Journal of Experimental Psychology:
Learning, Memory, and Cognition
1986, Vol. 12, No. 4, 623-634

0278-7393/86/$00.75

Premonitions of Insight Predict Impending Error

Janet Metcalfe
Indiana University

Five experiments explored the dynamic metacognitions that accompany the problem- and anagram-solving processes. Subjects repeatedly rated how "warm" or close they were to solution. High feelings of warmth before an answer indicated that the answer would be incorrect. Moderately low warmth ratings characterized correct responses. The data suggest that the high warmth ratings may result from a process wherein subjects convince themselves that an inelegant but plausible (wrong) answer is correct. No gradual rationalization process precedes the correct response to insight problems. The warmth-rating data also indicate that when the correct answer was given to the problems and anagrams used in this study, there was usually a subjectively catastrophic insight process.

In this article, I investigate dynamic metacognitions that lead up to the production of correct or incorrect solutions in solving insight problems and anagrams. A technique is used in which subjects are asked to give judgments about how close they feel to the solution of problems—called feeling-of-warmth judgments—repeatedly during the course of problem solving. These judgments are called "warmth" judgments after the searching game in which one person hides an object and then directs others who do not know where the object is by telling them that they are getting warmer—closer to the object, or colder—farther away (see Simon, Newell, & Shaw, 1979). Subjects are asked to assess their subjective warmth, or to indicate how near they believe they are, to the solutions to problems. As will become clear during the development of this article, these feeling-of-warmth judgments provide an indication of whether subjects will produce the correct solution or an incorrect solution to insight problems and to anagrams.

Feelings of warmth, in problem solving, seem analogous to the tip-of-the-tongue phenomenon in memory retrieval (Brown & McNeill, 1966; Eysenck, 1979; Freedman & Landauer, 1966; Hart, 1967; Yarmey, 1973) wherein a subject can almost, but not quite, remember the desired information. Metacognitions such as the tip of the tongue and feeling of knowing are often a good indication that correct memory performance will ensue (Hart, 1965, 1967; Nelson, 1984; Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982). Thus, if the analogy holds, we might expect high feelings of warmth to provide an accurate index of impending problem solution. However, the analogy to memory may not hold. Metcalfe (1986) has found that the feeling of knowing produces different results in the memory and problem-solving domains. Although feelings of knowing were predictive of memory, they did not predict problem-solving performance with insight problems.

People's warmth ratings also potentially bear on the question of whether (certain) problems are solved by insight or by accumulative processes. If insight problems are solved suddenly, by a perceptual-like shift in gestalt (Ellen, 1982; Maier, 1930, and see Dominowski, 1981), then we might expect that subjects would show fairly constant warmth ratings before solving and then sudden increases upon solution.

However, these problems might be solved incrementally (see Bowers, 1985), and the main process of interest could be one of searching through various levels of a solution maze (as Weisberg & Alba, 1981, 1982, suggest) or of gradual accumulation of information. Perkins (1981) notes that the solution to insight problems, such as are studied in the present paper, seems to provide a minor version of the experience of the grand insights of thinkers such as Darwin or Poincaré. The method Perkins uses to study insights is retrospective report. He argues, though, that the process of insight may be "less leaplike" and less mysterious than might have been thought. These views indicate that we should expect gradual increases in warmth ratings up to the time of solution. Simon et al. (1979) have discussed, in some detail, a similarity-matching heuristic in problem solving that should be reflected in subjects' ratings of how warm or cold they are. They conceptualize the solving of algebra, logic, and chess problems as a search through a maze of potential pathways that eventually leads to the solution to the problem. "Examination of paths produces clues of the 'warmer-colder' variety" (p. 154). The clues will indicate increasing warmth if the state that transpires by virtue of taking a particular path is more similar to the goal state than is the state that existed before the problem solver traversed the pathway in question. If the state becomes more similar to the goal state, then the person takes that pathway or applies that operator. This heuristic of reducing differences from the goal is fundamental to the general strategy of means-ends (functional) analysis which has far-reaching applications. Simon et al. (1979) provide a think-aloud protocol (see

I wish to thank Jennifer Campbell, Roger Dominowski, Eric Eich, Art Glenberg, Gordon Logan, Thomas Nelson, Lorene Oikawa, John Pinel, Henry Roediger, Paula Ryan, Don Sharpe, and James Steiger for their help. This research was supported by a University of British Columbia Social Sciences and Humanities Research Council grant and by Natural Sciences and Engineering Research Council of Canada Grant A0505 to Janet Metcalfe.

Correspondence concerning this article should be addressed to Janet Metcalfe, Department of Psychology, Indiana University, Bloomington, Indiana 47405.

Ericsson & Simon, 1984) for an algebra problem that suggests that this heuristic is used. The main advantage to the heuristic is that it greatly reduces the number of pathways or expressions that must be searched, making the problem-solving process goal directed (though still mechanistic), rather than random or systematically exhaustive. Thus, the warmth clues, which subjects presumably monitor, seem to be central to the problem-solving process. Indeed, it could be argued that many problems would be unsolvable, because the search time would simply be too great, were it not for these clues. Simon et al.'s (1979) discussion of the subjective warmth phenomenon and its theoretical importance for problem solving is the only reference I have found to feelings of warmth. There have been no explicit experimental investigations of the accuracy of feeling-of-warmth judgments.

Experiment 1

Method

Procedure. Subjects were given a sheet of paper on which was written the "B.C. problem" (given below under Materials). They were told that they would be asked to write down a number between 0 and 10 (where 0 meant that they were *cold* about the problem, that is, they had no glimmering of the solution; 10 meant they were certain they had the solution; and intermediate ratings indicated intermediate warmth) every 10 s on the sound of a click. When subjects had achieved the solution, they were to write it down clearly, so that the experimenter would know that it was correct.

Materials. The problem was "A stranger approached a museum curator and offered him an ancient bronze coin. The coin had an authentic appearance and was marked with the date 544 B.C. The curator had happily made acquisitions from suspicious sources before, but this time he promptly called the police and had the stranger arrested. Why?"

Subjects. The subjects were 134 introductory psychology students who participated in the experiment as part of a classroom demonstration on problem solving. Thirty-six students wrote answers that were uninterpretable or failed to record any warmth ratings. These were excluded from further consideration.

Results

In all of the analyses that will be presented in this article, a probability level of $p < .05$ will be taken as the criterion for statistical significance. The particular probability level for a reported effect will not be stated, unless it is greater than this value.

Forty-three students got the right answer to the problem and 44 students wrote down a wrong answer. Eleven students did not reach a conclusion in the 5 min allocated to solve the problem. To be scored as having gotten the right answer, a subject had to indicate that he or she was aware of the fact that the coin could not have been authentic because it would have been impossible for someone who had actually lived in 544 B.C. to know that eventually the calendar would change to mark a date that itself was still 544 years in the future. Thus, the date is a giveaway. The wrong answers that were listed included explanations such as "bronze was not invented yet in 544 B.C." and "the curator knew the man was a thief."

