confirmed another finding from the previous experiment, in that participants who gave the same response to the four indicative selection tasks were of higher ability than those who gave different responses.

Clearly this difference between consistent and inconsistent responders is a robust one. There are a number of possible reasons why it might occur. A rather mundane explanation is that those of lower ability were responding randomly. This seems unlikely (though not impossible) given that inconsistent responders showed the classic difference between deontic and indicative selection tasks (50% versus 7% correct), $t(76) = 11.25$, $p < .01$, and the typical effects of logic, $F(1, 76) = 34.13$, $p < .01$, and belief, $F(1, 76) = 39.89$, $p < .01$, on syllogisms (though the interaction effect was not significant, $F(1,76) = 1.60$). Another possible explanation is in terms of transfer effects, in that the more able participants may simply remember what their response was to the earlier task and repeat it. However, this would presumably lead to them give the same response to all tasks, whether indicative or deontic, which clearly did not happen.

A more likely explanation is that the more able students recognized similarities between the different problems that the less able people did not. Thus the more able participants realized that the different indicative problems had something in common—the use of the conditional rule—and hence gave the same response to them all. Less able participants treated each problem on its own merits and failed to recognize that they possessed a common underlying structure. According to this explanation, the more able people abstracted the underlying form of the rule but the less able simply responded on the basis of whatever contextual cues were available in the problem.

What we are suggesting in effect is that there are two different types of decontextualization involved in solving indicative selection tasks. First, the conditional rule must be decontextualized from its scenario; those who achieve this will respond to all indicative rules in the same way, usually by choosing the two cards named in the rule. Second, invited inferences, such as the assumption that conditionals are biconditionals, must be resisted; those who do this will give the correct response. This latter suggestion is somewhat speculative since there were few if any participants in our studies so far who have achieved this higher level of decontextualization. Those of higher ability achieved the first level of decontextualization only, whereas those of lower ability did not achieve even this.

It may seem a little surprising to talk of contextual cues with indicative problems, but there is much evidence that even the most abstract scenarios produce probabilistic, knowledge-based cues (e.g., Evans, 1995; Schroyens, Schaeken, Fias, & d'Ydewalle, 2000). The contextual cues are even more compelling on the deontic tasks, and on the basis of the differences between the deontic tasks we might speculate that the contextual cues are particularly strong on the drinking age and cigarettes problems. Clearly, in order to test the claim that resisting invited inferences is a key factor in explaining the differences between ability levels on indicative tasks we need an independent measure of the ability to resist invited inferences. This was a major purpose of the next experiment.

Turning now to the data from the syllogistic reasoning task, there was clear evidence that ability correlated with logical performance on syllogisms. This showed up in a correlation both between ability and logic scores and between ability and conflict problems. Thus participants who resisted the response cued by the believability of the conclusion and reasoned logically were of higher ability. There was little indication of any relationship between measures of thinking style and syllogistic reasoning, apart from correlations between cognitive motivation and scores on nonconflict problems, and between the TDC and logic scores. Ability was unrelated to any of the measures of thinking style.

The findings from the two experiments reported so far show consistent discrepancies from previous research. The most plausible explanation for these discrepancies lies in the ability level of our samples, which seemed to be lower than what might have been expected in such populations. Experiment 3 was designed to further test the replicability of this finding and also to investigate other individual differences, including the ability to resist invited inferences and the ability to generate alternative representations and explanations.

A range of inferences are invited by a standard conditional of the form *if p then q*. These inferences may include inferring the converse, *if q then p,* or the inverse, *if not p then not q* (Geis & Zwicky, 1971; Rumain, Connell, & Braine, 1983). On the selection task, inferring the reverse conditional will lead to the selection of the *q* card, whilst the reverse contrapositive will lead to the selection of the *not-p* card. Resisting these inferences is crucial to identifying the correct selections on the selection task. Hence we would expect measures that reflect the tendency to resist these inferences to correlate with indicative selection task performance. In contrast, a deontic rule, such as "If a person is drinking beer then they must be over 18 years of age", does not invite these inferences. People do not infer that "If a person is over 18 then they must be drinking beer". Hence we would not expect a relationship here.

As a measure of the tendency to resist invited inferences, in Experiment 3 we introduced a conditional inference task. This task involves participants being presented with a conditional statement (*if p then q*) followed by an assertion or denial of one of the terms; the participant's task is to indicate whether a specified conclusion definitely follows or definitely does not follow, or whether there is insufficient information to decide. There are two valid inferences: modus ponens and modus tollens. Under modus ponens (MP), the assertion of *p* allows one to infer that *q* is true, while modus tollens (MT) allows one to infer from *not q* that *not p* is the case. There are no other valid inferences. One can deduce nothing from the denial of *p*, nor from the assertion of *q*.

Previous research (see, e.g., the summary in Evans, Newstead, & Byrne, 1993) has shown that almost everyone draws the modus ponens inference (MP), and a majority draw modus tollens (MT). However, people often believe that *not q* follows from *not p* (the denial of the antecedent, DA, fallacy), and that *p* follows from *q* (the affirmation of the consequent, AC, fallacy). These fallacies follow from the invited inferences described above. Hence we make the strong prediction in Experiment 3 that performance on indicative selection tasks will be correlated with the tendency for people to resist the AC and DA fallacies. Further, we predict that this tendency will be related to ability.

