Confidence and accuracy in deductive reasoning

JODY M. SHYNKARUK and VALERIE A. THOMPSON
University of Saskatchewan, Saskatoon, Saskatchewan, Canada

In two experiments, we investigated the relationship between confidence and accuracy in syllogistic reasoning. Participants judged the validity of conclusions and provided confidence ratings twice for each problem: once quickly and again after further deliberation. Correlations between confidence and accuracy were small or nonexistent. In addition, confidence and accuracy were mediated by different variables. Confidence judgments appeared to reflect external cues, so that confidence was greater when the participants were allowed additional time to think about the problem, as well as when the conclusion was either believable or unbelievable, rather than neutral. In contrast, accuracy changed little as a function of the amount of time available and did not differ for believable and neutral problems. These data support a model in which initial decisions are made quickly, on the basis of heuristic cues, and analytic processes are used to justify or rationalize the earlier decision.

The ability to make accurate confidence judgments is an important element of our ability to introspect regarding our knowledge (Allwood & Granhag, 1996). Adaptive decision making requires the ability to judge the certainty of one's knowledge and to take actions that appropriately reflect that certainty or uncertainty. Surprisingly, however, the evidence from a large number of domains, including vocabulary tests (Stankov & Crawford, 1997), general knowledge questionnaires (Zakay & Tuvia, 1998), recognition memory (Busey, Tunnicliff, Loftus, & Loftus, 2000), free recall (Roediger & McDermott, 1995), and perceptual judgments (Petrusic & Baranski, 2003), indicates that confidence and accuracy are only loosely related.

The purpose of this article is to extend our understanding of the relationship between accuracy and confidence to a novel domain, that of deductive reasoning. Presumably, as in other domains, reasoners make inferences with more or less certainty and take actions that reflect this certainty. Surprisingly, however, relatively little is known about reasoners' confidence in their deductive inferences or about the extent to which this confidence reflects the accuracy and validity of their inferences. In addition, although some theories of reasoning posit confidence as a critical element for the implementation of basic processes, such as belief-based inferences (Klauer, Musch, & Nauwer, 2000; Quayle & Ball, 2000), there is no evidence to

support the view that inferential processes produce uncertainty. Moreover, none of the major theoretical views offers a coherent account of the variables that determine subjective feelings of confidence or uncertainty.

Confidence, Accuracy, and Memory

Given the relative absence of findings in the reasoning domain, we decided to frame our research questions by analogy to another domain. There is an extensive literature on the relationship between confidence and accuracy in memory, from which three general findings have emerged. First, although it is clear that reported confidence poorly reflects the actual accuracy of memory (Leippe, 1980; Speelman & Kirsner, 1997; Stankov & Crawford, 1997), these variables nonetheless tend to be correlated. For example, it is common to report small but statistically significant correlations between participants' accuracy and their self-assessed confidence (e.g., Deffenbacher, 1980; Koriati, Lichtenstein, & Fischhoff, 1980; Sporer, Penrod, Read, & Cutler, 1995), although clearly there is large variability, due to task demands and testing conditions (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Koehler, 1994; Leippe, 1980; Sporer et al., 1995).

Second, when confidence and accuracy diverge, it tends to be in the direction of overconfidence, rather than underconfidence (Busey et al., 2000; Butterfield & Metcalfe, 2001; Koriati et al., 1980; Pulford & Colman, 1996; Stankov & Crawford, 1997; Zakay & Tuvia, 1998). Butterfield and Metcalfe attributed this overconfidence to the fact that high-confidence items are more easily retrieved from memory than are low-confidence items, even if they are incorrect. This overconfidence effect is more pronounced for unskilled participants than for expert participants (Kruger & Dunning, 1999) and tends to be associated with difficult items (Baranski & Petrusic, 1999; Gigerenzer et al., 1991) and items subject to large biases (Baranski & Petrusic, 1999).

Finally, there is evidence to suggest that confidence and accuracy may be sensitive to different variables, so that

This research was supported by the Natural Sciences and Engineering Research Council of Canada. The authors thank Jamie Campbell, Jonathan Evans, Janeen Loehr, Mike Oaksford, Kealee Playford, and two anonymous reviewers for the thoughtful comments they provided on an earlier version of the manuscript. Correspondence regarding this manuscript should be sent to V. A. Thompson, Department of Psychology, University of Saskatchewan, 9 Campus Drive, Saskatoon, SK, S7N 5A5 Canada (e-mail: valerie.thompson@usask.ca) or to J. M. Shynkaruk (e-mail: jody.shynkaruk@usask.ca).

Note—This article was accepted by the previous editorial team, when Colin M. MacLeod was Editor.

factors that increase accuracy do not necessarily increase confidence and vice versa. Busey et al. (2000) found that for face recognition, longer study durations, more rehearsal, and higher luminance all improved the participants' confidence in their memories. These variables, however, had different effects on their recognition accuracy, so that, for example, luminance actually decreased accuracy. Similarly, eyewitness confidence is increased by such variables as reinforcement and repeated questioning, although these variables do not produce concomitant increases in accuracy (see Wells, Olson, & Charman, 2002, for an overview). The fact that confidence and accuracy can be affected differently by different variables suggests that confidence ratings are based not only on information from memory, but also on the participant's consideration of aspects of the test and study conditions.

Extension to Deductive Reasoning

By analogy, there are three general questions one can ask about the relationship between confidence and accuracy in reasoning: (1) Are confidence and accuracy correlated? (2) Do reasoners tend to be over- or underconfident? (3) Do variables that increase confidence increase accuracy and vice versa? To answer these questions, we employed a commonly used paradigm—namely, syllogistic reasoning. Syllogisms are problems consisting of two premises and a conclusion, each of which contains one of the qualifiers *all*, *no*, *some*, or *some . . . not* (e.g., Newstead, Handley, & Buck, 1999). The middle term is repeated in the two premises, and the participants' task is to draw a conclusion linking the other two terms. A conclusion is valid if it follows necessarily from the premises, regardless of the believability of either the premises or the conclusion (see, e.g., Galotti, Komatsu, & Voelz, 1997; Thompson, 1996).

On the basis of the few studies available, we expected that confidence and accuracy would be poorly correlated on this task: When confidence has been measured directly, reasoners have consistently exhibited high confidence for erroneous deductions (e.g., Johnson-Laird & Savary, 1999; Quayle & Ball, 2000). For example, Johnson-Laird and Savary found that participants expressed equal confidence in correct and incorrect responses, which indicates that a high degree of confidence does not necessarily reflect accurate reasoning. We also expected that reasoners would tend to be overconfident, rather than underconfident, in their responses. Kruger and Dunning (1999) found that unskilled or novice participants tended to be more overconfident than did expert or experienced participants. Because syllogisms are difficult to solve, and because the participants in the present study had no prior experience with logic, it was reasonable to expect that the participants would be overconfident with these problems.

Time, Accuracy, and Confidence

As is the case for memory, we anticipated that confidence and accuracy of reasoning decisions would be affected by different variables. In particular, time to recon-

sider an initial decision should affect confidence but not accuracy. This prediction was derived from data obtained from the Wason selection task, which has demonstrated that much of the time spent reasoning about a problem appears to involve rationalizing, rather than reconsidering, one's first choice (Ball, Lucas, Miles, & Gale, 2003; Evans, 1996). In these studies, participants were asked to indicate which of several alternatives they were considering at a given time; the reasoners spent most of their time considering alternatives that they were going to choose and spent little time considering alternatives that they were going to reject. Evans (1996) argued that the reasoners' initial decisions were guided by heuristic cues and that the time they spent considering each alternative was used to rationalize those initial decisions, even though those responses were often logically invalid.

Although Evans's (1996) hypothesis has not previously been extended to syllogistic reasoning, there is nonetheless good reason to draw parallels to the Wason selection task. Specifically, for the type of syllogisms used here, there is little evidence that reasoners search for alternatives to their initial conclusion (Evans, Handley, & Harper, 2001; Evans, Legrenzi, & Girotto, 1999; Klauer et al., 2000; Polk & Newell, 1995; Thompson, Streimer, Reikoff, Gunter, & Campbell, 2003). Thus, even though the opportunity to reconsider one's initial response should allow reasoners to engage in the type of decontextualized, analytic thinking that allows one to overturn heuristic or belief-based choices (Chater & Oaksford, 1999; Evans, 2003; Evans & Over, 1996), there is little evidence to suggest that reasoners are inclined to engage that type of reasoning on this particular task. Indeed, Chater and Oaksford suggested that the ability to test the validity of a putative conclusion is not well developed in most participants.