Warmth as a function of correctness. The warmth ratings were segmented depending upon whether the answer given was correct or incorrect. In the first analysis, all subjects who had three or more warmth ratings before the 10 response were included and the means of the last three intervals before the interval with which the answer was given were examined. The pattern, shown in Table 1, suggested that correct responders gave lower warmth ratings than did incorrect responders, although this effect did not quite reach a significant level, $F(1, 50) = 2.81$, $MS_e = 61.68$, $p = .09$. There was an effect of interval, such that warmth increased as the solution interval approached, $F(2, 100) = 24.38$, $MS_e = 1.86$. The interaction between interval and correctness was not significant.

When the last two (rather than three) warmth ratings were analyzed, the higher warmth for incorrectly than for correctly solved problems was significant, $F(1, 59) = 6.48$, $MS_e = 28.63$. The difference in warmth between correct and incorrect problems was also significant when only the last rating was analyzed, $F(1, 68) = 15.00$, $MS_e = 9.08$. The mean warmth on the last interval for subjects who had one or more rating was 2.71 for correct responses and 5.52 for incorrect answers. The marginal effect when three intervals were included might be because there were few subjects in the analysis (19 correct and 33 wrong) or because more of the very first warmth ratings are included in the data. There was no difference in the first warmth ratings

Table 1
Warmth Ratings on the Last Three Intervals Before Solution for Correct and Incorrect Responses in Experiments 1, 2, 3, 4, and 5

Response	N	Interval rating: Third last	Second last	Last	Solution
Experiment 1 (problem)					
Correct	19	2.05	2.42	3.47	10
Incorrect	33	2.92	3.57	5.25	10
Experiment 2 (problems)					
Correct	29	3.52	3.73	4.60	10
Incorrect	29	4.16	4.64	5.02	10
Experiment 3 (anagrams—10 not required)					
Correct	16	2.06	2.19	2.62	9.78
Incorrect	16	2.10	2.16	2.53	7.86
Experiment 4 (anagrams—no guessing)					
Correct	23	1.29	1.50	2.20	10
Incorrect	23	1.78	1.91	2.63	10
Experiment 5 (anagrams—guessing)					
Correct	20	1.92	2.25	2.81	10
Incorrect	20	2.61	2.67	3.83	10

Note. In Experiment 1, subjects are divided into two groups on the basis of their performance on the one problem; in the other experiments, subjects' correct and incorrect responses on all problems are considered.

Table 2
Proportion of Problems or Anagrams Showing an Insight or an Incremental Pattern of Warmth Ratings

	Insight pattern		Incremental pattern		*N*	
Condition	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Experiment 1	.46	.14	.21	.62	24	37
Experiment 2	.66	.56	.11	.08	56	80
Experiment 3	.79	.66	.06	.06	227	32
Experiment 4	.76	.76	.15	.22	385	55
Experiment 5	.78	.44	.14	.32	185	108
Overall	.76	.52	.12	.25	877	312

Note. For the insight pattern, the last warmth rating before the rating with which the response was given was no more than 1 point greater than the first rating given in the protocol. For the incremental pattern, the last warmth rating before the rating with which the response was given was at least 5 points greater than the first rating given.

between problems that would be solved correctly or incorrectly, $t(68) = .06, p = .91$. The means were .97 and 1.00, respectively.

Insight pattern analysis. It was decided that warmth rating sequences that showed no more than a 1-point increase from the first to the last rating would be said to demonstrate an "insight" pattern of warmth ratings. (To be very strict, one might wish to argue that there should be a zero increase from start to end. However, in this and subsequent experiments, subjects sometimes wrote or were told to write their first warmth rating—usually a zero—before even reading the problem, so I thought that a 1-point increase was conservative enough.) Protocols were included that had two or more ratings. The top row of Table 2, on the left, shows the proportion of correctly and incorrectly solved problems that showed such an insight pattern. Nearly half of the correctly solved problems showed this pattern. A difference in proportions test for the correctly and incorrectly solved problems showed that the correctly solved problems were more likely to show the insight pattern (.46) than were the incorrectly solved problems (.14), $z = 2.78$.

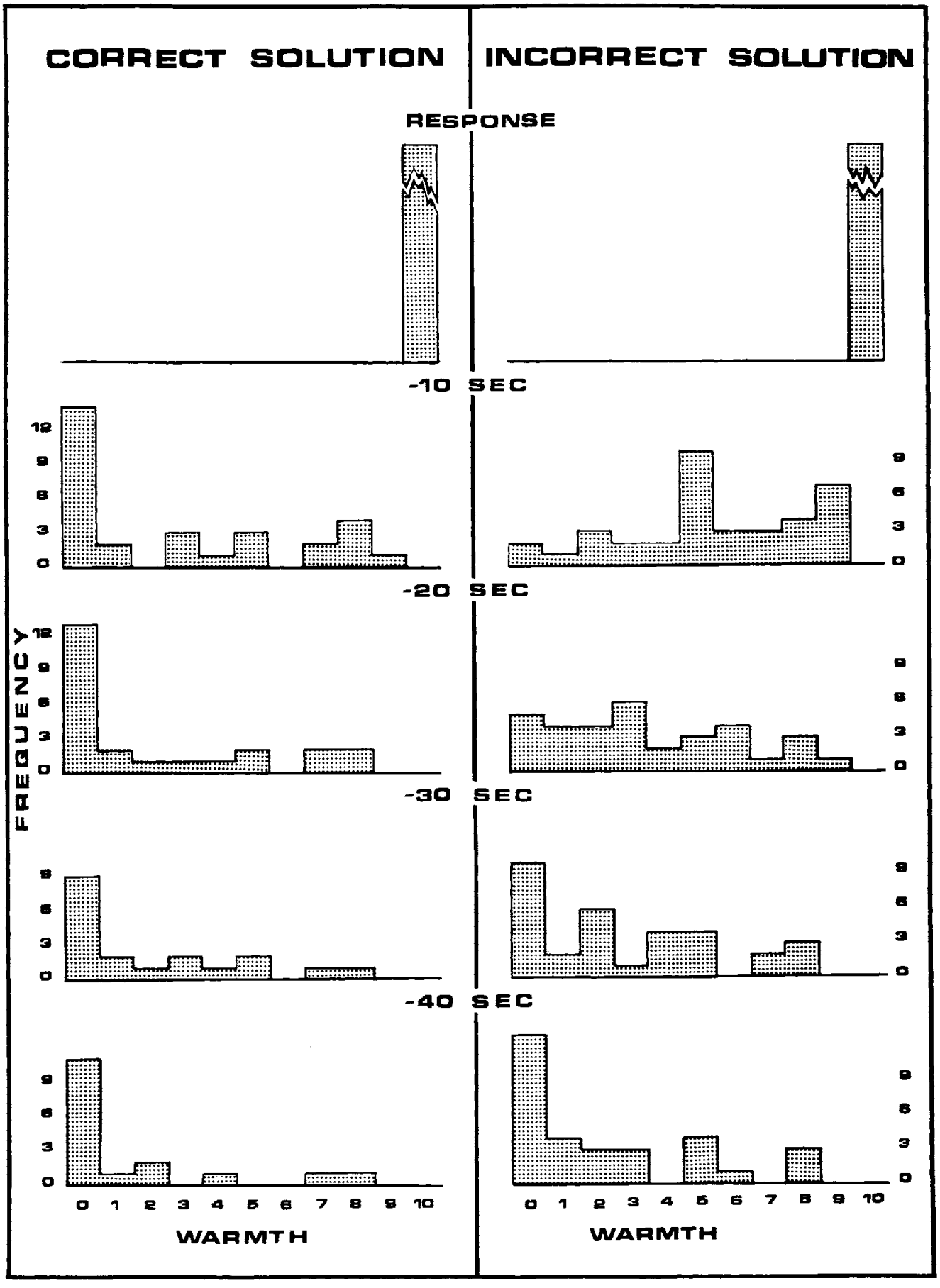
An incremental pattern was defined (somewhat arbitrarily) as one that showed at least a 5-point difference between the first and last warmth rating. In this experiment, the initial warmth rating was quite low, so a 5-point increment did not seem excessively demanding as an indication that there was a gradual increase in warmth. Table 2 shows the proportion of correctly and incorrectly solved problems whose warmth protocols conform to this incremental pattern. The results on the insight and incremental proportions are not independent of one another, of course, but rather give two slices through the distributions of warmth protocols.