A second main aim of Experiment 3 was to examine the relationship between the selection task, syllogistic reasoning, and a new variable: the tendency or ability to generate and consider alternatives. Stanovich and West (1998b) have argued that thinking disposition may reflect

the tendency to consider alternatives or counterexamples to initial conclusions. The lack of any consistent correlations between thinking style and reasoning performance in our studies may reflect the fact that self-report measures of thinking style only indirectly and poorly reflect the variable of alternatives generation.

Theoretical accounts of human reasoning also point us in the direction of alternatives generation as a candidate predictor of deductive reasoning. For example, according to the mental models theory (Johnson-Laird & Byrne, 1991), syllogistic reasoning involves generating alternative representations of information in order to falsify putative conclusions. Torrens et al. (1999) have demonstrated that the ability to generate alternative representations of categorical syllogisms correlates positively with logical performance on conditional syllogisms and negatively with the susceptibility to belief bias, and in some ways it is a better predictor than measures of intellectual ability (see also Newstead, Thompson, & Handley, 2002). Similarly Markovits (e.g., 1984) has argued that resisting the fallacies on a conditional inference task crucially depends upon the number of alternative antecedents that can be generated. Given this analysis, it is possible that alternatives generation is a key variable, possibly independent of cognitive ability, which will predict aspects of reasoning performance. Hence, in addition to the measures used in previous studies (AH4 and REI), Experiment 3 used a number of measures of alternatives generation in order to explore the importance of this individual differences variable in reasoning.

EXPERIMENT 3

Method

Participants

Participants were 70 psychology students from the University of Plymouth who participated for course credit.

Materials

The AH4 (Parts 1 and 2), the REI, the indicative selection tasks (letter–number and destination), deontic selection tasks (Debenhams and drinking age), and the syllogisms were exactly the same as those used in the previous experiment. The conditional reasoning task involved conditional statements such as "If there is a square, then there is a blue shape", followed by an assertion and a conclusion, for example: "There is a square. Therefore there is a blue shape". This is a modus ponens inference of the form: *if p then q*; *p*; therefore, *q*. The task was to say whether this conclusion definitely followed or definitely did not follow, or whether there was insufficient information to tell. In this example the logically correct answer would be that *q* definitely follows. The other valid inference, modus tollens, involves *not q* as the second premise, from which *not p* follows. The denial of the antecedent (*not p*) and the affirmation of the consequent (*q*) lead to no valid conclusion, hence it would be logically correct to indicate that no conclusion follows. In addition to the double affirmative conditional (*if p then q*) we also used the forms *if p then not q*, *if not p then q*, and *if not p then not q*. Each of these was paired with one example of each of the inferences (MP, MT, DA, and AC), making 16 problems in all.

The possible diagrams task was modelled closely on that of Torrens et al. (1999) and involved participants generating as many different ways of representing syllogisms as possible. The example given in the instructions used the premises "All of the jeans are blue and all of the jeans are denim" (a multiple model syllogism), and it was explained that the task required drawing a diagram showing how jeans, blue things,

and denim related to each other. A sample diagram was given in which the first premise was represented as a subset and then combined with a subset diagram representing the second premise such that there was an overlap relationship between blue things and denim things. It was pointed out that other alternatives were possible, and a diagram presenting an overall subset relationship was given. The instructions indicated that participants should draw as many diagrams as they could until they could think of no more. The task itself used four pairs of quantified premises using neutral content, and the responses were scored in terms of the number of different but correct diagrams produced.

The uses of objects test asked participants to give as many uses for a paper clip and a brick as they could within a 2.5-min period for each object. The score was simply the total number of responses generated after obvious repetitions had been eliminated.

The possible antecedents task was modelled on that used by Markovits (1984). Participants were given a conditional statement such as "When David has homework to do, he gets into a bad mood. I saw David after school today and he was in a bad mood. Can you imagine what would have put David in a bad mood?" They then had 30 s in which to produce as many possibilities as they could, and the score was the total number produced. The other situations used in this task were: "When it rains, the pavements become wet. On the way home from school today Sophie noticed that the pavements were wet"; "When Mark's dog has fleas it scratches constantly. When Mark came home from work yesterday he noticed that his dog was scratching constantly"; and "When Catherine eats mushrooms she is sick. This afternoon Catherine was sick".

The unlikely scenarios task was modelled on suggestions of Johnson-Laird (1993a, 1993b). Participants were given a scenario such as: "A murder was committed in Soho yesterday. The police have a prime suspect but the suspect was on a train travelling to Scotland at the time of the murder. Can you think of any way in which the suspect could have committed the murder?" The other scenario involved an old man bitten by a poisonous snake for which there is no known antidote, and yet the man did not die. In each case participants were asked to list as many possibilities as they could within two minutes, and the score was the number of possible scenarios produced.

The experimental confounds task involved a brief description of an experimental study based loosely on the cola example in Huck and Sandler (1979, p. 11). It described a study carried out a by a cola manufacturer in which participants had to choose between the company's cola in a blue container labelled "B" and their chief rival's cola in a red container labelled "R". On the basis of the finding that 90 people out of 100 preferred their brand, the company claimed that 90% of people preferred their cola. The task was to indicate as many reasons as possible in two minutes for this claim not being valid.