Although the participants were not predicted to change their responses often, when they did change, we predicted that they would have a tendency to accept more conclusions on their second than on their first response opportunity. This prediction was based on the hypothesis that reasoners, when presented with a conclusion, attempt to formulate a model that integrates the conclusion with the premises (Klauer et al., 2000; Polk & Newell, 1995; Thompson et al., 2003). In other words, participants search for a model in which the conclusion is consistent with the premises. When this attempt is successful, reasoners accept the conclusion. Thus, allowing extra time to reason about the problem should increase the probability that the attempt at integration is successful and should result in an overall increase in the number of accepted conclusions, regardless of whether the conclusion is valid or not.

In contrast, whereas time was predicted to have little effect on accuracy, it was predicted to have a large effect on confidence. Specifically, if participants use extra time to rationalize their initial choice, their confidence should increase between the first and the second response opportunities. Indeed, when forced to make a decision under uncertain conditions, reasoners tend to look for ways to rationalize or explain that decision (e.g., Shafir, Simon-

son, & Tversky, 1993), a process that increases their confidence in the decision (see Koehler, 1991, for a review) by, for example, focusing attention on evidence that supports the chosen answer at the expense of contradictory evidence (Koriat et al., 1980). Consequently, although responses should tend to stay the same over time, confidence should increase, especially when the initial answer does not change.

Believability, Accuracy, and Confidence

In order to ensure sufficient variability in accuracy, the believability of the conclusions was varied. Believable and unbelievable conclusions are known to produce large differences in rates of accuracy. Specifically, reasoners are more accurate reasoning about unbelievable than about believable conclusions (e.g., Evans, Barston, & Pollard, 1983; Evans, Newstead, Allen, & Pollard, 1994; Klaczynski & Robinson, 2000; Klauer et al., 2000; Newstead, Pollard, Evans, & Allen, 1992; Thompson et al., 2003), so that when the conclusions are unbelievable, reasoners discriminate valid from invalid conclusions. However, believable conclusions tend to be accepted regardless of validity. Neutral conclusions are accepted at the same rate as believable conclusions (Evans et al., 1994). This pattern is thought to arise from a process of disconfirmation, whereby participants accept believable and neutral conclusions with little critical analysis but search for an alternative when a conclusion is unbelievable. This results in more logical analysis of unbelievable than of believable conclusions, with the effect that reasoners discriminate more between logically valid and invalid conclusions when the conclusions are unbelievable.

How would believability affect confidence? Although current theories of reasoning do not specifically address the relationship between confidence and beliefs, we can make tentative hypotheses on the basis of the literature from other areas. There are at least two reasons why confidence might differ as a function of the believability of the conclusion. First, belief effects are ubiquitous in all forms of human reasoning (see Nickerson, 1996, for a review); thus, it seems likely that reasoners will be more confident in their decisions when they are able to rely on a cue, such as believability, that informs so many of their decision processes. That is, familiar cues tend to produce the feeling that one knows what the answer will be (Chandler, 1994; Metcalfe, 1994; Miner & Reder, 1994). As a result, participants should be more confident when familiar cues, such as belief, are available than when such cues are absent. Second, believable and unbelievable conclusions are unambiguously true or false, relative to neutral conclusions. Given that unambiguous stimuli produce high levels of confidence (McCarthy, Turner, Oakes, & Haslam, 1993; Peterson & Pitz, 1988), one might expect higher levels of confidence for decisions guided by belief-based cues than for ones made in the absence of such cues. In summary, on the assumption that belief-based cues are familiar and unambiguous, they should produce higher confidence than do situations in which such cues are absent, such as ones with neutral conclusions.

An additional prediction is possible in the case of unbelievable conclusions. As was explained above, these conclusions are thought to motivate reasoners to search for alternatives; however, given the difficulty of the task, reasoners may be aware that they need to reformulate their representation of the premises but be unable to do so, due to working memory constraints (Johnson-Laird & Byrne, 1991). This state of uncertainty may result in lower confidence when reasoning about unbelievable conclusions than when reasoning about believable ones (Quayle & Ball, 2000), even though they are more accurate with the former than with the latter.

Summary

The aim of the present research was to study the relationship between confidence and accuracy in deductive reasoning. Overall, we predicted that confidence and accuracy would be poorly calibrated and that the participants would express a greater degree of confidence than would be justified by the accuracy of their answers. We also predicted that confidence and accuracy would be mediated by different variables. Specifically, (1) accuracy would differ little between the first and the second responses, because the participants were expected to make their decisions early and use extra time more for rationalization than for a reevaluation of their initial choice; (2) the participants would be more confident for their second, reevaluated answer than for their first answer, because they would have more time to rationalize this response; (3) the participants would be more confident when they could reason on the basis of prior knowledge than when such cues were absent, so that higher confidence was expected for unbelievable and believable than for neutral problems; (4) the participants would be more confident in their responses to believable problems than in those to unbelievable problems, because unbelievable conclusions should produce a state of metacognitive uncertainty in the participants; and (5) the participants would be more accurate for unbelievable problems than for believable problems, because the former were more likely to be logically analyzed.

EXPERIMENT 1

In this experiment, the participants were provided with 12 syllogisms that varied orthogonally with respect to the believability and validity of their conclusions. The participants were asked to provide two answers for each syllogism. The first answer was to be made within a 10-sec deadline. The participants were then provided up to 1 min to rethink and possibly change their initial answer.

The goal of the deadline was to allow sufficient time to read and comprehend the premises and conclusions but, nonetheless, to require the participants to respond with the first answer that came to mind. To this end, we computed the mean response time (RT) for each of the 128 participants from an earlier study (Thompson et al., 2003) in which the same syllogisms were employed ($M = 20.7$ sec) and subtracted a time approximately equal to one standard deviation. Fewer than 20% of the responses

were made in less than 10 sec; thus, posing a 10-sec deadline should obligate the participants to respond before they have completed a full logical analysis of the problem. In order to provide sufficient time to complete this full analysis for their second response, the participants were then allowed an additional minute to reconsider their decision (note that in Thompson et al.'s, 2003, study, 98% of the responses were made in less than 1 min, indicating that 1 min should be ample time for reasoners to reconsider their initial option).

Finally, we divided the sample into *good* and *poor* reasoners on the basis of their overall accuracy. We did this because a large percentage of our sample performed essentially at chance levels, suggesting that many of the participants may have been guessing, rather than trying to reason through the problems. However, a valid test of our predictions required that the participants have an adequate opportunity to engage in analytical thinking. Given that individual differences in deductive reasoning are associated with the ability to generate alternative representations or models of the premises (e.g., Newstead, Thompson, & Handley, 2002; Thompson et al., 2003; Torrens, Thompson, & Cramer, 1999), it was possible that only the good reasoners would have the inclination and/or the ability to reanalyze their initial conclusions and show an increase in performance between their first and second response opportunities.

Method

Participants.¹ The participants (74 women and 54 men) were recruited from the University of Saskatchewan campus and received either \$5 or a research-related course credit. The mean age was 22.76 years ($SD = 6.59$).

Materials. Twelve logical reasoning problems (see the Appendix) were presented on a high-resolution computer monitor. Four of the problems led to believable, unbelievable, and neutral conclusions, respectively. Within each conclusion type, two problems were valid, and two were invalid. Problems appeared on the monitor one at a time; the premises and conclusion were presented simultaneously. The two response options, *yes* and *no*, appeared below each problem. The participants selected their response by choosing either the 1 or the 3 on the computer keyboard. The number corresponding to each response alternated across participants, so that for half of the participants 1 corresponded to *yes*, and for the remainder, 1 corresponded to *no*. After a response had been given, a 7-point confidence scale appeared, where 1 corresponded to *not at all confident* and 7 corresponded to *extremely confident*. The keys on the keyboard corresponded to the numbers on the scale.