There is a conflict between the conclusions one would draw about insight depending upon the analysis used. The mean warmth ratings show a significant increase in warmth over the last three intervals, suggesting insight does not occur. The insight pattern analysis given above indicates that insight does occur a large proportion of the time. When the warmth ratings are averaged, the high ratings contribute disproportionately. Thus, one subject who shows a very high warmth value offsets many subjects showing very low values. Figure 1 provides information about the distributions of warmth ratings that may not be obvious from Tables 1 and 2. The figure shows the frequency of each value of warmth rating given 10, 20, 30, and 40 s before a solution was given. All subjects who had data in the interval depicted are included. By reading the figure from bottom to top, the trends in distributions as subjects approach giving an answer can be seen. There is increasing warmth for those subjects who got the wrong answer, shown on the right half of the figure. The modal value increases from zero, at four intervals (40 s) before a response is given, to five, in the interval immediately preceding response. In contrast to this upward drift in warmth ratings for incorrect answers was the lack of drift shown by those subjects who got the correct solution. Like the people who gave an incorrect answer, the modal warmth value 40 s before the correct response was zero. However, the modal value of warmth immediately preceding the correct solution was still zero. Note that the means do increase, though, because the tail of the distribution for correct answers becomes more skewed. The finding shown by the insight pattern analysis and the frequency distributions—that warmth increased abruptly at time of solution—is what might be expected if the problem were solved by insight.

Discussion

There were two findings of interest in Experiment 1. First, when the problem was correctly solved, the modal warmth ratings increased from the minimum to the maximum in one interval. This suggests that the B.C. problem is an insight problem. The insight pattern analysis also revealed that on nearly half of the protocols of correctly solved problems the increase in warmth preceding solution did not exceed 1 point. (However, there is an incremental pattern over the last three intervals when the analysis is computed on means.) Second, with the problem used in Experiment 1, a strong feeling of warmth predicted failure to solve the problem rather than success.

The finding that high warmth predicted the wrong answer rather than the right one was one I found surprising. To determine if it was surprising to others, I conducted a survey to investigate people's expectations about the relation between warmth ratings and problem solving. The subjects were third-year psychology students at the University of British Columbia who were given an explanation of warmth judgments and a descrip-


CORRECT SOLUTION
INCORRECT SOLUTION
RESPONSE
-10 SEC
-20 SEC
-30 SEC
-40 SEC
FREQUENCY
0 3 6 9 12
0 1 2 3 4 5 6 7 8 9 10
WARMTH

tion of the experiment. They were then asked whether (a) the correct solutions would be preceded by lower warmth judgments than the incorrect solutions, (b) warmth judgments would not discriminate between right and wrong answers, (c) correct solutions would be preceded by higher warmth judgments than would wrong answers, or (d) someone had already told them the results of the experiment (because the survey was conducted after the experiment). Most respondents expected, wrongly, that if people thought that they were getting warm, they would get the correct solution (48) or that there would be no difference (37). Several subjects (3) had heard the actual results of the experiment. Very few students (3) correctly predicted the results: that correct answers are preceded by lower warmth ratings than incorrect answers.

Insight. The most interesting interpretation of the results of Experiment 1 relates to the processes and phenomenology involved in insight. It may be that the cognitive processes needed to solve insight problems are not incremental but rather are all or none. Those people who solved the problem used a strategy that resulted in a function of warmth ratings that looked more like what would be predicted by the insight hypothesis than did those people who gave an incorrect solution. Those people who failed to solve the problem may have used a successive approximations strategy, a "satisficing" strategy (Simon, 1979) in which a "good-enough" answer was gradually accepted, or an accumulative strategy. Several potential concerns about this interpretation are mentioned below.

Intelligence differences. It is possible that the more intelligent or experienced subjects (who also tend to get the answers right) differ in their monitoring from less intelligent subjects. In particular, the more intelligent subjects might give lower ratings. Oskamp (1962) has shown that although experienced judges gave more accurate clinical assessments, their confidence was lower than less experienced judges. This pattern resembles the one found in Experiment 1. Because the difference in warmth may be a function of problem-solving expertise or of intelligence, a within-subjects design was used in Experiment 2.

Demand for progress. If it took longer to solve the problem incorrectly than correctly, then those people who erred might be more prone to implicit pressure to give indications of progress. Among those subjects who had at least one warmth rating, a significant difference occurred in elapsed time before giving the correct solution (5.35 intervals) relative to an incorrect solution (8.08 intervals), $t(68) = 2.49$. Because the demand for progress hypothesis of the locus of the warmth differences is viable, time differences are examined in Experiment 2.

Attention. Perhaps those subjects who correctly solved the problem showed lower warmth ratings than those who incorrectly solved it, because the former gave default ratings of zero on the task rather than really making judgments of their progress. Making judgments of warmth is an attention-demanding control process, whereas simply giving default values may require little effort (see Kahneman, 1973; Shiffrin & Schneider, 1977). Automatic responding to the rating task would leave more cognitive energy available for problem solving which could increase the chance of getting the right answer. Subjects, when tested individually by human experimenters, report making the best assessments of warmth that they can, although some find the task annoying. If subjects were not making effortful judgments on correctly solved problems there would be no reason to expect a difference in warmth between those problems and problems that are not solved at all. A within-subjects comparison of warmth on correctly solved and unsolved problems is presented toward the end of this article on the four experiments that follow. Also, I reanalyzed the data excluding any subject who gave only 0 as a response because such subjects might be suspected of not doing the rating task. This did not change the results reported here.

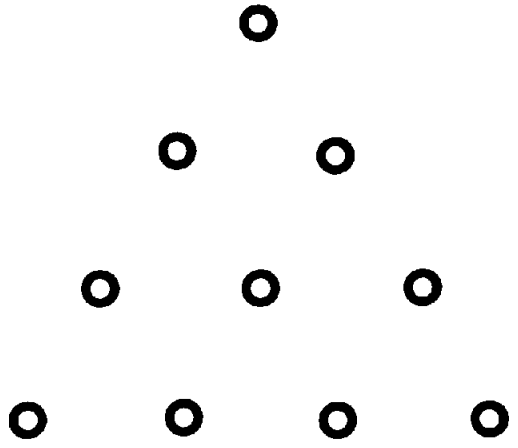


Figure 2. The triangle problem. Show how you can move three circles to get the triangle to point to the bottom of the page.