Procedure

Participants were run in groups of differing sizes. They were given the AH4 first, and then the uses of objects, possible antecedents, experimental confounds, and unlikely scenarios tasks in random order. Then came the syllogisms task, the REI, conditional inference, and possible diagrams tasks. Finally the selection tasks were presented in random order.

Results

Intelligence scores. The mean score on the AH4 amongst our sample was 98.8 ($SD = 14.4$). In contrast to previous studies the mean score was not significantly different to the norms for university students on the test ($M = 96.36$, $SD = 15.01$, $n = 726$), $t(794) = 1.29$, $p > .1$. The sample scored significantly higher than the participants in Experiment 2, $t(220) = 2.63$, $p < .01$.

Measures of alternatives generation. The various measures used to assess the ability to generate alternatives have not been previously used together, and indeed some were devised specifically for this study. We carried out a principal components factor analysis to see what dimensions underlie these measures. We constrained the analysis to two factors. The results of this can be seen in Table 7. Factor 1 corresponds in general to the possible diagrams questions, although one of the uses of objects items and one of the scenario problems also loaded moderately on this factor. Factor 2 corresponds to the uses of objects, unlikely scenarios, experimental confounds items, and possible antecedents items. On the basis of these results we felt justified in using two measures of alternatives production: possible diagrams and a composite of all the other alternatives generation measures. These latter measures all involve generating alternative verbal explanations or uses (as opposed to spatial representations in the possible diagrams task) and will be referred to collectively as the verbal alternatives measure. The reliabilities of each one of these scales was acceptable: possible diagrams, $\alpha = .76$, verbal alternatives, $\alpha = .71$.

The second new task that we introduced in this experiment was the measure of conditional reasoning. Recall that our prediction here was that the tendency to resist the fallacious AC and DA inferences would be correlated with performance on the indicative selection tasks but not with that on the deontic tasks. Each participant received four problems corresponding to each of the inferences, MP, MT, AC, and DA. For MP and MT participants were awarded one point for correctly drawing the inference, whilst for AC and DA they were awarded one point for correctly responding that it was not possible to tell what followed. The overall pattern of

TABLE 7
Factors loadings on measures of alternatives generation

	Factor	
	1	*2*
Antecedents 1		.55
Antecedents 2		.73
Antecedents 3		.58
Antecedents 4		.52
Uses 1		.42
Uses 2	.48	.41
Scenario 1		.62
Scenario 2	.40	.42
Confounds		.56
Diagrams 1	.80	
Diagrams 2	.79	
Diagrams 3	.54	
Diagrams 4	.73	
Eigenvalue	2.75	2.78
%variance	21%	21%

Note: Only factor loadings greater than .4 have been included.

results was similar to that found in previous published studies. The great majority of participants (92%) correctly endorsed the MP inference but a much smaller percentage (44%) endorsed the equally valid MT inference. The correct response to the DA and AC inferences is to say that nothing follows, and this response was given by 43% and 30% of the participants, respectively.

As expected, scores on AC and DA were highly correlated, $r = .56$, $p < .001$. Also as predicted, both of these inferences were correlated with AH4 scores, $r = .43$, and $r = .50$, respectively, ps $< .001$. These patterns of correlation suggest that those participants who resist AC are more likely to resist DA, and further that participants who resist both these inferences are of higher ability than those who do not. AC and DA also correlated with the possible diagrams task, $r = .52$ and $r = .27$, respectively, ps $< .05$, and DA also correlated with the verbal alternatives task, $r = .30$, $p < .05$. Given the high correlation between AC and DA, and given that resisting these inferences can be viewed as reflecting the ability to resist the pragmatic inferences that conditionals cue, in remaining analyses we combined the data for DA and AC into an overall fallacy index. The fallacy index is thus a measure of the ability to resist the common fallacies.

The patterns of correlations for MP and MT were unexpected. MP did not correlate with any of the other conditional inferences (rs range from .02 to .12), but there was a significant correlation with AH4 scores, $r = .35$, $p < .01$. In contrast, MT tended to correlate negatively, although nonsignificantly, with AH4 scores, $r = -.14$, $p > .1$, and negatively and highly significantly with the fallacy index, $r = -.52$, $p < .001$. Thus resisting the AC and DA fallacies is associated with rejection of MT, suggesting that MT may be supported by pragmatic inferences, in much the same way as fallaciously accepting the AC and DA inferences.

Selection tasks. There was the usual facilitation of deontic over indicative tasks (50% versus 14%). The relationship between ability and response patterns is presented in Table 8. In this study there was for the first time in this series of studies a suggestion of a relationship between ability and logical correctness on the indicative tasks. Those high in ability produced 19% correct answers on these tasks compared to 9% of those of low ability, though this trend did not reach significance ($\chi^2 = 1.14$ on the letter-number task and $\chi^2 = 0.24$ on the destination task, ps $> .1$). There was also a trend for the high ability group to produce more *p* responses than those of low ability (34% vs. 6%), which achieved significance on the letter–number task, $\chi^2 = 7.39$, $p < .01$, but not on the destination task, $\chi^2 = 1.65$, $p > .1$.