Each problem consisted of two premises, followed by a conclusion of the form "Some ___ are not ___." One premise was of the type "No ___ are ___," and one premise was of the type "Some ___ are ___." There were two conclusion forms ("Some C are not A" and "Some A are not C"). The believability of the conclusions was established in a rating study by Evans et al. (1983).

For the believable and unbelievable problems, content was assigned to the A, B, and C terms in such a way that the A and C terms referred to familiar categories (e.g., judges, police dogs), and the B term was a nonsense term (e.g., Rewons, Likels). For the neutral problems, the A and C terms were nonsense terms, and the B term referred to familiar categories. This was done in order to avoid any effects that might be attributed to the believability of the premises (Thompson, 1996). The believable and unbelievable problems were identical to those used by Thompson et al. (2003). The neutral

problems were created for the purposes of this experiment and were of the same syntactic structure as the believable and unbelievable problems.

Procedure. The participants were tested individually. The following instructions, adapted from Quayle and Ball (2000), were presented on a computer screen:

This experiment is designed to examine how people solve logical reasoning problems. Your task is to decide whether the conclusion given at the end of the problem follows logically from the information given within that problem.

For each problem, we want you to provide two answers. The first answer should be the first answer that comes to mind, so the answer that is your first inclination or first instinct. You will be given a 10-second deadline in which to give your first response. If you do not answer before this deadline, you will hear a "beep" instructing you to answer immediately. Your second answer should be your final answer. For this answer, you will be given up to 1 minute to fully think through the question. Make sure that you have the logically correct answer before giving your final answer. To choose an answer, please press either '1' or '3' on the keyboard. You must not make any notes or diagrams of any kind to aid you in this task.

To answer each question, you must assume that all information which you are given is true; this is very important. If, and only if, you judge that a given conclusion logically follows from the information given should you choose "YES." If you think that the conclusion given does not necessarily follow from the information given you should choose "NO." If you have any questions, please ask now.

After the instructions, the participants were given one practice problem. All of the participants received the same practice problem, and they were given as much time as they required to familiarize themselves with the problem before the timed trials began. The practice problem was a neutral syllogism containing a "Some C are not A" conclusion. The participants were also provided with feedback following the practice problem. The importance of logical necessity was reiterated, and the participants then evaluated the remaining 12 syllogisms. The participants pressed the space bar to display each problem. If the first response was not made within the 10-sec deadline, there was an audible beep from the computer, and a statement flashing "please respond now" appeared on the screen. Timing of the first response began when the problem appeared and ended when the *yes/no* response was given. After the initial response had been given, the 7-point confidence scale appeared.

After the confidence rating had been entered, the initial response and confidence rating disappeared from the screen and were replaced by the question "What is your final answer?" The participants were given up to 1 min to reevaluate their initial decision. Timing for the second response began with the onset of this question and ended when the participants entered their response. The participants were then prompted to provide a second confidence rating. After completing all 12 problems, the participants were asked to state how many of these problems they believed they had answered correctly. Testing lasted an average of 15 min.

Counterbalancing. For counterbalancing, the 12 syllogisms were divided into three blocks of 4 syllogisms. Two of these blocks consisted of problems leading to believable or unbelievable conclusions. Each of these blocks consisted of one problem in each belief by validity cell. The order of these conditions was repeated across blocks for each participant, and this order was counterbalanced across participants, using a Latin square.

The third block of four problems led to neutral conclusions. Two of the neutral problems were valid, and two were invalid. The order of these four problems was counterbalanced, using a Latin Square. The arrangement of the block of neutral problems within the believable and unbelievable blocks was also counterbalanced, so that they appeared first for half of the participants and last for the remaining participants.

Content was assigned to the problems so that no problem content was repeated during the 12 problems. Each of the problem contents

was assigned equally often to each of the conditions. Conclusion type ("Some C are not A" and "Some A are not C") was also counterbalanced across conditions, so that each conclusion type occurred equally often with both valid and invalid conclusions and equally often with believable, unbelievable, and neutral conclusions across blocks.

Results and Discussion

Acceptance of conclusions. Conclusions were scored as accepted when the participants responded *yes* and were scored as rejected when the participants responded *no*.² These data are presented in Table 1. Overall, the participants accepted 61% of the conclusions, providing evidence for a strong *yes* bias [$t(127) = 8.27, p < .001$]. Between the first and the second response opportunities, the participants changed 28% of their answers (mean number of changes per participant = 3.4 out of 12 problems). Of these changes, more were changed from *no* to *yes* (16%) than from *yes* to *no* [12%; $t(83) = 29.74, p < .001$]. This is consistent with our hypothesis that participants attempt to find a model that integrates the premises and the conclusion and that affording more time increases the probability that this process will be successful. Our findings are also consistent with the hypothesis that valid conclusions are easier to integrate than invalid conclusions (Thompson et al., 2003): The participants were more likely to change their response to *yes* (19.4%) than to *no* (7.4%) when the conclusions were valid; when the conclusions were invalid, there was no difference in answer changes to *no* (13.3%) or to *yes* (13.9%). Finally, the only problems for which the answer was more likely to change to *no* (16.8%) than to *yes* (8.6%) were those with invalid, unbelievable conclusions.

Response accuracy. As has been pointed out above, the participants in the present experiment exhibited a *yes* bias; consequently, *yes* responses to valid problems do not necessarily represent a correct judgment about the logical status of the conclusion. Therefore, to analyze accuracy, we computed a logic index similar to that used by Evans et al. (1994); accuracy was operationalized as the proportion of valid conclusions accepted minus the proportion of invalid conclusions accepted (% accept valid – % accept invalid). This resulted in a possible range of scores from

–1 to +1, where –1 represented perfect inaccuracy (i.e., responding *no* to valid syllogisms and *yes* to invalid syllogisms) and +1 represented perfect accuracy (i.e., responding *yes* to valid syllogisms and *no* to invalid syllogisms). A score of 0 represented an inability to discriminate valid and invalid conclusions.

The sample was divided into *good* and *poor* reasoners on the basis of a median split on the accuracy scores. The left panel of Figure 1 presents the mean accuracy scores as a function of conclusion believability, response opportunity, and reasoning ability. Accuracy was analyzed using a $3 \times 2 \times 2$ (belief $\times$ response opportunity $\times$ reasoning ability) repeated measures ANOVA. In this and all the subsequent analyses, when the assumption of sphericity was not met for the ANOVA, the *dfs* were adjusted using the Greenhouse–Geisser correction.

Not surprisingly, there was an effect of reasoning ability [$F(1,126) = 207.66, MS_e = 0.14, p < .001$], so that good reasoners ($M = .39$) were more accurate than poor reasoners ($M = .00$). In order to determine whether the participants were performing above chance levels, we examined whether accuracy was significantly in excess of zero, which corresponds to chance performance. One-sample *t* tests revealed that whereas the poor reasoners' accuracy scores did not differ from chance [$t(63) = -0.07, p = .95$], the good reasoners were performing well above chance levels [$t(63) = 20.74, p < .001$].

We also observed the standard effect of belief [$F(2,252) = 14.84, MS_e = 0.26, p < .001$], whereby the participants were more accurate for unbelievable conclusions ($M = .336$) than for believable conclusions [$M = .10; t(127) = 5.45, p < .001$] and neutral conclusions [$M = .15; t(127) = 3.70, p < .001$]. There was no difference in accuracy between the believable and the neutral conclusions [$t(127) = -0.96, p = .34$]. These main effects were qualified by an interaction between reasoning ability and belief [$F(2,252) = 6.52, MS_e = 0.26, p < .01$]: The poor reasoners were more accurate for unbelievable [$t(63) = 3.29, p < .01$] and neutral [$t(63) = 2.84, p < .01$] conclusions than for believable conclusions, with no difference in accuracy between the former [$t(63) = 0.49, p = .63$]. In contrast, the good reasoners showed the

Table 1
Mean Proportion of Conclusions Accepted in Experiments 1 and 2, as a Function of Validity, Believability, and Response Opportunity (With Standard Errors)

Believability	Response 1				Response 2			
	Valid		Invalid		Valid		Invalid	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Experiment 1								
Believable	.70	.03	.66	.03	.81	.02	.65	.03
Unbelievable	.58	.03	.37	.03	.74	.03	.29	.03
Neutral	.65	.03	.56	.03	.75	.03	.55	.04
Experiment 2								
Believable	.83	.03	.67	.03	.88	.02	.69	.03
Unbelievable	.54	.04	.33	.04	.61	.04	.37	.04
Neutral	.78	.03	.70	.03	.78	.03	.76	.03

Figure 1. Mean accuracy, confidence, and response times in Experiment 1 as a function of conclusion believability, response opportunity, and reasoning ability. Error bars reflect the standard error of the mean.

more typical pattern (Evans et al., 1994) and were more accurate for unbelievable conclusions than for believable [$t(63) = 4.43, p < .001$] and neutral [$t(63) = 5.08, p < .001$] conclusions, with no difference in accuracy between the latter [$t(63) = 1.28, p = .21$].