Correlational nature of the study. Although I have labeled the studies "experiments," they are in fact correlational in nature. Some of the interpretive problems outlined above may stem from the correlational, rather than manipulated, nature of the effects under investigation.

Generality. The results of Experiment 1 apply to one problem only. To study the generality of the phenomenon a number of problems are used in the experiment that follows. In Experiments 3, 4, and 5, anagrams are used.

Experiment 2

The major change between the first experiment and the second was that a number of problems were given to each subject. The warmth ratings, depending upon whether a subject got a problem right or wrong, were examined as a within-subjects rather than as a between-subjects factor.

Figure 1. Distributions of warmth ratings for four intervals leading up to correct (left panel) and incorrect (right panel) responses to the B.C. problem, Experiment 1.

Method

Procedure. Before doing the problem-solving task, subjects performed a memory task and a feeling-of-knowing task that are reported elsewhere (Metcalfe, 1986, Experiment 1). They saw eight problems and were asked to give the solution to these. Three of the problems were very easy, and most subjects solved them with no difficulty. The remaining five problems were selected from those presented under Materials below and were fairly difficult. These usually were not solved immediately unless it happened that the subject was familiar with the problem before coming to the experiment. Subjects were asked to rank order the first five problems they had not solved immediately in terms of how difficult they were. In addition they were asked to give an expectation score, on a scale from 0 (*sure will not solve*) to 10 (*sure will solve*), indicating how likely they were to solve each problem in a 5-min solution interval. The unsolved problems were reshuffled and presented to subjects one by one for solution. Every 15 s, the experimenter asked the subject to write down a warmth rating on a scale from 1 to 10 where 1 meant *cold* and 10 meant *certain about the correct solution.* Once a *10* had been given, the subject wrote down the answer. If the experimenter was unsure about the meaning of the response the subject had given, he asked the subject to clarify it. Subjects were tested individually.

Subjects. The participants in the experiment were 44 introductory psychology students at the University of British Columbia, who received a small bonus course credit for their participation.

Materials. The following problems were typed or drawn on 3 by 5-in. index cards. Answers, as well as warmth ratings, were written in an 8.5 by 11-in. answer booklet.

1. The horse trading problem: A man bought a horse for $60 and sold it for $70. Then he bought it back for $80 and sold it for $90. How much money did he make (or lose) in the horse trading business?

2. The chain problem: A woman has four pieces of chain. Each piece is made up of three links. She wants to join the pieces into a single closed ring of chain. To open a link costs 2 cents and to close a link costs 3 cents. She has only 15 cents. How does she do it?

3. The gardener's problem: A landscape gardener is given instructions to plant four special trees so that each one is exactly the same distance from each of the others. How would you arrange the trees?

4. The oil and vinegar problem: A small bowl of oil and a small bowl of vinegar are placed side by side. You take a spoonful of the oil and stir it casually into the vinegar. You then take a spoonful of this mixture and put it back into the bowl of oil. Which of the two bowls is more contaminated?

5. The postcard problem: Describe how to cut a hole big enough to put your head through in a (three by five inch) postcard.

6. The triangle problem (see Figure 2): The triangle points to the top of the page. Show how you can move three circles to get the triangle to point to the bottom of the page.

The correct solutions for the problems are as follows.

1. The horse trading problem: $20

2. The chain problem: Open all three links in one of the sections (cost $3 \times 2 = 6$ cents). Use these three links to join the three remaining sections. Closing cost is $3 \times 3 = 9$ cents.

3. The gardener problem: Plant them in a tetrahedron. Three trees are planted in an equilateral triangle. The fourth tree is either planted at the top of a hill in the middle of the triangle or is in a hole in the middle of the triangle.

4. The oil and vinegar problem: Both bowls are equally contaminated.

5. The post card problem: Cut the postcard into a long strip (by making a spiral, for instance). Then put a slit through the long strip.

6. The triangle problem: The three points of the triangle are rotated around the central rosette (see Figure 3).

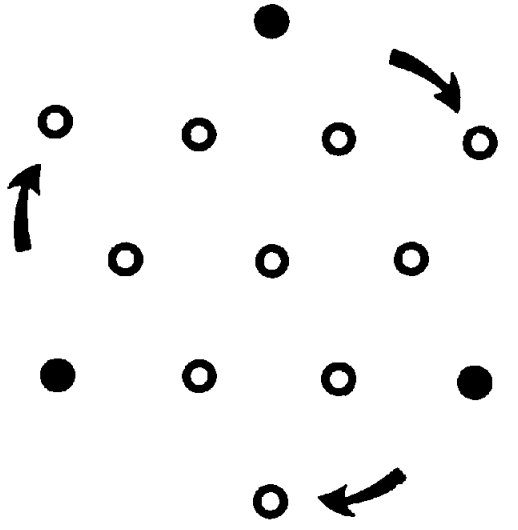


Figure 3. The triangle problem solution. The three points of the triangle are rotated around the central rosette.

Results

These problems are very difficult. The average proportions calculated on all 44 subjects who participated in the experiment were .28 correct, .49 incorrect, and .23 with no solution given. Subjects with at least one correct answer, with three or more warmth ratings preceding the answer, and at least one incorrect (but warmth rating = 10) answer with three or more warmth ratings were included in the following analysis. There were 29 such subjects.

Warmth as a function of correctness. An average warmth rating for each of the three intervals under consideration was computed for the correct and for the incorrect answers for each subject if he or she had more than one answer in the class. As is shown in Table 1, the warmth ratings on the incorrectly solved problems were higher than on the correctly solved problems, $F(1, 28) = 4.80$, $MS_e = 23.16$. The mean warmth rating over the three intervals for the correctly solved problems was 3.95; it was 4.61 for the incorrectly solved problems. There was an effect of interval such that warmth increased toward the solution, $F(2, 56) = 18.37$, $MS_e = .76$. The interaction between interval and correctness of answer was not significant. Thus, this experiment, like the first experiment, showed that correct answers to insight problems were preceded by lower warmth ratings than were incorrect solutions.

Insight pattern analysis. Table 2 shows the proportion of correct and incorrect problems (summed over all subject-problems with at least two ratings) that conform to the insight and incremental patterns, as outlined following Experiment 1. The difference in proportions between the correct (.66) and incorrect (.56) problems showing an insight pattern was not significant ($z = 1.2$), although the direction of difference in this experiment was the same as in the first experiment. A majority of the problems showed an insight pattern. The lack of significant difference on this measure might be due, in part, to the fact that the warmth ratings started, and stayed, much higher in this

experiment than in the previous one. This high starting value (which will be discussed below) constrains the scale and makes it difficult for differences in the pattern to emerge.

