*Memory & Cognition*
*1996, 24 (6), 687-699*

# Partial matching in the Moses illusion: Response bias not sensitivity

ELEEN N. KAMAS, LYNNE M. REDER, and MICHAEL S. AYERS
*Carnegie Mellon University, Pittsburgh, Pennsylvania*

Previous research has demonstrated that people have enormous difficulty in detecting distortions in such questions as, "How many animals of each kind did Moses take in the Ark?" Reder and Kusbit (1991) argued that the locus of the effect must be the existence of a partial-match process. Other research has suggested that this partial-match process operates at the word level and that, with adequate focus on the relevant word, the Moses illusion is greatly diminished. The present experimental results argue that those conclusions were based on a shift in response criterion with no concomitant change in ability to detect distortions. Furthermore, the data suggest that the matching process operates below the word level, at the level of distinctive features.

Understanding the nature of access to memory—how we parse questions, query memory, and decide that the requisite information has been found—is one of the central issues in understanding memory. An important premise in our work is that the matching process in memory must be a partial-matching process. This may not seem obvious until one considers, for example, that information is often not queried in exactly the same form as originally presented or encoded, and that people do not look exactly the same from one occasion to the next; thus, person recognition must be flexible. Given that we believe that the memory-match process must be a partial-match process, the questions become: How much must the memory probe overlap with the memory trace to be accepted as a match? What portions of the memory probe are used to match to memory?

One fruitful approach to these questions is to look at instances in which this partial-matching system leads people astray. For example, when asked, "How many animals of each kind did Moses take in the Ark?" most people immediately respond with "Two." This confident answer comes even from those who know that Noah, not Moses, built the Ark (Erickson & Mattson, 1981). When a term in a sentence or question (the "critical" term) is replaced with a semantically similar but incorrect term (the "distorted" term), people sometimes respond as if this distortion were not present. This tendency to overlook distortions in statements is known as the *Moses illusion*. Studying when the Moses illusion occurs and what factors influence it can shed light on the underlying memory processes in question answering and text comprehension.

The work reported here was sponsored by Grant BNS-8908030 from the National Science Foundation to L.M.R. and by a National Science Foundation Graduate Research Fellowship to E.N.K. We thank G. Kusbit for her assistance in conducting Experiments 1 and 3 and Jason Wyse for comments on the manuscript. Correspondence concerning this paper should be addressed to L. M. Reder, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 (e-mail: reder+@cmu.edu).

This paper continues the line of research begun by Reder and Kusbit (1991), in which they concluded that the best explanation for the Moses illusion was imperfect matching of queries to memory. They arrived at this conclusion after investigating several alternate hypotheses. These included: (1) the cooperative principle, that is, the listener notices the distortion, but ignores it to cooperate with the speaker; (2) imperfect encoding, that is, the person simply did not "hear" the incorrect term in the sentence; (3) imperfect memory retrievals, that is, the question is correctly heard but the information retrieved from memory is incomplete; and (4) imperfect matching of the question terms to memory.

Together, Reder and Kusbit's experiments and evidence from previous research eliminate all but the final hypothesis. The cooperative principle generates the hypothesis that the illusion results from subjects' cooperating with the experimenter and overlooking the distortions. If true, subjects should find it easier to detect distortions than to ignore them and answer the gist of the question. To the contrary, Reder and Kusbit's subjects found it easier, not harder, to answer the gist of a question than to detect distortions. Response times were faster and errors fewer in the gist condition than in the literal condition, in which detection of distortions was required. Furthermore, rather than using a question-answering task, several other experimenters (e.g., Bredart & Modolo, 1988; van Oostendorp & de Mul, 1990) used a sentence-verification task, which should not be subject to the cooperative principle. In those studies there was also a high rate of illusion.

The second hypothesis was that the Moses illusion was caused by failure to encode the distorted element, that is, the critical term simply was not read or heard. If true, subjects should spend less time reading the distorted term when they fail to notice the distortion. Word-by-word reading times collected while subjects tried to answer questions and notice distortions gave no support to this hypothesis. If anything, subjects took longer to read the distorted term when the distortion went unnoticed than

when it was detected (Reder & Kusbit, 1991, Experiment 4). This is consistent with van Oostendorp and de Mul's (1990) results, in which failures to detect distortions were found to be slower than detection. They had speculated that this result was an artifact of their experimental procedure; however, the Reder and Kusbit data indicate otherwise.

The third hypothesis Reder and Kusbit considered was that the failure to detect distortions results from imperfect memory retrieval, such that the retrieved memory trace does not include the information needed to detect the distortion. A manipulation intended to make relevant information more accessible by having subjects study the queried facts prior to answering the questions succeeded in significantly improving the subjects' knowledge of answers and speed of responding; however, it did not increase the rate of detection of the distortions (Reder & Kusbit, 1991, Experiments 2, 3, and 5), suggesting that access to the requisite knowledge was not the problem.

The hypothesis that remained was that the illusion results from an incomplete or partial-match strategy. As a question is read, the terms or concepts are matched to memory so that the answer may be retrieved. Not every word or concept in the question will be matched exactly to a corresponding memory structure, however. A criterion level will be set for a given situation, and the concepts in the question will be checked for overlap with the remainder of the sentence, with the criterion level determining how much overlap must be present. For example, since Moses is a biblical character and is thus loosely related to Noah, he will sometimes be accepted in the question about the ark, while Nixon, a modern politician, will never be accepted in this sentence.

Using the partial-match hypothesis as our premise, we designed the following experiments in an attempt to answer several questions about this partial-match process. For instance: (1) Is the failure to notice mismatches caused by insufficient attention to the terms in the question or to aspects of the memory representation? (2) Is the partial match due to insufficient resource capacity or is it a race between competing processes? (3) At what level does the partial match occur? Is it at the word or concept level, or is it at the feature level? (4) How robust is the partial-match process? Can one make subjects more sensitive to distortions by encouraging them to attend to the distorted words, or do efforts to make subjects attend to the distorted terms only bias them toward calling any sentence distorted?

In this series of experiments, we employed a technique previously unused in this line of research. By using signal-detection analysis, we were able to determine whether differences found between our experimental and control conditions were due merely to a criterion shift or whether they reflected a true change in sensitivity to distortions in the questions.

## EXPERIMENT 1

Research by Bredart and Modolo (1988) and Bredart and Docquier (1989) suggested that people's ability to detect a distortion can be influenced by changing the focus of a sentence. Bredart and Modolo manipulated the focus by using cleft sentences (e.g., "It was Moses who took two animals of each kind . . ." versus "It was two animals of each kind that Moses took . . ."), while Bredart and Docquier varied whether MOSES or another word from the sentence to be verified was all capitalized. Both studies found that subjects were more likely to reject Moses when this distorted term was in the focus. Similarly, Baker and Wagner (1987) found that when a distortion was placed in the portion of the sentence syntactically marked as given, subjects more often overlooked the distortion. This result was predicted because given information in a sentence is not processed as thoroughly as new information, while new information, such as the information in the focus of a sentence, is examined more carefully (Hornby, 1974).