Contrary to our initial prediction, the participants were slightly more accurate on their second response opportunity than on their first response opportunity [mean increase = .15; $F(1,126) = 32.88, MS_e = 0.13, p < .001$]. Response opportunity also interacted with belief [$F(2,252) = 3.32, MS_e = 0.12, p < .05$]. Specifically, the participants' accuracy increased from their first to their second response opportunity for all the problems, but this increase was greater for unbelievable conclusions than for either believable [$t(127) = 2.08, p < .05$] or neutral [$t(127) = 2.36, p < .05$] conclusions; the increase in accuracy was comparable for believable and neutral problems [$t(127) = 0.26, p = .80$].³ Note that the unbelievable conclusions were the only conclusions for which the participants were more likely to change their answers from *yes* to *no* than vice versa, suggesting that accuracy on this task required overcoming the participants' predisposition

to confirm conclusions (i.e., by providing them conclusions they were inclined to reject).

Self-reported confidence. In order to make the analyses of the relative confidence ratings comparable to the accuracy scores, the confidence scores were averaged over valid and invalid problems. The center panel of Figure 1 presents the mean confidence ratings as a function of conclusion believability, response opportunity, and reasoning ability. The data were analyzed using a $3 \times 2 \times 2$ (belief $\times$ response opportunity $\times$ reasoning ability) repeated measures ANOVA. Despite the large differences in accuracy, good ($M = 4.82$) and poor ($M = 4.75$) reasoners were equally confident [$F(1,126) = 0.17, MS_e = 4.68, p = .68$]. A power analysis (Campbell & Thompson, 2002) indicated an 80% power to detect a difference as small as .13 between good and poor reasoners.

There was an effect of belief on confidence [$F(1.91,240.74) = 12.54, MS_e = 0.83, p < .001$]. Follow-up t tests revealed that the participants were more confident for both believable [$M = 4.84; t(127) = 3.41, p = .001$] and unbelievable [$M = 4.95; t(127) = 4.53, p < .001$] conclusions than for neutral conclusions ($M = 4.56$). These

data are consistent with our prediction that the participants would be more confident when they could base their decisions on preexisting knowledge and that confidence would be lowest for neutral problems, for which such cues were absent. They are also consistent with the hypothesis that confidence and accuracy would be affected differently by the believability of the conclusions: Whereas the participants were equally confident reasoning about believable and unbelievable conclusions, they were more accurate with the latter, and whereas they were equally accurate when reasoning about believable and neutral conclusions, they were more confident reasoning about the former. Contrary to the metacognitive uncertainty hypothesis, the participants were equally confident for believable and unbelievable conclusions [$t(127) = -1.52, p = .13$]. A power analysis revealed that we had 80% power to find a difference greater than or equal to .23.

Consistent with our predictions, there was also an effect of response opportunity on confidence [$F(1,126) = 398.02, MS_e = 1.28, p < .001$], so that the participants were more confident in their second responses ($M = 5.49$) than in their first responses ($M = 4.93$). Given that the reasoners' accuracy also increased over this interval, it is possible that the increase in confidence reflected accurate metacognition on behalf of the reasoners; this hypothesis would predict that increases in confidence should be tied to increases in accuracy. Alternatively, it is possible that the participants used the extra time on their second response in an attempt to rationalize and justify their initial answer and that this rationalization resulted in increased confidence. If the latter hypothesis were true, one would expect that increases in accuracy and confidence across trials would be independent of one another.

In accord with the latter hypothesis, we observed that the increases in confidence and accuracy occurred in different belief conditions. Specifically, whereas accuracy increased more for unbelievable than for believable conclusions, confidence increased equally for believable and unbelievable conclusions [$t(122) = 0.85, p = .40$]. Whereas accuracy increased equally for neutral and believable problems, the increase in confidence for neutral problems was less than that for either the believable [$t(122) = 2.99, p < .01$] or the unbelievable [$t(122) = 1.96, p = .05$] problems. These data therefore support the hypothesis that changes in confidence are not tied to changes in accuracy and are consistent with the conclusion that the participants spent the extra time justifying their initial answer.

Further support for this interpretation comes from a second analysis, in which we examined change in confidence as a function of change in accuracy. For this analysis, we computed the mean difference in confidence between the first and the second response opportunities for each of three accuracy change conditions (accuracy increased, decreased, or stayed the same). Confidence increased from the first to the second response to the same degree regardless of whether the participants' accuracy increased (mean increase = 1.87) or stayed the same [mean increase =

1.64; $t(86) = 1.39, p = .17$] and regardless of whether accuracy increased or decreased [mean increase = 2.23; $t(77) = 1.64, p = .12$]. Ironically, confidence increased more for trials on which accuracy *decreased* from the first to the second response opportunity [$t(109) = 4.59, p < .001$] than when accuracy stayed the same. Clearly, the omnibus increase in both confidence and accuracy was coincidental and was driven by different variables: On a more detailed analysis, it is apparent that increases in confidence and accuracy occurred independently and on different trials. These data are consistent with predictions from the heuristic-analytic approach: When the participants were able to rationalize their initial answer, confidence was higher than when they changed their answer, even if the change did not result in greater accuracy.

Although accuracy did not appear to be linked to confidence, it is possible that confidence may be linked to the reasoning process in another way. Specifically, it is possible that confidence may drive answer change, rather than vice versa. In other words, a participant's decision to change his or her answer may depend on his or her initial confidence in that answer. To test this possibility, the base confidence rates (i.e., the confidence for the first response opportunity) were examined and compared for trials on which the participants changed their answers and those on which they kept their answers the same. Indeed, there was a significant difference in base confidence ratings between these conditions [$F(2,154) = 14.70, MS_e = 1.00, p < .001$], so that initial confidence was higher on the trials for which the participants later kept the same answer ($M = 4.00$) than when they changed their answers, regardless of whether the answer changed to the correct answer [$M = 3.42; t(86) = 4.09, p < .001$] or to the incorrect answer [$M = 3.12; t(109) = 7.56, p < .001$]. These data suggest that the decision to change the answer depended, at least in part, on the participants' initial confidence in their answer: Answers were more likely to be changed when initial confidence was low, regardless of whether the initial answer was correct or incorrect.

Finally, in order to determine whether the participants were generally overconfident, relative to their actual accuracy, we examined whether confidence was significantly in excess of the *not at all confident* point on the scale. This analysis was done only for the poor reasoners, for two reasons: (1) As was discussed in the introduction, overconfidence is generally found for less skilled participants (Kruger & Dunning, 1999); and (2) good reasoners were performing well above chance levels, and their high levels of confidence were reflective of this, whereas the poor reasoners were performing very poorly, with the same high levels of confidence. We found that poor reasoners rated their confidence significantly in excess of the *not at all confident* point on the scale, even though their overall performance was at chance levels [$t(63) = 32.60, p < .001$].

Conclusions from accuracy and confidence analyses. It is clear that confidence and accuracy in syllogistic reasoning were not mediated by the same variables. Peo-

ple were more confident reasoning about conclusions that had a belief value (i.e., were either believable or unbelievable) than when reasoning about neutral conclusions but were more accurate only for unbelievable conclusions. Although both confidence and accuracy increased from the first to the second response opportunity, the increase in confidence was independent of the increase in accuracy. Confidence, therefore, did not reflect accurate introspection about the outcome of the reasoning process. However, confidence appeared to influence the outcome of the process, in that reasoners seemed more likely to change their answers when they had low confidence in their first answer, regardless of whether this answer was correct.