Time. Because the difference in time to solve had been a potential confounding factor in Experiment 1, the time to solve was analyzed in this experiment as a function of whether the answers were right or wrong. The average number of intervals for each subject, as included in the basic warmth analysis presented above, was 9.33 for correct answers and 6.56 for wrong answers. This difference was significant, $t(28) = 2.40$. Note that it was in the opposite direction to what was found in the first experiment.

Expectancies. The mean warmth ratings were considerably higher in the present experiment than they had been in the earlier experiment. It seemed plausible that the ranking and expectations task at the start of the present experiment, but not Experiment 1, might have provided a high anchor on the warmth ratings. The high-expectations subjects, based on a median split on the average expectation scores, did show higher warmth scores over the last three intervals (5.23) than did subjects with low expectations (3.22), $F(1, 27) = 6.84$, $MS_e = 153.14$. The difference in warmth between correct and incorrect answers was found with this analysis (with both expectation groups), as before, $F(1, 27) = 5.39$, $MS_e = 22.59$. The frequency histograms corresponding to Figure 1 are more uniformly distributed over the entire warmth scale. They are not presented because they are relatively uninteresting. There was no effect of expectations on proportion correct, $t(27) = .61$, $p = .55$. The proportion correct for the high-expectation subjects was .30, and .33 for the low-expectations subjects.

The results of this within-subjects experiment substantiate the findings of the first experiment. The warmth ratings were higher for the incorrectly solved problems than for the correctly solved problems. This effect is not attributable to the amount of time taken to provide correct or incorrect solutions because in this experiment (in contrast to the first), subjects took longer to solve problems correctly than incorrectly. The effect also does not appear to be attributable solely to individual differences because it occurred in a within-subjects design. In addition, most of the problems showed warmth protocols that indicated that subjects experienced insight.

It could be argued that the difference in warmth between correctly and incorrectly solved problems is attributable to the misleading nature of some insight problems. Weisberg and Alba (1981) have pointed out that one characteristic of insight problems is that the obvious solutions do not work. Earlier experience of subjects may work against obtaining the correct insightful solution (Luchins, 1942, Bartlett, 1958). Levine (1975) notes that subjects tend to explore the wrong solution domain with insight problems. Sternberg and Davidson (1982, 1983; and see Sternberg, 1985) point out that some insight problems invite the wrong encoding. Perhaps these special deceptive characteristics of insight problems are essential to the finding of higher warmth on wrong than right answers. If so, then the effect should disappear with other materials, such as anagrams, which are not deliberately misleading. Feelings of warmth in anagram solving are investigated in the experiment that follows.

Experiment 3

Method

Materials. The 20 to-be-solved anagrams (Sherman, personal communication, 1985) were ssoia, phmny, aeuvl, pmuoi, ttnua, rdcei, oapnr, tlcee, reckl, elcsa, piaot, ulipp, nitga, ocbna, ucrco, dtuai, mnaai, oocnl, aebrl, and augdr.

Procedure. Subjects were shown anagrams printed one to a page along with 30 blanks for warmth ratings. They were asked to assess their feelings of warmth, on a scale from 1 to 10, during the course of solving the anagrams, by writing down a number at the sound of a tap given by the experimenter every 10 s. The maximum amount of time allowed for any anagram was 5 min. Most, but not all, subjects completed all of the anagrams listed above in the 50-min experiment. Subjects were told to write down a *1* after turning each page to indicate that they were ready to start solving. When they had formed a word from the anagram, they were asked to "write down the word, and your warmth rating". (Owing to a procedural error, subjects were *not* told, in this experiment, only to write down the word when their warmth rating was 10, or when they were certain of the response.) Subjects were not allowed to write out permutations of the anagram letters during the course of solving but rather were required to compute the solution mentally.

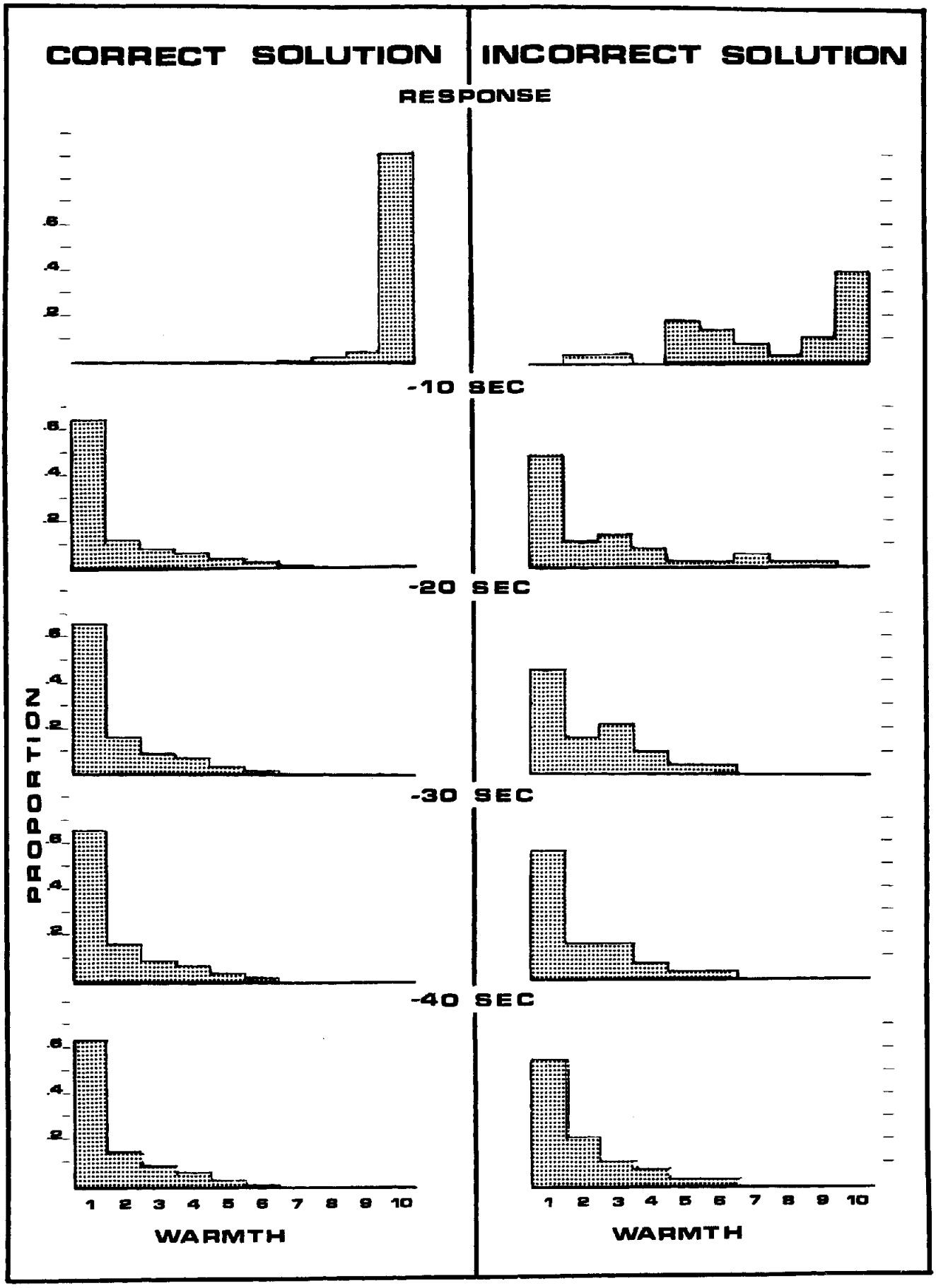
Subjects. The participants were 24 introductory psychology students who participated in return for a small bonus credit.