On the deontic tasks there was, again for the first time in this series of studies, no hint of a relationship between ability and correct performance, nor were there any other discernible trends in the data except that the low ability group produced more "other" responses, $t(68) = 2.07$, $p < .05$.

On our account, accurate performance on the indicative selection task depends upon the ability to decontextualize the problem from the accompanying scenario and subsequently resist the invited inferences that the conditional rule cues. Hence we predicted that a measure reflecting the tendency to resist the fallacies on the conditional inference task would correlate with performance on the selection task. In order to test this we used an accuracy index in which correct card selections (*p*, *not-q*) were given a score of +1 each while incorrect cards (*not-p*, *q*) were each given a score of –1. This gives a range of scores on each problem of +2 to –2. We

TABLE 8
Cards chosen on each selection task as a function of ability in Experiment 3

		p + *not q*		*p* + *q*		*p*		*All*		*Other*	
Task		*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*
Letter-number	Top	3	19	5	31	6	38	1	6	1	6
	Middle	11	29	17	45	4	11	–	–	6	16
	Bottom	1	6	5	31	–	–	–	–	10	63
Destination	Top	3	19	6	38	5	31	1	6	1	6
	Middle	9	24	17	44	4	11	1	3	7	18
	Bottom	2	13	4	25	2	13	1	6	7	44
Debenhams	Top	7	44	2	13	6	38	–	–	1	6
	Middle	19	50	5	13	6	16	1	3	7	18
	Bottom	8	50	1	6	2	13	–	–	5	31
Drinking age	Top	9	56	–	–	4	25	–	–	3	19
	Middle	18	47	5	13	6	16	–	–	9	24
	Bottom	9	56	–	–	3	19	–	–	4	25

Note: Top = top quartile on AH4 (n = 16); Middle = middle two quartiles (n = 38); Bottom = bottom quartile (n = 16).

recognize that this index has certain problems in that it can give identical scores to very different patterns of responses, but it provides a reasonable measure of accuracy and, importantly, a continuous range of scores.

Table 9 shows that there was a strong correlation between the index scores on the indicative selection tasks and the fallacy index, $r = .39$, $p < .01$. There was also a smaller but still significant correlation between the deontic tasks and the fallacy index, $r = .27$, $p < .05$. A direct statistical comparison of the correlation coefficients for the deontic and indicative tasks revealed a difference that approached significance, $t(67) = 1.40$, $p < .1$, one-tailed.

Of the two alternative generation measures employed in this study, only the possible diagrams task showed positive correlations with selection task performance (see Table 9). The correlation between this measure and performance on indicative tasks was significant but

TABLE 9
Correlations obtained in Experiment 3

	Indicative	*Deontic*	*Logic*	*Belief*	*Conflict*	*Non-conflict*
AH4	.30*	.15	.27*	–.11	.23	.10
Cognitive motivation	.05	–.06	.33*	–.01	.20	.24*
Experientiality	.07	.07	–.22	.08	–.18	–.09
Possible diagrams	.41**	.23	.35*	–.07	.25*	.19
Verbal alternatives	.11	.11	.26*	–.03	.17	.17
MP	.17	.04	.00	.11	–.07	–.10
MT	–.21	–.27*	–.08	.37**	–.30*	.26
DA/AC index	.39**	.27*	.27*	–.18	.27*	.05

Note: *$p < .05$; **$p < .01$.

marginally missed significance on the deontic tasks. In order to determine whether the possible diagrams task was predicting variance in performance on the indicative tasks independently of cognitive ability, we performed a standard multiple regression analysis with the index score on the indicative task as the dependent variable, and AH4 and possible diagram scores as the predictors. The analysis yielded a highly significant regression, $F(2, 67) = 8.94$, $p < .001$, adjusted $R^2 = .18$, and confirmed that the possible diagrams score was a unique predictor of performance, $t(67) = 3.17$, $p < .01$, with AH4 scores a much weaker predictor, $t(67) = 1.81$, $p < .1$.

Those who gave consistent responses on the indicative selection tasks were of higher ability ($n = 42$, AH4 = 102.7) than inconsistent responders ($n = 28$, AH4 = 92.9), $t(68) = 2.96$, $p < .01$. This difference was significant even when the 10 consistent responders who gave the logically correct answer to both of the indicative tasks were removed from the analysis, $t(58) = 2.56$, $p < .05$. This provides additional confirmation of the potentially important theoretical distinction between participants who give consistent and those who give inconsistent responses on the indicative tasks.

In order to explore these effects further we carried out a between-groups analysis of variance on ability scores and the fallacy index for the consistent correct responders, the consistent incorrect responders, and the inconsistent responders. On the basis of the theoretical analysis of the groups we predicted linear trends in these variables. A general linear model analysis revealed a significant linear trend in intelligence, $F(1, 68) = 7.96$, $p < .01$. There was also a significant linear trend on the fallacy index, $F(1, 68) = 12.24$, $p < .001$, with the consistent correct responders rejecting the fallacies more often ($M = 5.0$) than the consistent incorrect responders ($M = 3.18$) and the inconsistent responders ($M = 2.04$). A breakdown analysis revealed that consistent correct responders rejected the fallacies more often ($p < .05$) than the other two groups but that these did not differ from each other. This provides strong evidence that consistent responders are less able than consistent correct responders to resist pragmatically invited conditional inferences.