These results support the hypothesis that partial matching is responsible for the Moses illusion, and that subjects can be made more sensitive to distortions. The present experiment was intended to see whether subjects could be made more sensitive to distortions merely by influencing their memory structures. In Experiment 1, a study condition was used to vary the emphasis in memory rather than in the question to be answered. Subjects studied the relevant information ahead of time, with the critical information all capitalized. Of interest was whether the portion of the memory trace that had been emphasized in the study phase would be inspected more carefully during the questioning phase. To this end, subjects studied the fact relevant to answering the question with the critical term capitalized (e.g., "NOAH took two animals of each kind into the ark") with the answer capitalized (e.g., "Noah took TWO animals . . ."), or with no word capitalized.

### Method

**Design and Materials.** Seventy-two pairs of questions like the Noah/Moses question were used, including 45 of the pairs listed in Reder and Kusbit (1991). Some of the earlier questions were discarded because they were outdated or because subjects could not answer the undistorted form. Twenty-seven new pairs of questions were constructed according to the following criteria: The query tested common knowledge, the distorted term was semantically and syntactically similar to the term it replaced, and the distorted form of the question could not be answered literally (see Reder & Cleeremans, 1990; Reder & Kusbit, 1991).

The form of the sentences used for study was the complete sentence answers to the undistorted form of the questions, containing as many as possible identical clauses to the questions (but never the distorted term). Regardless of condition, the study form of the fact did not vary; that is, all subjects studied "Noah took two animals of each kind on the ark." The only difference was whether "NOAH" was capitalized, "TWO" was capitalized, or nothing was capitalized.

For each subject, question pairs were assigned randomly to condition (i.e., either normal or distorted) and to one of the three capitalization procedures. Twelve questions were assigned to each of the six conditions.

**Procedure.** Subjects were presented sentences one at a time for study in random order on an IBM PC monitor. They were instructed to study each statement carefully and to pay special attention to any capitalized words. They were told that they would be responsible for this information in a later part of the experiment. The subjects

controlled the speed of presentation of the statements, with a minimum exposure of 5.5 sec per sentence.

After studying the statements, the subjects began the question-answering phase. They were instructed to respond as quickly and as accurately as possible to each question. They were told to respond "can't say" to distorted questions, to give the answers to undistorted questions, or answer "don't know" if they did not know the answer. Distorted questions were defined as those for which "literally there is no answer to the question (e.g., On what holiday do people sing carols, decorate a tree and top it with a BUNNY RABBIT?)." The subjects initiated the display of each question by pressing a button on a button box attached to the computer. They spoke their answers into the microphone connected to a voice key. Once the subject's vocal response tripped the voice key, the correct answer (e.g., "can't say" or "Christmas," depending on whether the question was distorted or normal) appeared on the screen. The experimenter noted invalid response time (RT) measurements that occasionally occurred when the voice key failed to trip or when there was ambient noise. Invalid RTs were excluded from the analyses. The experimenter also coded each answer as correct, incorrect, "don't know," or "can't say." The entire experiment took approximately 40 min.

**Subjects.** Twenty-nine Carnegie Mellon University undergraduates participated either to partially fulfill a course requirement or for pay.

## Results and Discussion

Two subjects' data were excluded from the analyses, one due to a technical failure and the other because of a stuttering problem that invalidated the articulation RTs. Below, we report the data in two ways in order to facilitate comparison with prior studies, some of which reported one measure, some the other. One is the proportion of distorted questions correctly identified; the other is the rate of illusion, which is the proportion of times the distortion was overlooked. Table 1 reports the proportion correct and proportion of each error type: illusions, "can't say" responses to undistorted questions, "don't know" responses, and incorrect responses. Illusion rate is also shown in Figure 1.

An analysis of variance (ANOVA) was performed on the accuracy data and the latency to give the correct answer, using as factors whether the question was distorted or not and capitalization during study (none vs. answer vs. target). Separate analyses were performed that collapsed over questions (which had been assigned randomly to condition) and over subjects, treating questions as random effects. RTs did not differ among the conditions (all $F$s < 1) and will not be discussed further.

As expected, distorted questions were answered less accurately than normal questions [over subjects, $F(1,26) = 132.39, p < .001$; over items, $F(1,71) = 68.48, p < .001$]. Of interest was the effect on accuracy of the type of capitalization in the studied fact. Although there was no main effect of study condition on accuracy ($F$s < 2), there was a significant distortion × capitalization interaction [over subjects, $F(2,52) = 6.31, p < .01$; over items, $F(2,142) = 4.19, p < .05$]. Studying a statement with the critical term capitalized *increased* the accuracy for the *distorted* questions as compared with capitalization of the answer term, whereas study with the critical term capitalized tended to *decrease* accuracy for *undistorted* questions. Bonferroni's method for multiple comparisons was used to compare performance on distorted questions on the basis of study condition, with $\alpha = .05$. The detection of distortions was reliably better only when the studied sentences had the critical term capitalized as opposed to when the answer term was capitalized; however, neither of these conditions differed significantly from the no capitalization condition.

Analysis of the illusion rate (the proportion of distorted questions that were answered as if undistorted) showed a pattern that was similar to the accuracy data. The rate of illusion was highest when the answer term had been capitalized in the study sentence (33%), followed by no special capitalization (29%). When the critical term had been capitalized in the study sentence, the illusion rate was lowest (24%). A Tukey contrast among illusion rates, with $\alpha = .05$, revealed a significant difference between the condition in which the critical term had been capitalized during study and the condition in which the answer term had been capitalized, although neither of these terms differed from the no-capitalization control.

Given that the capitalization manipulation improved performance on distorted statements by lowering the illusion rate but raised the error rate for undistorted questions by increasing the number of spurious "can't say" re-

**Table 1**
**Mean Proportion Correct, Proportion of Error Types, and Mean Latency (in Seconds, in Parentheses) for Questions in Experiment 1 as a Function of Capitalization During Study and Whether the Question Was Distorted**

| Part of Study Sentence Capitalized | Correct | Types of Errors* | | | |
|---|---|---|---|---|---|
| | | Undistorted | CS | DK | Incorrect |
| **Distorted Questions** | | | | | |
| Critical term ("NOAH") | .75 (3.28) | .24 (3.44) | – | .02 (4.59) | 0 |
| Answer ("TWO") | .63 (3.30) | .33 (3.47) | – | .02 (3.64) | .02 (5.94) |
| None | .70 (3.19) | .29 (3.38) | – | .01 (5.05) | 0 |
| **Undistorted Questions** | | | | | |
| Critical term ("NOAH") | .90 (3.33) | – | .04 (3.88) | .04 (4.99) | .02 (3.66) |
| Answer ("TWO") | .94 (3.28) | – | .03 (4.20) | .02 (3.73) | .01 (2.98) |
| None | .93 (3.28) | – | .03 (5.18) | .03 (4.66) | .01 (4.00) |

*The types of errors are answering distorted questions as if they were undistorted (Undistorted), "can't say" responses to undistorted questions (CS), "don't know" (DK), and incorrect responses.

sponses, we wanted to determine whether the subjects were actually better at detecting distortions or had simply shifted their bias for calling a question distorted on the basis of the sentence form during study. We calculated nonparametric measures of sensitivity and bias (see Snodgrass & Corwin, 1988, and Donaldson, 1992), $A'$ and $B''d$, for each subject by defining hits as the proportion of correctly detected distorted questions (i.e., the proportion of "can't say" responses to distorted questions) and false alarms as the proportion of "can't say" responses to undistorted questions.[1] The mean sensitivity scores in the three conditions were not significantly different (critical term capitalized, $A' = 0.896$; answer capitalized, $A' = 0.865$; no capitalization, $A' = 0.888$). This suggests that performance differences reflected a change in bias, not in sensitivity; calculation of $B''d$ supported this view. The subjects appeared biased against calling a question distorted; they were less likely to judge a question as distorted if the study sentence had the answer term capitalized ($B''d = 0.760$) than if the target term ($B''d = 0.580$) or nothing ($B''d = 0.642$) had been capitalized. This difference in bias approached significance [$F(2,52) = 2.57$, $p < .09$]. These results suggest that our manipulation of capitalization of the critical term during study only affected subjects' bias toward calling any question distorted and not their sensitivity to the distortions. We found this surprising and considered these measures further in the following experiments.