Response time data. The right panel of Figure 1 presents the participants' mean RTs as a function of conclusion believability, response opportunity, and reasoning ability; these data were analyzed using a $3 \times 2 \times 2$ (belief $\times$ response opportunity $\times$ reasoning ability) repeated measures ANOVA. RT was analyzed primarily to ensure that the participants were adhering to the 10-sec deadline; a detailed analysis of these data will not be reported. The analysis indicated that the mean RT for the first response opportunity was 9.89 sec ($SE = 0.40$). Thus, initial responses were made within the deadline, for the most part. The results also indicated that the participants spent more time on their second response opportunity than on their first response [mean increase = 9.09 sec; $F(1,126) = 66.04$, $MS_e = 240.32$, $p < .001$], indicating that they were reevaluating, or at least reconsidering, their initial answer.

Interestingly, for our examination of good and poor reasoners, there was no effect of reasoning ability [$F(1,126) = 0$, $MS_e = 248.95$, $p = .98$]. In other words, good and poor reasoners spent the same amount of time reasoning about the problems. This indicates that the poor reasoners were likely not just guessing. Instead, it seems that they were simply using a less successful strategy than were the good reasoners. These data are consistent with the hypothesis that the amount of time spent reasoning about problems is not necessarily tied to accuracy, in that good and poor reasoners spent the same amount of time thinking about the conclusions but the good reasoners were much more accurate than the poor reasoners.

Correlations between accuracy and confidence.

A second goal of the present experiment was to explore whether confidence and accuracy are correlated, so that higher confidence ratings reflect higher accuracy scores. To do so, a one-tailed bivariate correlation of confidence and accuracy was performed using the mean scores for each participant; confidence was not related to accuracy for either good or poor reasoners [$r(56) \leq .14$, $p \leq .30$, one-tailed]. When this correlation was computed separately for each of the 24 conditions in the experiment (3 belief $\times$ 2 validity $\times$ 2 responses $\times$ 2 blocks of trials), the mean correlation between confidence and accuracy was marginally different from 0 on a one-tailed test [$r = .06$; $t(23) = 1.63$, $p = .058$].

We also computed a correlation between each participant's mean accuracy score and his or her estimate of the

number of problems (out of 12) that he or she believed had been answered correctly. The data in response to this question were not available for 16 participants, because of an error in the computer program. The responses for the remaining participants were converted into proportions; their accuracy scores were also converted into proportions of correct responses (i.e., the number of correct responses divided by 12), in order to be comparable to the aggregate proportions. Good and poor reasoners differed in terms of how well calibrated they were. For the poor reasoners, the relationship between their aggregate assessments and actual accuracy was not significant [$r(56) = .03$, $p = .41$, one-tailed]. For the good reasoners, this relationship was significant [$r(56) = .35$, $p < .01$, one-tailed].

Finally, we compared the number of problems estimated to be correct to the number of problems each participant solved correctly. As was the case with the confidence ratings, the good and the poor reasoners did not differ in estimated accuracy ($M = .54$ and $.55$, respectively; $t < 1$). The good reasoners, who solved, on average, 70% of the problems correctly, significantly underestimated their performance [$t(55) = 5.12$, $p < .001$], whereas the estimates made by the poor reasoners were comparable to their overall accuracy of 50% [$t(55) = 1.86$, $p = .07$]. In other words, whereas the good reasoners underestimated their overall performance, the estimates of the poor reasoners were in line with their actual accuracy.

Collectively, these findings suggest that the good reasoners were sensitive to the overall level of task difficulty: Those who thought that they had done well tended to do better than those who thought that they had not. They also tended to underestimate the number of problems solved correctly, again suggesting that they recognized the difficulty of the task. Neither the good nor the poor reasoners, however, were able to provide an accurate assessment of their performance on a trial-by-trial basis, suggesting that the mechanisms that give rise to feelings of confidence are distinct from those that produce accurate reasoning behavior.

EXPERIMENT 2

Before discussing the broader theoretical implications of our findings, we first sought to replicate them with a more traditional methodology (i.e., the paper-and-pencil evaluation task). Once again, the participants were asked to provide two responses and two confidence judgments. The participants were tested in a single group and were, therefore, not timed. Except as noted below, the procedure was identical to that in Experiment 1.

Method

Participants. One hundred twenty-three introductory psychology students from the University of Saskatchewan were tested in a single testing session during their regular class time. The students participated in partial fulfillment of a course requirement.

Materials and Procedure. The same 12 problems as those in Experiment 1 were used. The problems were also counterbalanced in the same way as in Experiment 1. The syllogisms were presented in booklet form; on each page, the two premises and one conclusion

were followed by the two response options: *yes* and *no*. This was followed by the 7-point scale, where 1 corresponded to *not at all confident* and 7 corresponded to *extremely confident*.

Testing was completed in a single session. The following instructions, which were modified from Experiment 1, were presented on the second page of the test booklets:

This experiment is designed to examine how people solve logical problems. On the following pages are twelve problems, one on each page. For each problem, you will complete two tasks. These two tasks are described below.

1. Does the conclusion follow?

The first task is to decide whether the conclusion given follows logically from the information (premises) given within each problem. In order to answer the problems, you must assume that all information you are given is true; this is very important. If, and only if, you judge that a given conclusion follows logically from the information given will you circle "YES." If you think the conclusion does not follow logically from the information given, you should circle "NO." Do not draw any diagrams or make any other kinds of notes to help you make your decisions.

2. Confidence Ratings

The second part of the task asks that you rate your confidence in the correctness of your answer. A seven-point rating scale, similar to the one below, will follow each problem that you solve.

Not at all Confident 1 2 3 4 5 6 7 Extremely Confident

You would circle 7 if you were very confident that your answer was correct, 4 if you were not sure, and 1 if you were very confident that your answer was incorrect.⁴ Thank you very much for participating.

Testing lasted an average of 30 min.

Results and Discussion

Acceptance of conclusions. Conclusion acceptance was examined in order to verify the *yes* bias found in Experiment 1 and to look at the pattern of answer changes from the first to the second response opportunity. These data are presented in Table 1. Replicating our results from Experiment 1, the participants exhibited a strong *yes* bias, so that 66% ($SE = 0.01$) of the conclusions were accepted overall. In comparison with Experiment 1, the participants changed their answers less often (i.e., 12.5% of the answers were changed in the present experiment, in comparison with 28% in Experiment 1). Once again, how-

ever, the participants were much more likely to change their answers to *yes* than to *no* on their second response opportunity. Specifically, 63% of answers were changed from *no* to *yes*, and 37% were changed from *yes* to *no*. Unlike in Experiment 1, this trend was also found for the invalid, unbelievable problems.

Response accuracy. As in Experiment 1, accuracy was operationalized as the difference in the rates of acceptance between valid and invalid conclusions. The left panel of Figure 2 presents the mean accuracy scores as a function of conclusion believability and reasoning ability. When the accuracy data were examined, it became obvious that the overall accuracy scores were much lower than those found in Experiment 1 ($M = .15$, $SE = .02$). A median split was again performed to separate the participants into good ($n = 65$) and poor ($n = 58$) reasoners. The accuracy data were analyzed using a $2 \times 2 \times 3$ (reasoning ability $\times$ response opportunity $\times$ belief) repeated measures ANOVA.

Not surprisingly, an effect of reasoning ability was found once again [$F(1,120) = 284.24$, $MS_e = 0.11$, $p < .001$], so that the good reasoners were more accurate than the poor reasoners. Reasoning ability also interacted with belief [$F(2,240) = 8.24$, $MS_e = 0.26$, $p < .001$]. The form of this interaction suggests that only the good reasoners evidenced differences in accuracy between belief conditions. The good reasoners were more accurate for unbelievable problems than for believable problems [$t(64) = 2.08$, $p < .05$] and were more accurate for believable problems than for neutral problems [$t(64) = 3.30$, $p < .01$]. However, there was no difference in accuracy across the different belief conditions for the poor reasoners [$t(57) \leq 0.91$, $p \geq .37$]. In fact, as Figure 2 shows, the poor reasoners' accuracy was at about chance levels regardless of the belief condition.