Syllogisms. The overall pattern of results on syllogisms was very similar to that obtained in Experiment 2, as Table 6 shows. An analysis of variance revealed significant main effects of logic and of belief subsumed by a significant interaction between these factors, $F(1, 69) = 12.31$, $MSE = 0.81$, $p < .001$, reflecting a greater effect of belief on invalid problems. The correlations, too, were similar to previous experiments, as can be seen in Table 9. The logic index correlated positively with AH4 and also, in this study, with cognitive motivation ($r = .33$). Performance on nonconflict syllogisms correlated with cognitive motivation, and the correlation between the AH4 and conflict syllogisms approached significance. There were also correlations with some of the new measures introduced for the first time in this study, notably positive correlations between the logic index and both of the alternative generation measures, and between conflict syllogisms and possible diagrams. The fallacy index correlated significantly with the logic index and with scores on the conflict syllogisms. In sharp contrast, MT inferences correlated positively with the belief index, negatively with conflict syllogisms, and positively with nonconflict syllogisms.

Correlations between measures. For the first time in this series of experiments cognitive motivation correlated with the AH4, $r = .43$, $p < .001$, and it also correlated marginally with

the alternatives generation measure, $r = .23$, $p < .06$. The verbal alternatives measure correlated with the AH4, $r = .33$, $p < .01$, as did the possible diagrams task, $r = .28$, $p < .05$.

Discussion

In this study, the relation between ability and performance on indicative selection tasks, so elusive in our earlier studies, has to some extent returned. The overall ability level of the participants was in line with the norms for the test and significantly higher than that in Experiment 2. In this experiment we have fortuitously sampled a population of higher ability, and this is reflected in more accurate performance on the indicative selection tasks (21% correct) compared to a much lower performance in preceding studies (6% correct overall). There was no evidence in Experiment 3 of any relationship between ability and performance on the deontic tasks, a finding that was clearly apparent in Experiments 1 and 2. These findings lend strong support to the claim that the unusual results of the first two experiments are primarily attributable to the lower ability levels of the sample tested.

The conditional reasoning task was introduced primarily to provide a measure of the ability to resist invited inferences, which we assumed would be a key factor in accurate performance on the selection task. The findings from the present study provide strong support for this aspect of our proposals. As predicted, those who were more able to resist these inferences performed better on the indicative selection tasks.

A second aim of Experiment 3 was to examine the relationship between measures of alternatives generation and performance on reasoning tasks. The rationale for the introduction of measures of this kind was in part based on the proposal that thinking style reflects the tendency to consider alternative possibilities. The possible diagrams and the verbal alternatives tasks were introduced in an attempt to provide a process-oriented measure reflecting this construct. In some respects this attempt was unsuccessful as neither cognitive motivation nor experientiality correlated significantly with the alternatives generation measures. However, alternatives generation (especially possible diagrams) was related to selection task performance, and in addition both measures correlated with logic scores on the syllogisms task, with possible diagrams also predicting performance on the conflict syllogisms. The success of these measures (especially diagram production) in predicting reasoning performance is of interest in its own right. Mental models theory claims that an essential part of reasoning is the construction of alternative representations, and the predictive success of this measure lends support to this claim: Those who are better at constructing different mental models are better reasoners.

The conditional inference task has not been widely used in studies of individual differences, and the findings are of considerable interest in their own right. Logically correct responses on three of the inferences (MP, AC, and DA) correlated significantly with ability but MT inferences did not. Similarly, MP and the AC/DA fallacy index correlated positively with the measures of alternatives generation. MT correlated significantly with none of these measures; only the correlation between MT and possible antecedents even approached significance. Similarly, while AC and DA correlated with each other, they both correlated negatively with MT. This pattern of findings shows very clearly that those participants who resist AC and DA also fail to make the MT inference. It is notable that Markovits, Doyon, and Simoneau (2002) found a similar pattern of results using measures of working memory rather than intellectual ability. These findings suggest that the pragmatic processes that lead

reasoners into drawing the fallacies may also lead them into drawing the valid MT inference. Resisting these inferences impacts not only on the fallacies but also on MT.

GENERAL DISCUSSION

A principal rationale for the present research was to increase the range of individual differences measures that have been used in reasoning research and thereby throw light on the theoretical debates surrounding this area. With some of the measures the results have been disappointing. Cognitive motivation and experientiality, as measured by the REI, have proved to be extremely reliable measures but have produced few correlations with the measures of reasoning used in these studies. Experientiality did not, as we had predicted, correlate with performance on deontic selection tasks and with belief bias effects on syllogisms. There was a suggestion in Experiment 1 that experientiality was related to the frequency with which correct responses were given to deontic tasks, but this did not replicate in the other studies. It seems likely that the kind of intuitive responding measured by the experientiality scale is not the type of intuition that leads to these two contextual effects.