Given that subjects were significantly less likely to report a distortion in a question when the answer term had been capitalized during study than when the critical term had been capitalized, but neither manipulation differed reliably from the no-capitalization control, two different explanations seem viable; the difference in performance could be due to (1) making the critical term more salient or (2) making the answer more available. Experiment 3 tested the hypothesis that distortion detection depends on making salient the critical term in the question.

The answer-availability explanation involves the assumption that the question-answering process and the distortion-detection process operate in parallel, and that the subject's response depends on which process is completed first. This explanation might account for the results of Experiment 1, because it predicts that when the answer is made more salient during study, the process that searches for the answer would more often be completed before the process that searches for distortions. Experiment 2 was designed to test this parallel-processing hypothesis.

## EXPERIMENT 2

If the process that searches for the answer competes with the process that searches for distortions in the question, removing the process that tries to report the answer to the question should allow completion of the inspection process that searches for distortions. That there should be two separate processes for question answering and distortion detection is intuitively logical, and this idea is supported by instances in which subjects answered distorted questions and immediately afterwards noticed the distortions. To examine this empirically, some subjects in this experiment were asked only to monitor for distortions and not to answer the questions. Our hypothesis was that without the competing process of trying to answer the questions, these "single-task" subjects should be more accurate at detecting distortions than those required both to monitor for distortions and to answer undistorted questions.

### Method

**Materials.** Sixty-six of the pairs of test questions used in Experiment 1 were used; six other questions were eliminated for datedness and lack of knowledge by undergraduates. A posttest question was constructed for each of the test questions to verify that the subject knew the relevant information that might be distorted (e.g., "Who brought two animals of each kind into the Ark?").

**Design and Procedure.** The design and procedure were similar to those of Experiment 1. Questions were assigned randomly to be distorted or undistorted, and the subjects were assigned randomly to either the single-task condition or the standard experimental task, yielding a 2 × 2 mixed design.

Subjects assigned to the standard task attempted to answer every undistorted question and respond "can't say" to distorted questions. In the single-task condition, rather than responding orally, the subjects responded by pressing a button on a button box. One button was labeled "DISTORTED," for distorted questions, and another was labeled "NORMAL," for undistorted questions. RTs from the presentation of the question to the buttonpress were recorded. After the subject's response was recorded, the correct answer appeared on the screen below the question. Subjects proceeded to the next question at their own pace.

After all test questions were presented, subjects in both conditions answered the posttest questions. These questions were presented by computer, and the subjects typed their answers. This portion was self-paced and RTs were not recorded.

**Subjects.** Fifty-three Carnegie Mellon undergraduates participated in partial fulfillment of a course requirement or for pay. Twenty-six subjects were assigned randomly to the standard-task condition, and 27 were assigned to the single-task condition.

### Results and Discussion

The data from 1 subject in the single-task condition were discarded because the subject had previously participated in a related experiment. Approximately 2% of the trials were discarded due to inaccurate timing (e.g., inadvertent vocalizations or other noises which stopped the clock prematurely or answers spoken too softly to register). The subjects answered approximately 76% of the posttest questions correctly. Table 2 presents the means for accuracy performance, error rates, and RTs. Only the data for the questions for which the corresponding posttest question was answered correctly are reported here.

An ANOVA of the accuracy and response-time results revealed that undistorted questions were answered much more accurately [over subjects, $F(1,50) = 109.27$, $p < .001$; over questions, $F(1,64) = 47.04$, $p < .001$] and more quickly [$F(1,50) = 5.91$, $p < .05$] than were distorted questions. Accuracy was greater in the single-task condition than in the question-answering condition [over subjects, $F(1,50) = 38.77$, $p < .001$; over questions, $F(1,64) = 54.07$, $p < .001$], but subjects in the single-task condition responded to the questions more slowly [$F(1,50) = 6.20$,

**Table 2**
**Mean Proportion Correct, Proportion of Error Types, and Mean Latency (in Seconds, in Parentheses) for Questions in Experiment 2 as a Function of Task and Whether the Question Was Distorted**

| Task | Correct | Types of Errors* | | | |
|---|---|---|---|---|---|
| | | Undistorted | CS | DK | Incorrect |
| Distorted Questions | | | | | |
| Single task | .62 (5.21) | – | – | – | .38 (6.59) |
| Question answering | .40 (4.38) | .37 (4.41) | – | .19 (4.92) | .04 (4.86) |
| Undistorted Questions | | | | | |
| Single task | .89 (4.95) | – | – | – | .11 (5.56) |
| Question answering | .69 (3.91) | – | .05 (5.16) | .23 (4.80) | .04 (4.43) |

*The types of errors are answering distorted questions as if they were undistorted (Undistorted), "can't say" responses to undistorted questions (CS), "don't know" (DK), and incorrect responses.

$p < .05$]. Question type and task condition did not interact ($F$s < 1).

Clearly, the accuracy rates reveal a great improvement in performance when only detection of the distortion is required—supporting our hypothesis. On the other hand, an examination of the error rates gives a different story. Consider the illusion rate, that is, the proportion of responses to distorted questions as if undistorted. In the question-answering condition, the illusion rate was 37%, which is indistinguishable from the 38% error rate in the single-task condition. It is unclear how these data should be interpreted, because the tasks were quite different. Specifically, subjects gave binary responses in the single-task condition, so questions to which the subject might not know the answer could be scored as correct or as an error.

Given these equivocal results, we calculated sensitivity scores for each subject. For subjects in the single-task condition, hits were defined as correct identification of a distorted question as distorted, and false alarms as the incorrect identification of a normal question as distorted. For subjects in the standard-task condition, hits and false alarms were defined as in Experiment 1. Mean sensitivity to distortions did not differ between the two tasks [$A'$ = 0.792 for question answering and $A'$ = 0.827 for single task; $F(1,50) = 2.21, p > .10$]. In contrast, response bias did significantly differ for the two tasks [$B''d$ = 0.573 for the single-task condition vs. $B''d$ = 0.889 for the standard task; $F(1,50) = 34.31, p < .05$].[2] This means that eliminating the "distraction" of answering the question did not make subjects more sensitive to distortions, but only increased their bias to call any question distorted. This was true despite the fact that subjects were much slower in the single task than in the conventional task. It is worth noting here that even when subjects shifted their criterion to detect more distortions, they were still strongly biased toward giving the undistorted answer ($\beta$ = 5.3). This once again underscores the robustness of the phenomenon and the difficulty of the distortion-detection task.