Consistent with our predictions, but contrary to Experiment 1, there was no effect of response opportunity on accuracy [$F(1,120) = 0.06$, $MS_e = 0.07$, $p = .81$]. In other words, additional time to solve the problems did

Figure 2. Mean accuracy and confidence in Experiment 2, as a function of conclusion believability, response opportunity, and reasoning ability. Error bars reflect the standard error of the mean.

not result in higher accuracy. A power analysis revealed that we had an 80% power to find a difference greater than or equal to .07 between the first and the second response opportunities. Thus, it is 80% likely that the real difference in accuracy between the first and the second response opportunities was less than or equal to .07. Finally, response opportunity did not interact with reasoning ability [$F(1,120) = 0.28$, $MS_e = 0.07$, $p = .60$] or belief [$F(2,240) = 1.99$, $MS_e = 0.07$, $p = .14$].

Self-reported confidence. The right panel of Figure 2 presents mean reported confidence as a function of conclusion believability and reasoning ability. The data were analyzed using a $2 \times 2 \times 3$ (reasoning ability $\times$ response opportunity $\times$ belief) repeated measures ANOVA. Although the good reasoners were much more accurate than the poor reasoners, the two groups did not differ in confidence [$F(1,121) = 2.15$, $MS_e = 5.88$, $p = .15$], replicating our findings from Experiment 1. A power analysis revealed that we had an 80% power to find a difference greater than or equal to .31 on our 7-point scale, suggesting that the null finding could not be attributed to lack of power. Reasoning ability also did not interact with response opportunity [$F(1,121) = 0.43$, $MS_e = 0.89$, $p = .52$] or belief [$F(2,242) = 1.71$, $MS_e = 0.97$, $p = .18$].

As in Experiment 1, the participants' confidence increased from the first to the second response opportunity [$M = .56$; $F(1,121) = 64.81$, $MS_e = 0.89$, $p < .001$], even though they were not more accurate for their second response. This is consistent with the idea that participants' rational processes are often engaged to justify or rationalize their initial decisions, rather than to reanalyze those choices. There was also an interaction between belief and response opportunity [$F(1.87,225.97) = 4.51$, $MS_e = 0.18$, $p < .05$], so that the increase in confidence between the first and the second response opportunities was larger for the believable [$t(122) = 2.99$, $p < .01$] and unbelievable [$t(122) = 1.96$, $p = .05$] conclusions than for the neutral conclusions.

Finally, confidence varied as a function of belief [$F(1.69,204.03) = 32.38$, $MS_e = 1.15$, $p < .001$]. The pattern of this effect replicates the pattern in Experiment 1, so that the participants were more confident for believable [$t(122) = 6.81$, $p < .001$] and unbelievable [$t(122) = 5.94$, $p < .001$] conclusions than for neutral conclusions. Once again, there was no difference in confidence between believable and unbelievable conclusions [$t(122) = 0.48$, $p = .64$]. A power analysis revealed that we had an 80% power to find a difference greater than or equal to .10 on the 7-point scale.

Correlations between confidence and accuracy. As was the case in Experiment 1, we computed a mean accuracy and a mean confidence score for each participant. The overall correlation between confidence and accuracy was not significant [$r(108) = .10$, $p = .15$]. Because the good and the poor reasoners differed in terms of accuracy, this correlation was examined for the good and the poor reasoners separately. However, confidence and accuracy did not correlate for either the good [$r(46) = -.09$, $p =$

.27] or the poor [$r(49) = .08$, $p = .29$] reasoners. Thus, the participants who were more accurate were not more confident, and vice versa.

We also conducted a trial-by-trial analysis. For this, a correlation between confidence and accuracy was computed for each of the 24 conditions in the experiment (i.e., 3 levels of belief $\times$ 2 levels of validity $\times$ 2 blocks $\times$ 2 response opportunities). The average correlation across the 24 conditions was of the same magnitude as the one reported in Experiment 1 ($r = .075$); although small, this value was reliably greater than 0 [$t(23) = 1.84$, $p = .04$, one-tailed].

Summary. The data from Experiment 2 replicated the main phenomena in Experiment 1 concerning the relation between confidence and accuracy in syllogistic reasoning. First, the reasoners were once again more confident for problems leading to believable and unbelievable conclusions than for those leading to neutral conclusions but were more accurate only for the unbelievable conclusions. Second, the participants were more confident in their second than in their first response opportunity, although, unlike in Experiment 1, there was no increase in accuracy across responses. Once again, however, confidence and accuracy were influenced differently by the believability of the conclusion and the amount of time allowed to evaluate the conclusion. Third, the correlations between accuracy and confidence were very low. For both the good and the poor reasoners, confidence and accuracy did not exhibit a reliable correlation; the correlations between confidence and accuracy averaged less than .10. In terms of reasoning ability, the good and the poor reasoners were equally confident, despite large differences in accuracy. Overall, it seems that the good reasoners in Experiment 2 were performing more like "typical" reasoners, showing the traditional effects of validity and belief and their interaction. It is telling, however, that despite the near-chance performance of the poor reasoners, they were no less confident in their answers than were the good reasoners.

GENERAL DISCUSSION

In two experiments, we examined the relationship between self-reported confidence and reasoning accuracy. To this end, we examined (1) whether variables known to affect the accuracy of reasoning performance would likewise affect reasoners' confidence judgments, and vice versa; (2) whether reasoners would be over- or underconfident; and (3) whether confidence and accuracy would be correlated. Our evidence clearly supports the conclusion that reasoners' subjective confidence in the accuracy of their conclusion is a very poor indicator of the actual accuracy of their performance and that subjective confidence and objective accuracy are determined by different variables. The only occasion on which accuracy and confidence were correlated occurred when the participants were asked to provide an overall estimate of the number of problems answered correctly; on this measure, for the good reasoners, the estimates of the number answered

correctly and the number of problems actually solved correctly showed a robust correlation (.35 in Experiment 1).

It is important to note, however, that although these variables were correlated, the good reasoners tended to underestimate their performance, so that their overall estimates were not, in fact, reliable indicators of their actual performance. That is, whereas a problem-by-problem assessment appeared to produce systematic overconfidence, at least in the poor reasoners, the aggregate assessment produced underconfidence, at least in the good reasoners (see also Allwood & Granhag, 1996). In other words, neither measure of confidence appeared to be well calibrated with performance. This lack of calibration was not simply due to an attenuation of range on the confidence scale. Indeed, the mean scores for both confidence (Experiment 1 = 4.78; Experiment 2 = 5.23) and accuracy (Experiment 1 = .20; Experiment 2 = .14) were in the midrange of the scale. Thus, the small correlation between confidence and accuracy cannot be attributed to either ceiling or floor effects.

That having been said, it would be wrong to suggest that confidence played no role in reasoning. In fact, our results suggest that confidence determines the probability that participants will change their answers. In other words, it seems as though confidence determines what action will follow in the reasoning process (e.g., if a participant is more confident, he or she will be less likely to change his or her initial answer, regardless of whether this initial answer is correct). Because confidence seems to play such an important role in the reasoning process, it is necessary to understand how subjective confidence is determined.

Determinants of Confidence

Our results suggest that subjective confidence is mediated by factors extrinsic to the reasoning process itself and is cued by overt elements of the problem environment. The situation for reasoning appears to be analogous to that observed in studies of metamemory, in which participants' confidence about memory performance reflects their consideration of the conditions of study and test (see, e.g., Busey et al., 2000; Wells et al., 2002). Similarly, reasoners' confidence may be linked to variables that they think predict reasoning accuracy. First, the participants' confidence ratings were influenced by belief, or prior knowledge, so that they were more confident for unbelievable and believable problems than for neutral problems. This suggests that reasoners may believe that accuracy should vary with knowledge and, so, rate problems about which they have such knowledge with higher confidence than they rate problems for which such knowledge is lacking. Second, the participants were more confident when they were provided with extra time to think about a problem. This was true regardless of whether or not their first response was provided under a timed deadline and whether or not the extra time produced an increase in accuracy. Again, this suggests that confidence is mediated by variables external to the problems, such as the participants' beliefs about the relationship between time and accuracy.