In the Introduction it was pointed out that some authors have made the link between experientiality and the intuitive, automatic, unconscious, System 1 type of thinking. The present negative findings undermine this claim. There is in any case something inherently implausible about measuring unconscious processes using a self-report questionnaire. Experientiality appears to be a sound construct, and there is much independent evidence for the existence of two kinds of thought; our only claim here is that experientiality, as measured by the REI, is not a unitary measure of System 1 thinking. It is, of course, possible that there are different kinds of System 1 thinking and that experientiality measures some of these, but there is little evidence that it correlates with the kinds of bias studied in the present series of experiments.

We made no specific predictions about the relationship between cognitive motivation and reasoning tasks, but one effect emerged consistently: a positive correlation with performance on nonconflict syllogisms. In the Introduction we suggested that factors other than ability might predict performance on these syllogisms but did not specifically predict that cognitive motivation would be such a predictor. Cognitive motivation is in large part a willingness to engage in thinking tasks, and this correlation may simply reflect the fact that some people took the tasks more seriously than others. Those who took it seriously were able to perform well on those problems where there were fairly obvious and consistent cues to the correct response. However, in the presence of conflicting cues (i.e., on conflict syllogisms) motivation alone is not enough: It would seem that intellectual ability is required to pick one's way through such conflicting information. Cognitive motivation does not seem to be a good predictor of logical performance in general.

Interestingly, the process measure of cognitive style that was explored in the final experiment, possible diagrams, proved to be a rather better predictor of performance. It correlated with ability and predicted logical accuracy on indicative selection tasks, syllogisms, and conditional reasoning as well as, if not better than, standard measures of ability. The capacity to generate alternative representations is clearly an important aspect of a number of types of reasoning, a finding that will offer considerable comfort to those favouring the mental models approach. However, this process measure did not correlate with cognitive styles as measured

by the REI. This research thus poses a question mark over the use of self-report measures of cognitive style such as the REI and the TDC. They may be reliable but they do not serve as good predictors of performance on tasks such as those used in the present series of studies. Research might be better served by moving towards more process-based as opposed to self-report measures of cognitive style.

The most interesting findings from the present studies are those that we did not predict. We were confident that we would confirm the well-established relationship between ability and performance on indicative selection tasks. Experiments 1 and 2 showed a distinct lack of relationship, and we suggested that this might be due to our sample being of lower ability than that in some previous studies, and our analyses confirmed this. The findings of Experiment 3, which used a group of (serendipitously) higher ability, lent support to this contention. Table 10 presents the results from all of the experiments in comparison to the norms for the AH4 and AH5. It can be readily seen that there were more participants in the top 10% in Experiment 3 than in either of the other two experiments, and correspondingly fewer in the bottom 10%.

In order to provide a fuller picture of our results, we combined the data from the three studies that used the AH4 (Experiments 2 and 3 and the replication study in Experiment 1). The results can be seen in Table 11, and indicate that there is a slight tendency for those of higher ability to perform better on the indicative tasks, which approached significance for the letter–number task, $\chi^2 = 3.08$, $p = .08$. The tendency to choose more p cards alone was significant for the standard letter–number task, $\chi^2 = 8.77$, $p < .01$, but not for the destination task, $\chi^2 = 1.43$, $p > .1$.

There is thus a suggestion that the usual relationship between ability and performance on indicative tasks is present in our data overall. It is clear, however, that this is almost entirely due to the results of Experiment 3 where there were more people of higher ability. It seems reasonable to suggest that it is primarily at the very highest levels of ability that the relationship holds. The claim that correlations with ability confirm that the correct normative model has been applied can thus be upheld provided it is restricted to the higher levels of ability. There is perhaps a catch to this, however. Those at the highest levels of ability are likely to be those with the highest educational attainments. It might be argued that the normatively correct response is that given by people who have been educated in such a way as to lead them to accept the precepts of logical analysis. It is not clear how, or indeed whether, the effects of education and intellectual ability can be completely separated out.

TABLE 10
Ability scores of participants in Experiments 1, 2, and 3 relative to the norms for university students

	Experiment					
	1 (AH5)		*2 (AH4)*		*3 (AH4)*	
	Score	*%*	*Score*	*%*	*Score*	*%*
A (top 10%)	–	–	4	3	8	11
B (next 20%)	–	–	28	18	16	23
C (middle 40%)	4	4	65	43	30	43
D (next 20%)	16	16	38	25	11	16
E (bottom 10%)	78	80	17	11	5	7

TABLE 11
Cards chosen on each selection task as a function of ability (as measured by AH4 scores) in Experiments 1 (replication study), 2, and 3

		p + *not q*		*p* + *q*		*p*		*All*		*Other*	
Task		*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*	*No. of cards*	*%*
Letter-number	Top	7	11	33	52	16	25	3	5	4	6
	Middle	13	9	73	52	30	22	4	3	19	14
	Bottom	2	3	40	63	4	6	3	5	15	23
Destination	Top	6	10	35	56	14	22	1	2	7	11
	Middle	13	9	74	53	26	19	5	4	21	15
	Bottom	6	9	34	53	9	14	4	6	11	17
Debenhams	Top	33	52	12	19	10	16	–	–	8	13
	Middle	55	40	28	20	28	20	3	2	25	18
	Bottom	17	27	17	27	10	16	2	3	18	28
Drinking age	Top	46	73	3	5	8	13	–	–	6	10
	Middle	91	65	8	6	21	15	1	1	18	13
	Bottom	36	56	12	18	5	8	–	–	11	18

Note: Top = top quartile on AH4 (n = 63); Middle = middle two quartiles (n = 139); Bottom = bottom quartile (n = 64).