Both Experiment 1 and Experiment 2 showed little evidence that subjects can actually improve their detection of distortions. In Experiment 1, the improvement in detection when the critical term was emphasized during study was surprisingly small. Perhaps the manipulation was just too subtle. On the other hand, we were surprised that the difference in both experiments seemed attributable to a change in bias, rather than to true ability to detect mismatches (distortions) in the questions. Previous research had found an improvement in the rate of distortion detection with an explicit focus in a sentence-verification paradigm (Bredart & Modolo, 1988), but perhaps if those experiments had used undistorted controls for the critical items (e.g., "It was Noah who took" ... for the Moses question), the results would have shown only a change in bias.

Bredart and Modolo's design does not provide conclusive evidence for increased sensitivity to distortions, because it does not allow for analysis of response bias; however, we wondered whether there would be some conditions under which sensitivity increased, independent of changes in response criterion. Presumably, if subjects were explicitly told which word in a question might be distorted, their detection rate would improve. In Experiment 3 (A and B), we tested whether the detection of the distortion could be improved by capitalizing the distorted element within the question itself. Our expectation was that if subjects were asked, "How many animals of each kind did MOSES take on the ark?" they would have no difficulty detecting this distortion. The important question was, if such an improvement were found, would it be due to an improved ability to detect distortions or merely a shift in criterion for stating that a question was distorted?

## EXPERIMENT 3A

As in Experiment 1, subjects attempted to answer questions while monitoring for distortions. There was no study phase prior to question answering; for some of the questions, however, the critical term (e.g., "NOAH" or "MOSES") was capitalized. The ability to detect distortions in questions having the critical term capitalized was compared with that to detect distortions in questions in which the critical term was not capitalized. If the partial match process is one where some words are carefully matched while others are not, then the capitalization manipulation should increase the probability of matching

the target word and thereby improving the detection of distortions.

## Method

**Materials, Design, and Procedure.** The design and procedure were similar to the question-answering phase of Experiment 1. Subjects were asked to answer each undistorted question, but to respond "can't say" when they detected a distorted question. The 72 pairs of questions (distorted and undistorted) from Experiment 1 were used, with questions to be either distorted or undistorted randomly assigned and with the critical term being capitalized or not. The random assignment was done separately for each subject.

The subjects were informed that sometimes words in questions would appear in capital letters, and that this might help them in deciding whether the question was distorted but that it did not guarantee that the statement was distorted. Questions were presented on a computer screen, and the subjects spoke their answers into a microphone attached to a voice key that measured RTs. After the voice key was tripped, the correct answer appeared on the screen and the experimenter entered a code for the response. Subjects pressed a button on a button box to proceed to the next question. The experimenter was present for the entire experimental session, both to enter subjects' responses and to note when RTs were invalid due to either a premature vocalization or a vocalization that was too soft to trigger the voice key. The entire experiment took approximately 40 min.

**Subjects.** Forty Carnegie Mellon University undergraduates participated either for partial fulfillment of a course requirement or for pay. The subjects were native English speakers raised in the U.S.

## Results and Discussion

Less than 3% of the trials were discarded because of inaccurate timing due to inadvertent vocalizations or answers too quiet to trigger the voice key. ANOVAs were performed on both the accuracy data and the valid latency data, using as factors capitalization (critical term in capitals vs. nothing in capitals) and distortion (distorted vs. undistorted questions). In addition, an ANOVA was performed on the illusion rate data. Table 3 presents the accuracy of responses, types of errors, and latency of responses as a function of whether or not the question was distorted and whether or not there was capitalization in the question. As before, there were no reliable effects on latency, so our discussion will focus on the accuracy data.

Distorted questions were answered significantly less accurately than normal questions [over subjects, $F(1,39) = 59.66, p < .001$; over items, $F(1,71) = 32.63, p < .001$], but capitalization did not produce a main effect on accuracy. As in Experiment 1, capitalization interacted with distortion such that it improved performance on distorted questions but disrupted performance on the undistorted questions [over subjects, $F(1,39) = 12.70, p < .01$; over items, $F(1,71) = 14.95, p < .001$]. The illusion rate, which included only the distorted questions, was especially affected by capitalization [over subjects, $F(1,39) = 11.26$, $p < .01$; over items, $F(1,71) = 17.28, p < .001$].

It is important to note that the error rates for the undistorted questions also differed as a function of capitalization (see Table 3). Specifically, the proportion of "can't say" responses was higher when the question contained a word all in capitals (9%) than in the no-capitalization control condition (4%). Given that subjects' improvement in the detection of distortions with capitalization focus was offset by a decrement in accuracy for undistorted questions because of this increase in the "can't say" responses, it seems likely that capitalization was at least partially affecting response bias rather than probability of carefully matching the target word. To test this, we computed sensitivity ($A'$) and bias ($B''d$) measures for each subject. Capitalizing the critical term made subjects more sensitive to the distortions [$A' = 0.831$ with capitalization vs. $A' = 0.791$ with no capitalization; $F(1,39) = 5.87, p < .05$]. However, the subjects were also significantly more biased to call any question with capitalization distorted [$B''d = 0.642$ with capitalization vs. $B''d = 0.741$ without capitalization; $F(1,39) = 9.34, p < .01$].[3]

We were pleased that capitalizing the critical term in the sentence increased subjects' sensitivity to distortions, but disappointed that the increase in accuracy of detection of distorted questions was due at least as much to a shift in bias as to the change in sensitivity. While the first two experiments showed only a change in bias with no corresponding change in sensitivity, it is not very surprising that subjects would become more sensitive when the distortion was made very obvious with capital letters in the question. However, even with capitalization of the critical term, the task is clearly a very difficult one, so that subjects become more inclined to call a question distorted whenever they see capitalization. Perhaps this apparent shift in bias resulted from questions for which the sub-

**Table 3**
**Mean Proportion Correct, Proportion of Error Types, and Mean Latency (in Seconds, in Parentheses) for Questions in Experiment 3A as a Function of Capitalization in the Question and Whether the Question Was Distorted**

| | | Types of Errors* | | | |
|---|---|---|---|---|---|
| Capitalization | Correct | Undistorted | CS | DK | Incorrect |
| | | Distorted Questions | | | |
| Critical term capitalized | .62 (4.15) | .23 (4.66) | – | .13 (4.93) | .02 (6.39) |
| Nothing capitalized | .48 (4.10) | .36 (4.16) | – | .15 (4.89) | .02 (5.42) |
| | | Undistorted Questions | | | |
| Critical term capitalized | .74 (4.15) | – | .09 (5.54) | .12 (5.05) | .05 (4.57) |
| Nothing capitalized | .80 (3.90) | – | .04 (5.55) | .12 (5.29) | .03 (5.53) |

*The types of errors are answering distorted questions as if they were undistorted (Undistorted), "can't say" responses to undistorted questions (CS), "don't know" (DK), and incorrect responses.

jects did not have the knowledge required to discriminate between the distorted and undistorted versions of the question.