Determinants of Accuracy

Evans (1989, 1996) proposed that participants' initial decisions may be based on preconscious heuristics that help to direct attention to relevant features of the problems. Rational analysis then follows and is often used to justify or rationalize the initial decision. Although this heuristic-analytic theory was derived from his work with the Wason selection task, our results suggest that it may be applicable to syllogistic reasoning as well. In Experiment 1, the participants made many of their initial decisions within the 10-sec deadline, suggesting that these decisions were heuristically driven and were based on fairly superficial features of the problems. Moreover, this initial answer did not change on about 70% of the trials, suggesting that decisions were often made early and that the extra time did not necessarily lead to a reanalysis of the problem. Extra time did lead to a large increase in confidence, however, lending support to the hypothesis that the participants were using the time to rationalize their initial answers.

Under this analysis, the accuracy of participants' responses would depend on the degree to which these heuristic processes produced correct responses. For example, Chater and Oaksford (1999) suggested that participants respond on the basis of the informativeness of the premises provided, rather than on the basis of a logical analysis of the premises (see also Oaksford, Roberts, & Chater, 2002). Specifically, they proposed that participants should prefer conclusions having a quantifier (e.g., *all*, *some*, or *no*) of the same type as that in the least informative premise. Accuracy, in this model, would therefore depend on whether or not the conclusion produced by the *minimal informativeness* rule was valid or not. The conclusion that reasoning decisions are frequently made quickly and that rational processes are often engaged post hoc to rationalize or justify the initial decision is also consistent with the hypothesis proposed by Gilhooly, Logie, Wetherick, and Wynn (1993), who suggested that most participants use superficial, heuristic strategies, rather than working through the logic of the argument.

Analytic Processes

Finally, although our data suggest that many decisions may be based on fast, heuristic processes, in accord with the dual-process theory of reasoning (Evans, 2003; Evans & Over, 1996; Klaczynski & Robinson, 2000), we also found evidence that some responses are determined by rational, analytic processes. Specifically, in Experiment 1, the reasoners changed their responses on a substantial minority of trials, and these changes resulted in an overall increase in accuracy. Thus, it appears that although many of the participants engaged in automatic, belief-based, System 1 reasoning, some of the participants did use the extra time on their second response to switch to the more logical analysis of System 2 reasoning. However, the overall increase in accuracy was small, lending support to Chater and Oaksford's (1999) suggestion that this ability does not seem to be well developed, or well used, in most

participants. That being said, it is apparent that at least some of the participants used the extra time to reexamine the problem and to derive a solution on the basis of a logical analysis.

Our data suggest two candidates for these analytic processes. The first is an integration process suggested by Thompson et al. (2003). These authors proposed a modified version of the verbal reasoning theory (Polk & Newell, 1995) to explain conclusion evaluation with belief-laden materials. According to the modified verbal reasoning theory (MVRT), when reasoners are presented with a conclusion to evaluate, they do so by attempting to find a model of the situation in which the premises and the conclusion can be integrated. When such a model is found, the conclusion is accepted; otherwise, it is rejected. Because all models of the premises are consistent with valid conclusions, the integration process is assumed to be easier for valid than for invalid conclusions, meaning that more valid than invalid conclusions are accepted.

In accord with this model, we observed that providing a second opportunity to think about the problems increased the tendency to accept the conclusions (i.e., when answers were changed, they more often changed from *no* to *yes* than vice versa). This finding is consistent with the hypothesis that conclusions are accepted when reasoners are able to find a model that integrates the conclusion with the premise; in this view, allowing extra time to reason about the problem increases the probability that the attempt at integration will be successful. Overall, the tendency to accept more problems on the second than on the first opportunity was somewhat less likely for invalid than for valid conclusions (i.e., in Experiment 2 there was no change, whereas in Experiment 1 there was an overall increase). These findings are also consistent with MVRT, in that finding an integrating model for invalid problems is more difficult than finding one for valid problems, so that allowing more time to consider the problem is of less benefit for invalid than for valid conclusions.

We also found some evidence consistent with the mental models theory (MMT) of belief bias (Newstead et al., 1992). In Experiment 1, the reasoners were more likely to reject invalid, unbelievable conclusions on their second than on their first opportunity; for the other invalid conclusions, there was no difference. According to the MMT, reasoners discriminate valid from invalid conclusions by attempting to construct models of the premises that might falsify the presented conclusion. Reasoners are more likely to undertake such a search when the conclusion is unbelievable than when it is believable or neutral, so that the former should be rejected more often than the latter.

Analytic processes in 10 sec? One of the most striking features of our data is that the interaction between validity and belief was observed even when the participants were asked to respond in less than 10 sec. Both the MMT and the MVRT propose that this interaction arises from the application of analytic, rational processes. Thus, in order to explain the interaction in the short-deadline condition, these theories must make the assumption that

10 sec is sufficient time to engage in a successful search for counterexamples (as per the MMT) or to find a model integrating the premises and conclusion (the MVRT). We thought this unlikely, given that relatively few responses are spontaneously generated in 10 sec (as our reanalysis of Thompson et al., 2003, indicated). Alternatively, under the more probable assumption that the short deadline was sufficient to curtail extensive analysis, one must conclude that the interaction is not due to formal reasoning processes but, rather, arises from the application of fast and simple heuristics, which can be applied in about 10 sec.

Conclusions

As we outlined in the introduction, we had reason at the outset to expect that the calibration between confidence and accuracy in syllogistic reasoning would not be perfect, but we did not anticipate that the correlation would be so poor. Indeed, good and poor reasoners were equally confident, despite large differences in accuracy. One explanation for the poor relationship between confidence and accuracy in reasoning is that inexperienced reasoners lack the skills that would help them to differentiate between more and less difficult problems (Kruger & Dunning, 1999). This suggests that providing the participants practice in a domain, choosing problem domains that were more familiar to the participants, or using items with a lesser degree of difficulty should produce better calibration. Indeed, Kruger and Dunning have found that improving the skill of participants not only made them more accurate, but also made them more aware of the limitations of their abilities. Thus, although our results demonstrate a striking dissociation of accuracy and confidence in deductive reasoning, it remains possible that extensive practice and training would yield realistic intuitions about the accuracy of one's deductions.

REFERENCES

- ALLWOOD, C. M., & GRANHAG, P. A. (1996). Considering the knowledge you have: Effects on realism in confidence judgements. *European Journal of Cognitive Psychology*, **8**, 235-256.
- BALL, L. J., LUCAS, E. J., MILES, J. N. V., & GALE, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology*, **56A**, 1053-1077.
- BARANSKI, J. V., & PETRUSIC, W. M. (1999). Realism of confidence in sensory discrimination. *Perception & Psychophysics*, **61**, 1369-1383.
- BUSEY, T. A., TUNNICLIFF, J., LOFTUS, G. R., & LOFTUS, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, **7**, 26-48.
- BUTTERFIELD, B., & METCALFE, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1491-1497.
- CAMPBELL, J. I. D., & THOMPSON, V. A. (2002). More power to you: Simple power calculations for treatment effects with one degree of freedom. *Behavior Research Methods, Instruments, & Computers*, **34**, 332-337.
- CHANDLER, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, **22**, 273-280.
- CHATER, N., & OAKSFORD, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, **38**, 191-258.
- DEFFENBACHER, K. A. (1980). Eyewitness accuracy and confidence:

- Can we infer anything about their relationship? *Law & Human Behavior*, **4**, 243-260.
- EVANS, J. S. B. T. (1989). *Biases in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.
- EVANS, J. S. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology*, **87**, 223-240.
- EVANS, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, **7**, 454-459.
- EVANS, J. S. 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. S. B. T., HANDLEY, S. H., & HARPER, C. N. J. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, **54A**, 935-958.
- EVANS, J. S. B. T., LEGRENZI, P., & GIROTTI, V. (1999). The influence of linguistic form on reasoning: The case of matching bias. *Quarterly Journal of Experimental Psychology*, **52A**, 185-216.
- EVANS, J. S. B. T., NEWSTEAD, S. E., ALLEN, J. L., & POLLARD, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology*, **61**, 263-285.
- EVANS, J. S. B. T., & OVER, D. E. (1996). *Rationality and reasoning*. Hove, U.K.: Psychology Press.
- GALOTTI, K. M., KOMATSU, L. K., & VOELZ, S. (1997). Children's differential performance on deductive and inductive syllogisms. *Developmental Psychology*, **33**, 70-78.
- GIGERENZER, G., HOFFRAGE, U., & KLEINBÖLTING, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528.
- GILHOOLY, K. J., LOGIE, R. H., WETHERICK, N. E., & WYNN, V. (1993). Working memory and strategies in syllogistic-reasoning tasks. *Memory & Cognition*, **21**, 115-124.
- JOHNSON-LAIRD, P. N., & BYRNE, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
- JOHNSON-LAIRD, P. N., & SAVARY, F. (1999). Illusory inferences: A novel class of erroneous deductions. *Cognition*, **71**, 191-229.
- KLACZYNSKI, P. A., & ROBINSON, B. (2000). Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning biases. *Psychology & Aging*, **15**, 400-416.
- KLAUER, K. C., MUSCH, J., & NAUMER, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, **107**, 852-884.
- KOEHLER, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, **110**, 499-519.
- KOEHLER, D. J. (1994). Hypothesis generation and confidence in judgment. *Dissertation Abstracts International: Sec. B. The Sciences & Engineering*, **54**, 6486.
- KORIAT, A., LICHTENSTEIN, S., & FISCHHOFF, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 107-118.
- KRUGER, J., & DUNNING, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality & Social Psychology*, **77**, 1121-1134.
- LEIPPE, M. R. (1980). Effects of integrative memorial and cognitive processes on the correspondence of eyewitness accuracy and confidence. *Law & Human Behavior*, **4**, 261-274.
- MCGARTHY, C., TURNER, J. C., OAKES, P. J., & HASLAM, S. A. (1993). The creation of uncertainty in the influence process: The roles of stimulus information and disagreement with similar others. *European Journal of Social Psychology*, **23**, 17-38.
- METCALFE, J. (1994). A computational modelling approach to novelty monitoring, metacognition, and frontal lobe dysfunction. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition* (pp. 137-156). Cambridge, MA: MIT Press.
- MINER, A. C., & REDER, L. M. (1994). A new look at feeling of knowing: Its metacognitive role in regulating question answering. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition* (pp. 47-70). Cambridge, MA: MIT Press.
- NEWSTEAD, S. E., HANDLEY, S. J., & BUCK, E. (1999). Falsifying mental models: Testing the predictions of theories of syllogistic reasoning. *Memory & Cognition*, **27**, 344-354.
- NEWSTEAD, S. E., POLLARD, P., EVANS, J. S. 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.
- NICKERSON, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking & Reasoning*, **2**, 1-31.
- OAKSFORD, M., ROBERTS, L., & CHATER, N. (2002). Relative informativeness of quantifiers used in syllogistic reasoning. *Memory & Cognition*, **30**, 138-149.
- PETERSON, D. K., & PITZ, G. F. (1988). Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 85-92.
- PETRUSIC, W. M., & BARANSKI, J. V. (2003). Adaptive decision processes in perceptual comparisons: Effects of changes in global difficulty context. *Journal of Experimental Psychology: Human Perception & Performance*, **29**, 658-674.
- POLK, T., & NEWELL, A. (1995). Deduction as verbal reasoning. *Psychological Review*, **102**, 533-566.
- PULFORD, B. D., & COLMAN, A. M. (1996). Overconfidence, base rates, and outcome positivity/negativity of predicted events. *British Journal of Psychology*, **87**, 431-445.
- QUAYLE, J. D., & BALL, L. J. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, **53A**, 1202-1223.
- ROEDIGER, H. L., III, & McDERMOTT, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Memory, Learning, & Cognition*, **21**, 803-814.
- SHAFIR, E., SIMONSON, I., & TVERSKY, A. (1993). Reason-based choice. *Cognition*, **49**, 11-36.
- SPEELMAN, G. P., & KIRSNER, K. (1997). The specificity of skill acquisition and transfer. *Australian Journal of Psychology*, **49**, 91-100.
- SPORER, S. L., PENROD, S., READ, D., & CUTLER, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, **118**, 315-327.
- STANKOV, L., & CRAWFORD, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence*, **25**, 93-109.
- THOMPSON, V. A. (1996). Reasoning from false premises: The role of soundness in making logical deductions. *Canadian Journal of Experimental Psychology*, **50**, 315-319.
- THOMPSON, V. A., STREIMER, C. L., REIKOFF, R., GUNTER, R. W., & CAMPBELL, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin & Review*, **10**, 184-189.
- TORRENS, D., THOMPSON, V. A., & CRAMER, K. M. (1999). Individual differences and the belief bias effect: Mental models, logical necessity, and abstract reasoning. *Thinking & Reasoning*, **5**, 1-28.
- WELLS, G. L., OLSON, E. A., & CHARMAN, S. D. (2002). The confidence of eyewitnesses in their identifications from lineups. *Current Directions in Psychological Science*, **11**, 151-154.
- ZAKAY, D., & TUVIA, R. (1998). Choice latency times as determinants of post-decisional confidence. *Acta Psychologica*, **98**, 103-115.

NOTES

1. Sixty-four participants were originally tested, but another 64 participants were run in order to ensure that our results could be replicated. Because the results were the same across both groups of participants, the data were combined, resulting in a total of 128 participants.
2. Quayle and Ball (2000) found that participants were more confident for valid than for invalid conclusions. However, in their analysis, validity was confounded with response (*yes, no*). When we analyzed the effect of validity separately for *yes* and *no* responses, there was no difference in reported confidence between valid and invalid conclusions [$F(1,116) = 0.14$, $MS_e = 0.82$, $p = .71$]. This finding was replicated in Experiment 2.
3. Because the oft-found effect of belief on accuracy, usually attributed to a search for counterexamples for unbelievable problems (e.g., Newstead et al., 1992), was found even on the first response opportunity,

we were concerned that the participants were not adhering to the 10-sec deadline. Overall, the participants adhered to the deadline; only 29% had a mean RT greater than 10 sec for their first response opportunity. When the participants who exceeded a mean of 10 sec were excluded, the belief $\times$ validity interaction did indeed disappear [$F(2,174) = 1.33$, $MS_e = 0.11$, $p = .27$], but it reappeared when the exclusion criterion was increased to 11 sec [$F(2,210) = 2.98$, $MS_e = 0.12$, $p < .05$].

4. As can be seen in these written instructions, a typographical error occurred, which resulted in a contradiction between the values printed on the scale and the written instructions. Because of this, we did not conduct any analyses that required interpreting absolute values of the scale (i.e., to determine whether the participants were generally over- or underconfident) and, for this experiment, conducted only those analyses that made use of relative values.

APPENDIX

Examples of Believable, Unbelievable, and Neutral Syllogisms Used in Experiments 1 and 2

“Some C are not A” Conclusions	
<i>Valid/Believable</i>	<i>Valid/Neutral</i>
No police dogs are argonelles	No Welps are lawyers
Some highly trained dogs are argonelles	Some Abens are lawyers
Therefore, some highly trained dogs are not police dogs	Therefore, some Abens are not Welps
<i>Valid/Unbelievable</i>	<i>Invalid/Neutral</i>
No addictive things are ramadions	Some Delpads are bird watchers
Some cigarettes are ramadions	No Furels are bird watchers
Therefore, some cigarettes are not addictive things	Therefore, some Furels are not Delpads
<i>Invalid/Believable</i>	<i>Invalid/Unbelievable</i>
Some vitamin tablets are opprobines	Some rich people are Hudons
No nutritional things are opprobines	No millionaires are Hudons
Therefore, some nutritional things are not vitamin tablets	Therefore, some millionaires are not rich people
“Some A are not C” Conclusions	
<i>Valid/Believable</i>	<i>Valid/Neutral</i>
Some religious people are Selaciens	Some Rewons are bus drivers
No priests are Selaciens	No Likels are bus drivers
Therefore, some religious people are not priests	Therefore, some Rewons are not Likels
<i>Valid/Unbelievable</i>	<i>Invalid/Neutral</i>
Some astronauts are Lapitars	No Mobbles are teachers
No healthy people are Lapitars	Some Plicks are teachers
Therefore, some astronauts are not healthy people	Therefore, some Mobbles are not Plicks
<i>Invalid/Believable</i>	<i>Invalid/Unbelievable</i>
No well-educated people are Pennes	No deep sea divers are Sylvians
Some judges are Pennes	Some good swimmers are Sylvians
Therefore, some well-educated people are not judges	Therefore, some deep sea divers are not good swimmers

(Manuscript received September 16, 2004;
revision accepted for publication April 4, 2005.)