Once these very high performers are excluded, a different picture emerges. There was little evidence of any link between ability and correct performance on indicative tasks, not surprisingly since there were so few correct responses. However, two other very clear effects were obtained. First, those of higher ability tended to be more consistent in their responses to the indicative tasks, and second they tended to give the correct response to the deontic tasks. We cannot rule out the possibility that there is an element of random responding in the lower ability group, but our preferred explanation is that those of lower ability failed to detect the similarities between different indicative tasks. Those of somewhat higher ability noticed these similarities and responded consistently by choosing *p* + *q* on the indicative tasks and *p* + *not-q* on the deontic tasks. These higher ability participants are thus better able to utilize the contextual and pragmatic cues in the conditional statements.

There is evidence in the developmental literature that ability is related to the tendency to draw pragmatic inferences. Noveck (2000) has drawn attention to a number of curious findings whereby younger children appear to be more logical than older ones. The reason seems to be that younger children have not yet acquired the pragmatic interpretations of certain logical connectives and thus paradoxically respond in a way that is closer to the tenets of formal logic than do older children who have learned such interpretations.

In Experiment 3, the more able participants in this study responded logically on the MP, DA, and AC inferences, suggesting that they had acquired the ability to reason formally. But why was there not a similar correlation with the MT inference? We believe that there are two routes to drawing MT, one of which arises from the conversational pragmatics of conditionals and another that depends upon the sophisticated application of an explicit suppositional

reasoning strategy. We argued earlier that AC and DA fallacies are supported by invited inferences. The robust negative correlation between resisting these fallacies and drawing MT suggests that the same processes underlie the MT inference. A conditional (*if p then q*) may not only invite one to infer *if q then p* or *if not p then not q*, but also to pragmatically infer the logically entailed contrapositive, *if not q then not p*. This inference supports MT, but through a pragmatic process that happens, in this case, to map on to the logic of implication. Those participants who resist the invited inferences may also resist the pragmatic inferences that support MT—hence the absence of a correlation between ability and drawing the MT inference.

This argument is mirrored in the developmental literature where it has been suggested that MT is supported by pragmatic inferences amongst younger children. As children begin to resist these inferences in mid to late adolescence, a reduction in DA and AC fallacies is accompanied by a reduction in MT (O'Brien, Dias, & Roazzi, 1998; Rumain et al., 1983; Wildman & Fletcher, 1977; but see also Klaczynski & Narasimham, 1998, for conflicting findings). According to this account, some older children and adults will develop a more sophisticated suppositional strategy for MT, based upon a complex *reductio ad absurdum* line of reasoning. However, such indirect reasoning routines are only acquired by highly sophisticated reasoners. The individual differences data presented here map on to this developmental account well. Those people of higher ability resist the pragmatic inferences, but cannot find the indirect line of reasoning required for MT. Hence MT correlates in an entirely different way with our other measures, since the participants of lower ability who are more susceptible to pragmatic influences are the ones drawing the inference.

The prediction that ability would correlate positively with logical performance on syllogistic reasoning was confirmed, and the correlation was especially marked on conflict problems. Ability tended to correlate negatively, though not quite significantly, with belief scores. The picture is a little simpler than that with selection tasks, in that right across the ability range it would appear that more able participants are likely to decontextualize the problem and respond according to logic rather than belief. More able students are also more capable of coping with the conflicting cues when logic and belief pull in different directions. This indicates once again that the ability to abstract rules from their context is a key component in reasoning and closely related to intellectual ability.

A number of potential confounds exist in the present series of studies. One of the most important relates to the order in which the various tasks were presented. The timed tasks (such as the ability tests) were for practical reasons always presented first, which means that the reasoning tasks were presented later in the sequence. It is possible that intuitive System 1 thinking is more likely to be used when fatigued, possibly explaining why we failed to find convincing evidence of logical thinking on the indicative selection tasks. We tried to counter such criticisms by presenting the selection tasks in random order in Experiments 2 and 3, and by presenting them in two separate blocks of four in Experiment 2. It is also suggestive that the strongest support for logical reasoning, in the form of a correlation between ability and performance on indicative selection tasks, came in Experiment 3 when the selection tasks were presented last at the end of a long series of tasks. This, combined with the fact that our findings on all tasks were typical of those found in previous research, leads us to believe that order effects were not a major factor in the effects we obtained.

One of the aims of studying individual differences in reasoning is to shed light on the correct normative theory. Stanovich and West (2000) have argued that correlations with ability provide clues as to the correctness of the normative theory. As we have seen, this must be tempered somewhat in the light of the variable correlations found in the present studies. Based on the first two experiments this would lead to the conclusion that the normatively correct response to indicative selection tasks is the matching response, $p + q$, since this is the answer that some of the most able participants gave. Certain reasoning theories, as indicated in the Introduction, do claim that the incorrect normative model has been applied (e.g., Oaksford & Chater, 1994) and hence this finding lends some support to the claims of such theorists. However, if this argument is applied to MT it would cast doubt on the widely held view that MT is a normatively appropriate inference—a conclusion that Oaksford and Chater would endorse but many would challenge. Our findings suggest that a certain amount of caution must be used in interpreting the findings of individual differences in reasoning, especially if the sample does not span the whole range of differences.