This possibility motivated us to replicate the study with a posttest to ensure that we included only trials for which subjects could recall the information (e.g., Noah) that might be distorted.[4] In addition, we tested the possibility that the instructions mentioning the capitalization may have induced subjects to use a strategy different from what they would otherwise have used. To check for this possibility, the replication included some subjects who received the same instructions as in Experiment 3A and some who received instructions that did not mention the capitalization that appeared in half of the questions.

## EXPERIMENT 3B

### Method

**Materials**. Sixty-four of the 72 questions used in Experiment 1 were included. The other 8 were discarded because of datedness or lack of knowledge by undergraduates. A posttest question was constructed for each of the test questions to verify that subjects knew the relevant information that might be distorted (e.g., "Who brought two animals of each kind into the Ark?").

**Design and Procedure**. This experiment was identical to Experiment 3A, except for the inclusion of the posttest and the alternate instructions that did not comment on capitalization. Approximately half of the subjects received the same instructions used previously; the other subjects received instructions that did not mention the capitalization in the questions. The posttest was administered after the subjects completed the question-answering/distortion-detection phase.

**Subjects**. Thirty-seven subjects participated in this experiment. Sixteen were undergraduate Carnegie Mellon students who participated in order to fulfill a course requirement; 21 other members of the campus community, who were recruited from an advertisement posted on a Carnegie Mellon electronic bulletin board, were paid for their participation.

### Results and Discussion

The data from 1 subject were excluded because she did not answer the posttest questions. Approximately 3% of the trials were discarded due to inaccurate timing (e.g., inadvertent vocalizations or other noises which stopped the clock prematurely, or answers spoken too quietly to register). The subjects answered approximately 80% of the posttest questions correctly, confirming our intuition that the questions queried common knowledge.

The accuracy rate and RT data were analyzed for the questions that were answered correctly on the posttest.[5] ANOVAs were performed on both the accuracy data and the correct RT data, using as factors: instruction type (mention of capitalization vs. no mention), capitalization (critical term in capitals vs. nothing capitalized), and distortion (distorted vs. undistorted questions). The type of instruction had no main effect on the accuracy rates and did not interact with any of the other factors, so it will not be mentioned further. Table 4 presents the accuracy and RT data, as well as the proportion of types of errors, as a function of capitalization and distortion, for the questions answered correctly on the posttest. Again, no factor significantly affected response times (all $F$s < 2).

The subjects answered much more accurately when the question was not distorted [over subjects, $F(1,34) = 22.02, p < .001$; over questions, $F(1,63) = 14.73, p < .001$] and when the target terms were capitalized [over subjects, $F(1,34) = 9.00, p < .01$; over questions, $F(1,63) = 4.62, p < .05$]. However, the capitalization improved performance only for the distorted questions, yielding a significant interaction [over subjects, $F(1,34) = 10.13$, $p < .01$; over questions, $F(1,63) = 5.12$, $p < .05$]. This was in part because subjects continued to give more "can't say" responses to undistorted questions when a capitalized word was present (6% vs. 4%), yielding more errors with capitalization for undistorted questions. This tendency was less pronounced here than in Experiment 3A, but we still calculated sensitivity ($A'$) and bias ($B''d$) measures. Sensitivity was greater when the critical term was capitalized [$A' = 0.864$ with capitalization vs. $A' = 0.831$ without; $F(1,35) = 6.32$, $p < .05$]. Bias was also affected by the presence of capitalization: the subjects were significantly more likely to call a question distorted when the critical term was capitalized ($B''d = 0.624$) than when there was no capitalization [$B''d = 0.741$; $F(1,35) = 5.62, p < .05$].[6]

Experiment 3B replicated the results of Experiment 3A, indicating that our previous results were not simply an artifact produced by subjects' ignorance of the relevant

**Table 4**
**Mean Proportion Correct, Proportion of Error Types, and Mean Latency (in Seconds, in Parentheses) for Questions in Experiment 3B as a Function of Capitalization in the Question and Whether the Question Was Distorted, for the Questions Answered Correctly on the Posttest**

| | | Types of Errors* | | | |
|---|---|---|---|---|---|
| Capitalization | Correct | Undistorted | CS | DK | Incorrect |
| Distorted Questions | | | | | |
| Critical term capitalized | .66 (4.98) | .27 (5.14) | – | .06 (5.95) | .01 (7.04) |
| Nothing capitalized | .53 (5.06) | .32 (5.03) | – | .12 (6.14) | .03 (5.96) |
| Undistorted Questions | | | | | |
| Critical term capitalized | .77 (5.11) | – | .06 (6.69) | .14 (6.33) | .03 (5.72) |
| Nothing capitalized | .77 (4.82) | – | .04 (7.04) | .15 (5.86) | .04 (5.95) |

*The types of errors are answering distorted questions as if they were undistorted (Undistorted), "can't say" responses to undistorted questions (CS), "don't know" (DK), and incorrect responses.

facts. In both Experiment 3A and Experiment 3B, capitalizing the target term in the question provided a focus of attention that enhanced the detection of distortions by increasing sensitivity to the distorted questions, but this improvement also resulted, at least in part, from an increase in the tendency to guess that a question was distorted when capitalization was present.

Experiments 1 and 2 and previous research have indicated how difficult it is to detect distortions in questions. Our sensitivity analyses suggest that even when the ability to detect distortions appears to improve, the results are really due to a shift in response bias rather than to an increase in sensitivity. In Experiments 3A and 3B, where the critical terms were made much more salient, we were finally able to elicit an increase in sensitivity, but this effect was still accompanied by a shift in response bias. Furthermore, even though subjects were more sensitive to distortions when the critical terms were capitalized, the improvement was small compared with the size of the effect. Averaging over Experiments 3A and 3B, subjects still failed to detect the distortions 25% of the time when the distorted term was capitalized; the illusion rate was 34% when it was not capitalized.

Capitalization of the target term, either in a prestudied sentence or in the question itself, produced significant increases in the detection of distortions, but they were relatively small and due partially to changes in response bias. This led us to reconsider our conceptualization of how partial matching operated. Previously, we had assumed that with some probability, subjects would focus attention on the critical (i.e., distorted) word, and thereby notice the mismatch with the corresponding concept in memory. Experiments 3A and 3B provide some support for this concept, as capitalization of the critical term increased sensitivity to distortions. However, because this manipulation also increased bias, we have come to the conclusion that there must be another component to the partial-matching process. Because subjects so frequently fail to notice that the word does not match the memory structure even when highlighted, it seems that the matching must be going on at a lower level (i.e., at a feature level). Therefore, it seemed to us that the critical manipulation should not make a word more or less salient, but rather alter the salience of the semantic features of the word. We suspected that the distortion would be detected more readily if the semantic features that distinguished the distorted term from the undistorted term were emphasized. Consistent with this conjecture, van Oostendorp and de Mul (1990) found that similarity ratings of the critical and the distorted term (the number of shared attributes and the strength of relations between concepts) was predictive of distortion detection. Experiment 4 was designed to test this idea by manipulating the salience of particular features of the critical term.