Results

On average, 74% of the anagrams were solved correctly; 7% were given incorrect solutions; 19% were attempted but unsolved. Figure 4, the analogue to Figure 1, shows the distributions of warmth ratings for correct and incorrect solutions to the anagrams four, three, two, and one 10-s intervals before the solution was written down, as well as the warmth rating given with the response. Data from all subjects who had ratings in the appropriate intervals are included in the figure. As can be seen from the left panel of the figure, the modal warmth rating at all intervals before the solution was given was 1. The distribution changes very little from 40 s before a response to 10 s before a response. The shift to a warmth rating of 10 is abrupt. This pattern is typical of the correct-solution pattern for Experiments 4 and 5 as well. For those anagrams that were incorrectly solved, the modal warmth ratings were also 1. The main difference between the correct and the incorrect solutions occurs in the warmth rating given at the time the response is given, rather than in the ratings leading up to a response. The warmth rating given with the response is lower for incorrect responses than for correct responses.

An analysis of variance (ANOVA) was conducted on the last three warmth ratings of the 16 subjects who had at least one correct and one incorrect response with at least three warmth ratings. Although there was an effect of interval, $F(2, 30) = 5.84$, there was neither an effect of correctness of response, nor an interaction between type of response and interval, Fs < 1. The means for the last three intervals are provided in Table 1.

The insight pattern analysis, however, suggested a difference between correctly and incorrectly solved anagrams. The proportions of anagrams that showed the insight pattern was marginally significantly greater for correctly than for incorrectly solved anagrams, $z = 1.63$ (one tailed), as shown by the means


CORRECT SOLUTION
INCORRECT SOLUTION
RESPONSE
-10 SEC
-20 SEC
-30 SEC
-40 SEC
PROPORTION
.6
.4
.2
1 2 3 4 5 6 7 8 9 10
WARMTH
WARMTH

presented in Table 2. Of the anagrams that were correctly solved, 79% showed an insight-like pattern whereas only 66% of the incorrectly solved anagrams showed an insight pattern.

Finally, there was a significant difference between the mean warmth ratings given at time of solution, depending upon whether the solution given was right or wrong, $t(16) = 3.92$. The warmth rating upon correct solution was 9.78; upon incorrect solution it was 7.86. This difference in warmth at the time the solution is given may provide a hint about why the warmth ratings in the two previous experiments were higher for the incorrectly solved problems than for the correctly solved problems.

Discussion

People seemed to use the warmth ratings in this experiment to indicate an uneasiness about their wrong answers. This suggests that the high-warmth pattern that is sometimes found with incorrect solutions may have something to do with a satisficing strategy, noted by Simon (1979) as "aiming at the good when the best is incalculable, . . . some stop rule must be imposed to terminate problem-solving activity. The satisficing criterion provides that stop rule: search ends when a good-enough alternative is found" (p. 3). In this experiment, but not the two previous experiments, subjects were allowed to write an answer with a less than 10 warmth rating. It seemed possible that the high-warmth pattern with wrong answers might have been due to a similar strategy on the part of subjects, accompanied by a process wherein the subject convinces himself that a good-enough answer is acceptable and the warmth ratings increase with increasing conviction about the answer. Of course, in the problems and anagrams used in the present experiments, there was a correct answer, and the good-enough answers given by subjects were wrong (assuming the right answer was rarely given as a good-enough answer).

If this conjecture about the locus of the high warmth ratings for incorrect solutions is correct, then it should be the case that if people are required to produce a 10 with their response, the higher warmth for incorrectly than for correctly solved anagrams should show up, as it did in the problem-solving experiments. In Experiment 4, subjects are required to produce a 10 before responding.

Experiment 4

Method

This experiment was similar to Experiment 3 except that (a) the instructions were different, (b) the lower boundary of the warmth ratings was zero rather than one, and (c) the subjects may have represented a different population. The instructions differed from those of Experiment 3 in stating (after an explanation of what warmth ratings are) that "if you don't know what the word is write down a zero; if you are close to the word write down a number between zero and ten; when you have the word write down a ten and the word next to it." The subjects were summer students in a second-year psychology course at the University of British Columbia who received a small bonus course credit for participating, and other people who were recruited by notices posted around campus and who were paid $4 for participating. There were 41 subjects in the experiment. One subject was eliminated, because she had been in an experiment that used some of the same anagrams and remembered some of the solutions, leaving 40 subjects.

Results and Discussion

Subjects were included in this analysis if they had at least one correct and one incorrect answer with three or more warmth ratings. Twenty-three subjects provided usable data. There was an effect of response type such that warmth was higher on the incorrect responses (2.11) than on the correct responses (1.66), $F(1, 22) = 4.0$, $MS_e = 10.23$, $p = .055$. There was an effect of interval such that warmth increased over interval, $F(2, 44) = 13.82$, $MS_e = .73$. There was no interaction between response type and interval, $F < 1$. These results are shown in Table 1. They suggest that when subjects are required to give a confident response, they produce higher warmth ratings on the incorrect than on the correct responses.

As is shown in Table 2, the difference in proportion of anagrams that were solved by an insight-like process did not vary for the correctly and incorrectly solved anagrams. There was a tendency for more incorrectly (.22) than correctly (.15) solved anagrams to fall into the incremental pattern, but the difference did not reach a significant level, $z = 1.35$. The direction of the pattern, however, is the same as was found in the previous experiments, a tendency to show more of an insight pattern and/or less of an incremental pattern on the correctly solved problems than the incorrectly solved problems.

It appeared that encouraging subjects to provide a 10 response resulted mainly in the appearance of higher warmth ratings for incorrectly solved anagrams than correctly solved anagrams.

Experiment 5

The results of Experiments 3 and 4, taken together, suggest that subjects might sometimes be using a satisficing strategy, or convincing themselves to accept a good-enough answer when they make mistakes. It might be possible to increase the use of such a strategy by encouraging subjects to guess. It would be expected, under such encouragement, that the warmth ratings should be higher. In addition there should be an increase in the tendency to show a noninsight rather than an insight pattern of warmth ratings. Such a strategy would also, of course, result in more errors in the anagram-solving situation in which there are correct answers.

Method

The method was identical to that of Experiment 4 except that subjects were also told: "Some of the anagrams in this experiment are solved by rather uncommon words, so do not be afraid to guess . . . you will

Figure 4. Distributions of warmth ratings for four intervals leading up to correct (left panel) and incorrect (right panel) responses to anagrams, Experiment 3.

probably be right," and at the end of the instructions, "Again, do not be afraid to guess." Subjects were also told, in this experiment as in the last, that they had to write down a 10 when they had the answer.

The subjects in this experiment were 24 summer students at the University of British Columbia, who received a small bonus course credit for participating, and volunteers recruited on campus for pay. Although this experiment was conducted a few weeks after Experiment 4, the subjects were from the same pool and there were no obvious systematic differences in the subject population.