REFERENCES

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

Dominowski, R. L., & Dallob, P. I. (1991, September). *Reasoning abilities, individual differences, and the four card problem*. Paper presented to the British Psychological Society Cognitive Section Conference, Oxford, UK.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*, 390–405.

Evans, J. St. B. T. (1995). Relevance and reasoning. In S. E. Newstead & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of Peter Wason* (pp. 147–171). Hove, UK: Lawrence Erlbaum Associates Ltd.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, *11*, 295–306.

Evans, J. St. B. T., Newstead, S. E., Allen, J. L., & Pollard, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology*, *6*, 263–285.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Geis, M., & Zwicky, A. M. (1971). On invited inferences. *Linguistic Enquiry*, *2*, 561–566.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic materials effect in the Wason selection task. *British Journal of Psychology*, *72*, 407–420.

Handley, S. J., Newstead, S. E., & Wright, H. (2000). Rational and experiential thinking: A study of the REI. In R. J. Riding & S. G. Rayner (Eds.), *International perspectives on individual differences* (Vol. 1, pp. 97–113). Stamford, CO: Ablex.

Heim, A. W. (1967). *AH4 group test of intelligence* [Manual]. London: National Foundation for Educational Research.

Heim, A. W. (1968). *AH5 group test of intelligence* [Manual]. London: National Foundation for Educational Research.

Huck, S. W., & Sandler, H. M. (1979). *Rival hypotheses: alternative interpretations of data-based conclusions*. New York: Harper & Row.

Johnson-Laird, P. N. (1993a). *The computer and the mind*. London: Fontana.

Johnson-Laird, P. N. (1993b). *Human and machine thinking*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision-making. *Child development*, *72*, 844–861.

Klaczynski, P. A., Fauth, J. M., & Swanger, A. (1998). Adolescent identity: Rational versus experiential processing, formal operations, and critical thinking beliefs. *Journal of Youth and Adolescence*, *27*, 185–207.

Klaczynski, P. A., Gordon, D. H., & Fauth, J. (1996). Goal-oriented reasoning and individual differences in critical reasoning biases. *Journal of Educational Psychology*, *89*, 470–485.

Klaczynski, P. A., & Narasimham, G. (1998). Problem representations as mediators of adolescent deductive reasoning. *Developmental Psychology*, *34*, 865–881.

Markovits, H. (1984). Awareness of the "possible" as a mediator of formal thinking in conditional reasoning problems. *British Journal of Psychology*, *75*, 367–376.

Markovits, H., Doyon, C., & Simoneau, M. (2002). Individual differences in working memory and conditional reasoning with concrete and abstract content. *Thinking and Reasoning*, *8*, 97–107.

Newstead, S. E. (2000). Are there two different types of thinking? *Behavioral and Brain Sciences*, *23*, 690–691.

Newstead, S. E., Pollard, P., Evans, J. St. B. T., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition*, *45*, 257–284.

Newstead, S. E., Thompson, V. A., & Handley, S. J. (2002). Generating alternatives: A key component in human reasoning? *Memory & Cognition*, *30*, 129–137.

Noveck, I. A. (2000). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*, 165–188.

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

Oaksford, M., Morris, F., Grainger, B., & Williams, J. M. G. (1996). Mood, reasoning, and central executive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 476–492.

O'Brien, D. P., Dias, M. G., & Roazzi, A. (1998). A case study in the mental-models and mental-logic debate: Conditional syllogisms. In M. D. S. Braine & D. P. O'Brien (Eds.), *Mental logic*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*, 972–987.

Rumain, B., Connell, J., & Braine, M. D. S. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults. *Developmental Psychology*, *19*, 471–481.

Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *91*, 497–510.

Schroyens, W., Schaeken, W., Fias, W., & d'Ydewalle, G. (2000). Heuristic and analytic processes in propositional reasoning with negatives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26,* 1713–1734.

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

Stanovich, K. E., & West, R. F. (1998a). Cognitive ability and variation in selection task performance. *Thinking and Reasoning*, *4*, 193–230.

Stanovich, K. E., & West, R. F. (1998b). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*, 161–188.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–665.

Torrens, D., Thompson, V. A., & Cramer, K. M. (1999). Individual differences and the belief bias effect: Mental models, logical necessity, and abstract reasoning. *Thinking and Reasoning*, *5*, 1–28.

Valentine, E. R. (1975). Performance on two reasoning tasks in relation to intelligence, divergence and interference proneness. *British Journal of Educational Psychology*, *45*, 198–205.

Wildman, T. M., & Fletcher, H. J. (1977). Developmental increases and decreases in solutions of conditional syllogism problems. *Developmental Psychology*, *13*, 630–636.

Original manuscript received 17 September 2001
Accepted revision received 3 December 2002
PrEview proof published online 16 June 2003