## EXPERIMENT 4

In this experiment, features of the distorted term which differentiate it from the undistorted term were made salient in a question preceding the question of interest. Specifically, each critical question was immediately preceded by a question that (1) emphasized features shared by the critical and distorted terms, (2) emphasized features distinguishing the distorted term from the term it replaced, or (3) was irrelevant to the critical question. We hypothesized that the distortion in the question should be more obvious when the preceding question made the distinguishing features of the distorted term more salient.

### Method

**Materials**. Fifty-four pairs of questions were used, including 49 of the original 72 pairs. For each pair of questions, we created questions that were intended to make salient those features that were either common to the critical term and the distorted term or that distinguished them. For example, the question "What religions study the story of Moses?" focused on features common to both Noah and Moses. The criterion for acceptability of questions intended to emphasize the similarity between the distorted and original term was that either term could be used in the preliminary question and would produce the same answer. When this type of question preceded the distorted question, it contained the distorted term; before the undistorted question, it contained the undistorted term. The second type of preliminary question emphasized the differences between the original term and the distorted term. Two versions of this type of question were written, one that preceded the distorted question and queried information unique to the distorted term (e.g., "What sea did Moses part?") and one that preceded the undistorted question (e.g., "How many sons did Noah have?"). The irrelevant preceding question was merely another test question.

A posttest similar to that used in Experiment 2 was also used. This test was designed to check whether the subjects possessed the knowledge needed to detect the distortions in a question.

**Design and Procedure**. The design was similar to that of previous experiments. For each subject, questions were assigned randomly to the distortion condition. Half of the test questions were asked in their distorted form. One third of the distorted and one third of the undistorted questions were preceded by a question emphasizing similarities, one third were preceded by a distinguishing question, and one third by an irrelevant question (which was simply another, unrelated test question).

The procedure was the same as for the previous experiments. Both critical and preceding questions were presented by computer. The critical questions were immediately preceded by the type of question appropriate to their condition. All questions were presented in the same fashion, with nothing to distinguish among the types of questions. The subjects indicated, during debriefing, that they had been unaware that there were different types of questions. As before, the subjects initiated presentation of a question by pressing a button on a button box attached to the computer, then attempted to answer the question literally by saying "can't say" to distorted questions and giving the answer to undistorted questions. The experimenter remained in the room to record answers and to note trials that were invalidated by improper activation of the voice key. After all 90 context and test questions presented by computer had been answered in this fashion, the posttest questions were presented by computer and the answers typed by the subject.

**Subjects**. Forty-three Carnegie Mellon undergraduates participated in this study for partial fulfillment of a course requirement or for pay.

### Results and Discussion

Less than 3% of the trials were discarded due to inaccurate timing. The subjects answered over 80% of the questions on the posttest correctly. Only the data from trials that had correct posttest answers were included in the

following analyses. ANOVAs were performed on both the valid RT data and the accuracy data using as factors distortion (distorted vs. undistorted questions) and type of preceding question (irrelevant vs. distinctive vs. similar); illusion rates were also subjected to an ANOVA. Table 5 presents the mean accuracy of responses, RTs, and types of errors as a function of distortion and context type.

As before, the presence of distortions significantly decreased accuracy [over subjects, $F(1,43) = 130.43$, $p < .001$; over questions, $F(1,53) = 97.88$, $p < .001$] and slowed RTs [$F(1,43) = 14.12$, $p < .01$]. No other analyses of RTs were significant, so those data will not be discussed further. Importantly, test questions preceded by questions that emphasized a distinguishing feature of the distorted term were answered more accurately than those with an irrelevant preceding question or one that emphasized similar features [over subjects, $F(2,86) = 6.31$, $p < .01$; over questions, $F(2,106) = 7.10$, $p < .01$].

The interaction of distortion and type of preceding question was not significant ($F$s < 2). Nonetheless, the type of preceding question had a much larger effect on the distorted questions than on the undistorted questions. Bonferroni's method, with $\alpha = .05$, was used to compare performance on the distorted questions for each of the types of preceding question. The detection rate was significantly better when the question preceding the critical question distinguished features than when it emphasized similarity or was irrelevant. Detection rates did not differ between the latter two conditions. The illusion rate showed the same pattern as the accuracy data, with the illusion rate for distorted questions preceded by a distinctive question (30%) significantly lower than that for the other two conditions [41%–42%; over subjects, $F(2,86) = 3.74$, $p < .05$; over items, $F(2,106) = 4.73$, $p < .05$].

Our signal-detection analysis also indicates that the distinctive question condition facilitated detection of distortions. Calculation of sensitivity and bias measures showed that subjects became more sensitive to distortions when the critical question was preceded by a distinguishing question ($A' = 0.836$) than when those questions were preceded by similar questions ($A' = 0.795$) or irrelevant questions [$A' = 0.792$; $F(2,86) = 3.71$, $p < .05$]. In contrast, bias was not affected by this manipulation (distinctive preceding question, $B''d = 0.675$; similar preceding question, $B''d = 0.736$; irrelevant question, $B''d = 0.779$; $F < 2$).[7] Presumably, subjects did not alter their criterion for calling a question distorted on the basis of condition as they did in Experiment 3, because the manipulation here was much more subtle, precluding a conscious shift of strategy based on the preceding question.

The pair of results indicating that (1) detection rates improve when a preceding question emphasizes features that distinguish the original term and its replacement, and (2) detection rates are not impaired by an emphasis on the features of the two terms that are shared is particularly interesting in light of Barton and Sanford's (1993) results. Barton and Sanford investigated anomaly detection using the question, "When an airplane crashes, where should the survivors be buried?" Providing additional information to the question hurt performance when that information was relevant to the answer, but not when the information was irrelevant. In their case, the additional, relevant information was relevant to the answer, not to the term that was distorted. This result might seem analogous to our result in Experiment 1, in which the answer term had been capitalized during study. Barton and Sanford argue that if the context of the statement provides enough information to answer the question, the individual words will not be examined carefully, and thus the distortion, or anomaly, would not be detected. However, the present experiment finds that whether an anomalous question is detected as such can be affected by prior questions that have no bearing on the answer to the critical question.

## GENERAL DISCUSSION

We began by reviewing previous Moses illusion experiments that both illustrated the robustness of the illusion and ruled out a number of possible explanations for the effect. The current experiments explored aspects of the remaining hypothesis, the partial-match hypothesis.

**Table 5**
**Mean Proportion Correct, Proportion of Error Types, and Mean Latency (in Seconds, in Parentheses) for Questions in Experiment 4 as a Function of Type of Preceding Question and Whether the Question Was Distorted**

| Type of Preceding Question | Correct | Types of Errors* | | | |
|---|---|---|---|---|---|
| | | Undistorted | CS | DK | Incorrect |
| | | Distorted Questions | | | |
| Distinctive | .59 (4.50) | .30 (4.15) | – | .08 (6.13) | .02 (4.14) |
| Similar | .48 (4.41) | .41 (4.18) | – | .09 (4.61) | .02 (4.07) |
| Irrelevant | .47 (4.59) | .42 (3.85) | – | .11 (6.91) | .01 (3.71) |
| | | Undistorted Questions | | | |
| Distinctive | .86 (3.84) | – | .03 (5.66) | .08 (6.48) | .03 (5.91) |
| Similar | .83 (4.03) | – | .04 (5.72) | .10 (4.12) | .03 (5.30) |
| Irrelevant | .83 (4.09) | – | .04 (6.06) | .13 (5.22) | .01 (3.94) |

*The types of errors are answering distorted questions as if they were undistorted (Undistorted), "can't say" responses to undistorted questions (CS), "don't know" (DK), and incorrect responses.