Results

As in the previous experiments, an analysis was conducted on the last three warmth ratings for those 20 subjects who had at least one correct answer and one wrong answer with three or more ratings. As shown in Table 1, the warmth on the incorrect solutions (3.03) was significantly higher than was that on the correct solutions (2.33), $F(1, 19) = 6.85$, $MS_e = 13.11$. There was an effect of interval, such that warmth increased over the last three intervals, $F(2, 38) = 18.19$, $MS_e = .69$. There was no interaction between response type and interval. Thus, in this experiment, as in Experiments 1, 2, and 4, a high feeling of warmth was characteristic of the incorrect solutions.

A large proportion of the correctly solved anagrams showed an insight-like warmth pattern (.78). This proportion was significantly greater than that of the incorrectly solved anagrams (.44), $z = 5.93$. Thus, this analysis reveals that when subjects are encouraged to guess an insight-like pattern usually appeared with the correct not the incorrect answers. As might be expected, a larger proportion of incorrectly solved than correctly solved anagrams showed an incremental pattern of warmth ratings, as is shown in Table 2.

Comparison of Experiments 4 and 5

Because the subject populations were about the same and the experiments differed only in terms of instructions to subjects, it seemed appropriate to compare the results of Experiments 4 and 5 in order to look more directly at the effect of the guessing instructions on warmth ratings and on proportion correct.

The most striking discrepency between the two experiments was a difference in the proportion of incorrectly solved anagrams that were solved by an insight-like process, as shown in Table 2. In the no-guessing experiment, even incorrectly solved anagrams were solved by an insightlike process most of the time (i.e., .76). In contrast, in the guessing experiment, the incorrectly solved problems did not show the insight pattern in most cases (.44). The difference in these proportions was significant, $z = 3.88$. There was no difference in the proportion of correctly solved anagrams that were solved by an insight-like process. This finding supports the idea that the guessing instructions encouraged a satisficing strategy that is shown by the more incremental pattern of warmth ratings on the commission errors in Experiment 5 than in Experiment 4.

As might be expected, there was also a difference in the rate of correct and incorrect responses: The rate of correct responding was greater in the no-guessing experiment (.64) than in the guessing experiment (.55). The proportion of errors of commission—incorrect solutions—was smaller in the no-guessing experiment (.09) than in the guessing experiment (.27). And the proportion of anagrams for which no solutions were given was higher in the no-guessing experiment (.27) than in the guessing experiment (.18). An ANOVA comparing the proportion correct and incorrect in the two experiments showed that the interaction was significant, $F(1, 62) = 12.46$, $MS_e = .097$. These effects indicate that the instructions were effective at getting people to guess and to make errors of commission in the guessing experiment.

To investigate the effect of the guessing instructions on the magnitude of feeling of warmth judgments, a $2 \times 2 \times 3$ ANOVA with unequal n, in which the factors were experiment (guessing/no guessing—between subjects), type of response (correct/wrong—within subjects), and interval (third last/second last/last—within subjects) was conducted. There was a trend toward an effect of experiment, $F(1, 41) = 2.22$, $MS_e = 109.43$, $p = .07$ (one-tailed). The means showed higher warmth for the guessing experiment (2.68) than for the no-guessing experiment (1.89). As was the case in both experiments separately, there was an effect of type of response such that incorrectly solved anagrams (2.57) showed a higher anticipatory warmth than did correctly solved anagrams (1.99) $F(1, 41) = 11.03$, $MS_e = 11.56$. There was an effect of interval such that warmth increased over the last three intervals, $F(2, 82) = 31.82$, $MS_e = .711$. None of the interactions was significant (all Fs < 1).

The results of the comparison between Experiments 4 and 5 provide support for the idea that high and/or increasing warmth ratings accompany a satisficing strategy. In tasks that have a clear-cut answer (like those under investigation here), such a strategy frequently results in the wrong answer.

Additional Analyses

Nothing has been said so far about the warmth ratings that accompany the unsolved problems. It seemed likely that these ratings would be very low, because in many cases it would be expected that subjects would have no idea of even an approach to these problems. Krinsky and Nelson (1985) have found that feeling of knowing on errors of omission are lower than on errors of commission in a memory task. The omission errors might be analogous to the no-solution problems. The unsolved problems may also provide a quasicontrol comparison condition, revealing the warmth patterns under conditions in which subjects might be expected to have minimal information available. To compare the unsolved with the correctly solved problems, it is important to equate for time, because the unsolved problems necessarily have ratings for the maximum solving period, whereas the solved problems do not. Experiment 1 is not considered because the data were not within subjects. Subjects who had any answers correct with at least four warmth ratings and who had at least one unsolved problem were included in the analysis. Means over problems for each solution interval were computed on the correct solutions for each subject. These means were also computed on the unsolved problems, up to the last interval that was represented in the correctly solved protocols. Then the data for both correctly solved and unsolved problems were divided into four equal parts, and the mean warmth rating for each of the four parts was computed. These

Table 3
Temporal Comparison of Warmth Ratings for Correctly Versus Unsolved Problems or Anagrams in Experiments 2, 3, 4, and 5

	Quarters			
Type of problem	1	2	3	4
Experiment 2 (problems)				
Correct	3.69	3.86	3.96	4.53
No solution	3.04	3.19	3.05	2.84
Experiment 3 (anagrams)				
Correct	1.65	1.87	2.01	2.13
No solution	1.70	1.91	2.10	2.21
Experiment 4 (anagrams, no guessing)				
Correct	1.43	1.70	1.68	1.61
No solution	1.36	1.47	1.51	1.43
Experiment 5 (anagrams, guessing)				
Correct	1.54	1.64	1.87	2.45
No solution	1.26	1.50	1.49	1.52

means were analyzed by ANOVAs, and the results are shown in Table 3.

In Experiment 2 the correct responses showed higher warmth than did the no-solution problems, $F(1, 20) = 12.42$, $MS_e = 25.95$. There was no increase in warmth with quarter but there was an interaction between quarter and response, $F(3, 60) = 4.15$, $MS_e = .605$. In Experiment 3 there was no effect of response, nor was there an interaction between response and quarter. There was an effect of quarter, however, such that warmth increased, $F(3, 66) = 5.69$, $MS_e = .38$. In Experiment 4 the correctly solved anagrams showed higher warmth than the no-solution anagrams, $F(1, 38) = 5.39$, $MS_e = 3.19$. There was no effect of quarter nor was there an interaction between quarter and type of response (Fs < 1). In Experiment 5 there was no main effect for correctly solved versus no-solution anagrams. There was an effect of quarter, $F(3, 54) = 3.53$, $MS_e = .66$, and there was an interaction between response and quarter, $F(3, 54) = 3.33$, $MS_e = .34$.

Overall it does not appear that the correct-response warmth ratings were on floor, since they were often higher than were the temporally equated no-solution ratings. The low warmth ratings on the unsolved problems and anagrams are consistent with Krinsky and Nelson's (1985) feeling-of-knowing data. The difference between warmth on the correctly and incorrectly solved problems and anagrams is probably not the result of simple default values being given in the correct-response cases.