In four experiments, we tried to improve subjects' abilities to detect the questions that contained an illusion. In all four experiments, our manipulations produced significant increases in number of distortions reported, yet the distortion-detection rates were still far from perfect; subjects still failed to notice the distortion at least one time in four, even when the distorted word was set in CAPITAL letters. Figure 1 illustrates the illusion rate for each experiment based only on questions common to all of the experiments. Although the differences between the experimental and control conditions were significant in every case, the Moses illusion was clearly much stronger than our manipulations.

The results of these experiments help to provide answers to some of the questions posed in the introduction. Failure to detect distortions does not seem to result from insufficient allocation of attention to matching the critical term in the question to the memory representation. In Experiments 1 and 3, the critical term was capitalized in the question or in the corresponding study sentence, resulting in significantly more distortion responses; however, it also increased the tendency to label undistorted questions as distorted. Likewise, in Experiment 2, when subjects were required *only* to monitor for distortions and not to answer the questions, the absolute accuracy improved for distorted questions; there too, however, a concomitant increase in error rates for undistorted questions was found. Thus, failure to detect distortions does not seem to occur because another process that attempts to answer the question is completed before the distortion is noticed.

An important outcome from this research was the failure to replicate prior results showing significant improvements in distortion detection. The nature of our controls allowed us to determine that the improvement in absolute performance in Experiments 1 and 2 resulted not from improved sensitivity to distortions, but only from a shift in response bias. Subjects in Experiment 3 did show more sensitivity to the distortion with capitalization, but this also increased subjects' bias (i.e., capitalization also increased the tendency to respond "distorted" for undistorted questions). Only the manipulation in Experiment 4 affected subjects' sensitivity to distortions without also increasing their bias.[8]

So what do these experiments tell us about question answering and partial matching? First, partial matching is much more robust than originally thought. It seems that people cannot easily become more vigilant at detecting distortions, even when they try. Previous research as well as these studies can be interpreted as indicating that people are able to develop strategies to *try* to be on the lookout for tricky questions, but, in fact, they cannot easily change the basic nature of the partial-match process. The fact that our manipulations had a larger impact on bias than on ability to actually detect distortions suggests that other investigations of this phenomenon should also take care to distinguish shifts in bias from shifts in detection rate.

The results also indicated that subjects are effectively unable to adopt an explicit word-by-word checking procedure. Furthermore, the results of Experiment 4, in which sensitivity rather than bias was affected, suggest that the partial-match process operates at the feature level. Answering questions that required the answerer to attend to features of the distorted term that distinguished it from the original term were most effective in improving subsequent distortion detection.

How might this cognitive machinery be implemented? We have proposed elsewhere (Kamas & Reder, 1995) a semantic network of connected concepts in which activation spreads among concepts that are semantically related. When a person is asked a question, processes (e.g., productions) operate on the net to find the queried element. The speed with which the queried element can be

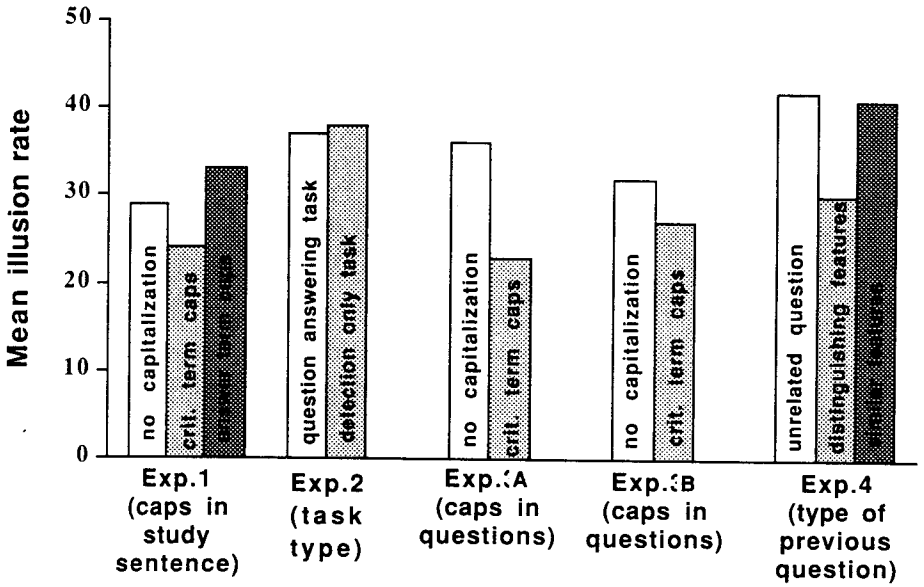


**Figure 1. Mean illusion rates from all experiments, including only those questions common to all experiments.**

located and given as an answer depends on the activation level of the proposition that contains the queried element. The more activation that accrues at a concept through its connections to the remainder of the concepts in the question, the more likely the person is to accept the retrieved concept as matching the question.

When a term in the question does not match the stored representation, the probability of detecting this mismatch is a function of the number and strength of connections from the distorted word to the schematic node that is queried. The more connections between the schema and the distorted term and the stronger those connections, the more likely it is that the distorted term will go undetected. For example, Noah and Moses are both biblical characters, old men with beards, and associated with biblical water stories. The large number of connections means that the substitution of Moses for Noah will go unnoticed. In contrast, Nixon has no obvious semantic connections with the Noah schema, so the mismatch will be detected. An additional necessary assumption is that there is a limit on total activation that is divided among all the concepts in the probe such that a mismatching word with no connections, such as Nixon, takes away activation that could be spread to the relevant script (see Anderson, Reder, & Lebiere, 1996) for further discussion of this assumption.

To relate this model to the present experiments, the first three experiments entailed manipulations at the word level, and we found little effect on sensitivity to distortions. The final experiment manipulated the salience of critical features of concepts, rather than entire words, and we found a change in sensitivity. According to our current ideas about partial matching, this should be the case. We should not expect manipulations at the word level to consistently affect the rate of distortion detection, because manipulations at this level should raise the salience of both the similar *and* the distinguishing features of the correct and distorted terms.

Assume that, prior to priming, the term Moses receives one-sixth of the total activation; however, after priming, Moses receives one-fourth of the total activation that can spread to the relevant script. In either case, that activation is divided among the component features of Moses that may or may not spread to the Ark script. Since priming in the first three experiments does not alter the relative distribution of activation from the word node to its constituent features, the proportion of activation sent from the similar features has not changed. The only hope for an increase in rejection is if the proportion of related features is sufficiently smaller than other terms in the question and, since that word is getting a greater share of the total activation, the probability of passing over the activation threshold at the script node is reduced. For related terms, such as Moses, this is not a very likely event (cf. Reder & Schunn, in press; Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, in press). Another way to think about it is that if the relative salience of the distinguishing features is the key to rejecting distortions, when we raise the activation of all of the features, the net effect on sensitivity should be zero because the change in *relative* salience is zero.