Conclusion

The results of these experiments bear upon whether problems and anagrams are solved by a sudden insight process or by more gradual accumulation of information. The data are not completely clear-cut, but some conclusions can be drawn. Most (76%) of the problems and anagrams that were correctly solved showed no more than a 1-point increase in warmth over the entire problem-solving interval, until the warmth suddenly jumped to 10 with the solution. Thus, most of the correctly solved problems exhibited a subjectively catastrophic process that could be defined as insight. It would be of great interest to show that other kinds of questions or problems do not show an insight pattern of warmth ratings. Within the present set of experiments there is a tendency—shown in all five experiments—for incorrectly solved problems or anagrams to show the insight pattern to a lesser extent or an incremental pattern to a greater extent than do the correctly solved problems. Of the correctly solved problems, 76% showed the insight pattern, whereas only 52% of the incorrectly problems and anagrams showed this pattern, $z = 7.97$. The incorrectly solved problems and anagrams (.25) were more likely to show an incremental pattern than were the correctly solved problems and anagrams (.12), $z = 5.73$. Thus it does not appear that the insight pattern is the only pattern that people will or can use; on the incorrectly solved problems, they are less likely to use it.

If the insight pattern was the rule, why then was there a consistent increase over the last three intervals when the warmth means were calculated? First of all, because most people started with a zero warmth, the scale itself constrained warmth to increase. Despite this constraint, most of the protocols show no increase in warmth. A minority (24%) of the correctly solved protocols did not conform to the definition, given above, of an insight pattern. Many of the nonconforming protocols were fairly low and consistent until the last interval in which the subject gave a high warmth value, just before overtly giving the solution. (As noted earlier, high warmth values disproportionately increase the means.) The reasons for this increase in warmth are not yet known. It could be that subjects were gradually accumulating information in this minority of cases. Alternatively, the insight may have already occurred, but not yet been articulated. This inarticulate (but solved) state might be given a high warmth rating by subjects. Perhaps some subjects do not provide the answer and a warmth rating of 10 until they have checked their answer for correctness. Thus, the shift from the unsolved to the solved state may not be instantaneous. How rapidly the shift in metacognition has to occur if the reflected process is to be considered to be insight is an open issue.

Four of the five experiments also indicated that the warmth ratings were higher just before subjects give the wrong answer to problems than just before they give the right answer. A strong premonition of solution is liable to be a marker that the wrong answer will follow. This result does not appear to be attributable to a simple increment in warmth over time with incorrect answers taking longer than correct ones; in Experiment 2 the correct solutions took longer than the incorrect ones, and yet still the effect obtained. The effect does not appear to be entirely attributable to individual differences; it obtains in within-subjects as well as in between-subjects designs. The misleading nature of some insight problems does not seem to be the reason for the effect, insofar as it obtains with anagrams as well as with insight problems. It also does not appear to be the result of a strategy of producing low-default warmth values thus leaving more cognitive resources for problem solving; the warmth was

often lower on unsolved problems than on problems to which solutions were given, indicating that the warmth on correctly solved problems was not consistently on the floor.

The high-warmth effect on the incorrectly solved problems may be attributable to a process in which the subject convinces him- or herself that an inelegant solution is adequate. The differences between Experiments 4 and 5 in which subjects were or were not encouraged to use such a satisficing strategy are relevant. When instructed to guess, subjects produced more wrong answers. More interesting, however, was the fact that their warmth ratings tended to be higher when subjects were encouraged to accept a good enough answer (Experiment 5). There was also an increased tendency, when satisficing was encouraged, to produce warmth protocols that did not conform to the insight pattern, especially on incorrectly solved anagrams. The satisficing strategy is probably normal in many real-life and social situations, although it leads to error in solving insight problems or anagrams, in which there are ideal solutions.

To conclude with the practical question of whether one should believe people when they say that they have almost got the answer to a difficult insight problem, results of the present series of experiments suggest that we should be wary: At least with the class of problems studied here, premonitions of insight predict mistakes.

References

Bartlett, F. C. (1958). *Thinking.* London: Allen and Unwin.

Bowers, K. S. (1985, May). *Memory, intuition, and Meno's paradox.* Paper presented at the meeting of the Society for Philosophy and Psychology, Toronto, Ontario, Canada.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5,* 325–337.

Dominowski, R. L. (1981). Comment on "An examination of the alleged role of 'fixation' in the solution of 'insight' problems." *Journal of Experimental Psychology: General, 110,* 199–203.

Ellen, P. (1982). Direction, past experience, and hints in creative problem solving: Reply to Weisberg and Alba. *Journal of Experimental Psychology: General, 111,* 316–325.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Eysenck, M. W. (1979). The feeling of knowing a word's meaning. *British Journal of Psychology, 70,* 243–251.

Freedman, J. L., & Landauer, T. K. (1966). Retrieval of long-term memory: "Tip-of-the-tongue" phenomenon. *Psychonomic Science, 4,* 309–310.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208–216.

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior, 6,* 685–691.

Kahnemen, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Krinsky, R., & Nelson, T. O. (1985). The feeling of knowing for different types of retrieval failure. *Acta Psychologica, 58,* 141–158.

Levine, M. (1975). *A cognitive theory of learning.* Hillsdale, NJ: Erlbaum.

Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monographs, 54* (6, Whole No. 248).

Maier, N. R. F. (1930). Reasoning in humans. I. On direction. *Journal of Comparative Psychology, 10,* 115–143.

Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 288–294.

Nelson, T. O. (1984). A comparison of current measure of the accuracy of feeling of knowing predictions. *Psychological Bulletin, 95,* 109–133.

Nelson, T. O., Leonesio, R. J., Shimamura, A. P., Landwehr, R. F., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 279–288.

Oskamp, S. (1962). The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs, 76,* (28, Whole No. 547).

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

Shiffrin, R. M., & Schneider, W. (1977). Control and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84,* 127–190.

Simon, H. A. (1979). *Models of thought.* New Haven: Yale University Press.

Simon, H. A., Newell, A., & Shaw, J. C. (1979). The process of creative thinking. In H. A. Simon (Ed.), *Models of Thought* (pp. 144–174). New Haven: Yale University Press.

Sternberg, R. J. (1985). *Beyond IQ.* Cambridge, England: Cambridge University Press.

Sternberg, R. J., & Davidson, J. E. (1982, June). The mind of the puzzler. *Psychology Today,* pp. 37–44.

Sternberg, R. J., & Davidson, J. E. (1983). Insight in the gifted. *Educational Psychologist, 18,* 51–57.

Weisberg, R. W., & Alba, J. W. (1981). An examination of the alleged role of "fixation" in the solution of several "insight" problems. *Journal of Experimental Psychology: General, 110,* 169–192.

Weisberg, R. W., & Alba, J. W. (1982). Problem solving is not like perception: More on gestalt theory. *Journal of Experimental Psychology: General, 111,* 326–330.

Yarmey, A. D. (1973). I recognize your face but I can't remember your name: Further evidence on the tip-of-the-tongue phenomenon. *Memory & Cognition, 1,* 287–290.

Received August 31, 1985
Revision received December 18, 1985 ■