Other data from our lab are consistent with this conception of how question answering and detection operate. We examined the posttest performance from two Moses experiments that included posttests. We were interested in determining whether knowledge of the correct information was affected by exposure to particular conditions. Given that questions were assigned randomly to conditions for each subject, any difference in posttest accuracy (e.g., Who took the animals on the Ark?) must be due to the experimental treatment. We discovered a very robust finding in the posttest performance, showing that subjects are more likely to know that Noah took the animals on the Ark when they received the undistorted question and less likely to know that when they were presented with the distorted version, strongly suggesting that world knowledge is affected by exposure to one question.[9] Differences in posttest performance in both experiments were significant by sign tests ($N = 22, p < .001$, and $N = 21, p < .001$, respectively).

Not only is posttest performance lower when the corresponding question was distorted, but the kinds of errors made are predicted by an activation-based model of memory. Specifically, subjects are much more likely to answer with the distorted term on the posttest if they had previously seen the distorted term in the corresponding question. For example, if the posttest question was "Who took the animals on the Ark?" subjects were much more likely to respond "Moses" if they had originally seen "How many animals of each kind did Moses take on the ark?" than if they had previously seen the same question with "Noah." Furthermore, as we explain below, the size of this effect is greater for questions that are more susceptible to the illusion than for those that do not find the distorted question "tricky."

One can ask whether this difference in posttet performance is due to inherent lack of knowledge or to interference from the manipulation. Another unpublished experiment from our lab directly examined posttest performance as a function of three types of questions: two types were the same as those used in Experiments 1–4 reported here; the third type was one that used versions of questions that did not mention either the distorted or the undistorted term (e.g., "How many animals of each kind were taken on the Ark?"). The results make clear that posttest-performance differences were not due to lack of knowledge, but rather to interference. Subjects gave the distorted term as the posttest answer more often when the original question had been seen in distorted form (13.8%) than when it had been seen in the undistorted (2.7%) or omitted (3.3%) form [$t(25) = 4.38, p < .001$, and $t(25) = 4.15, p < .001$, for undistorted and omitted forms, respectively]. It appears that this finding is due to the presence of the distorted term rather than the undistorted term, because we found virtually no difference in posttest performance between the undistorted and omitted forms [$t(25) = 0.53, p = .60$]. The differences were even greater among the questions that were more effective at induc-

ing the illusion. For posttest questions that were above the 50th percentile in terms of illusion rate, the distorted term was given on the posttest at rates of 21.5%, 5.0%, and 4.6%, respectively, for distorted, undistorted, and omitted forms of the original question.

Given that we found virtually no difference in posttest performance between the undistorted and omitted forms, it appears that the change in world knowledge was not due to subjects' not having the relevant information ahead of time, but was due to the presence of the distorted term's affecting their knowledge structure. Thus, the mere presence of the distorted term appears to affect schematic representations at least several minutes beyond the time the term is encountered. We suggest that this is because the features of the distorted term are already connected to the schema and that these links are strengthened by the presentation (cf. Reder & Schunn, in press; Schunn et al., in press).

These findings are related to those of Potter and Lombardi (1990) and Kelley and Lindsay (1993). Potter and Lombardi found that verbatim recall was distorted when a semantically similar word had been activated between the encoding of the sentence and the recall test. Subjects would inadvertently substitute the synonym into sentence recall because it fit semantically and was now more active than the original word. Kelley and Lindsay found that subjects were more prone to give the wrong answer to a question when a semantically similar one had been read earlier (e.g., "Hickcock" instead of "Cody" for "What was the last name of Buffalo Bill?").

In the Moses illusion paradigm, the distorted term shares semantic features with the undistorted term. Given that a subject fell for the illusion, we can safely assume that the distorted term was consistent with that subject's conceptual representation, because otherwise the distortion would have been detected. Because the distorted term was recently activated within the conceptual representation of the question (e.g., "Moses" within the representation of the Ark story) while the undistorted term was not, the distorted term was given as the answer on the posttest. We predict that as time passes between the presentation of the original question and the presentation of the corresponding posttest question, the probability of giving the distorted term as a response on the posttest decreases. This prediction follows from the forgetting functions that depend on the number of presentations made during acquisition.

Given the apparent difficulty people have in detecting distortions or inaccuracies in questions, it may seem that partial matching is a less-than-ideal way to process information. Why should the partial-match process be so robust, so difficult to override with a more exact/verbatim matching process? As we suggested earlier, this partial-matching (rather than exact matching) process is not only common and normal, but necessary, given the requirements of everyday information processing; queried facts often do not exactly match the stored information. Consider what would occur if we did try to exactly match to memory: Luria (1965/1968) describes the difficulties of S., the famous mnemonist who made a career out of remembering *exactly* what was presented to him. S. found it difficult to recognize voices on the telephone and faces, because "they're so changeable.... A person's expression depends on his mood and on the circumstances under which you happen to meet him. People's faces are constantly changing" (p. 64). S.'s experience suggests that exact matching is not an ideal way for the memory system to operate. We offer that partial matching is immutable because it is the most efficient way for memory to operate given the nature of the environment in which we live.

## REFERENCES


ANDERSON, J. R., REDER, L. M., & LEBIERE, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, **30**, 221-256.

BAKER, L., & WAGNER, J. L. (1987). Evaluating information for truthfulness: The effects of logical subordination. *Memory & Cognition*, **15**, 247-255.

BARTON, S. B., & SANFORD, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, **21**, 477-487.

BREDART, S., & DOCQUIER, M. (1989). The Moses illusion: A follow-up on the focalization effect. *Cahiers de Psychologie Cognitive*, **9**, 357-362.

BREDART, S., & MODOLO, K. (1988). Moses strikes again: Focalization effect on a semantic illusion. *Acta Psychologica*, **67**, 135-144.

DONALDSON, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General*, **121**, 275-277.

ERICKSON, T. D., & MATTSON, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning & Verbal Behavior*, **20**, 540-551.

HORNBY, P. A. (1974). Surface structure and presupposition. *Journal of Verbal Learning & Verbal Behavior*, **13**, 530-538.

KAMAS, E. N., & REDER, L. M. (1995). The role of familiarity in cognitive processing. In E. O'Brien & R. Lorch (Eds.), *Sources of coherence in readings* (pp. 177-202). Hillsdale, NJ: Erlbaum.

KELLEY, C. M., & LINDSAY, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory & Language*, **32**, 1-24.

LURIA, A. R. (1968). *The mind of a mnemonist* (L. Solotaroff, Trans.). New York: Avon Books. (Original work published 1965)

POTTER, M. C., & LOMBARDI, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory & Language*, **29**, 633-654.

REDER, L. M., & CLEEREMANS, A. (1990). The role of partial matches in comprehension: The Moses illusion revisited. In A. Graesser & G. H. Bower (Eds.), *The psychology of learning and motivation* (Vol. 25, pp. 233-258). New York: Academic Press.

REDER, L. M., & KUSBIT, G. W. (1991). Locus of the Moses illusion: Imperfect encoding, retrieval, or match? *Journal of Memory & Language*, **30**, 385-406.

REDER, L. M., & SCHUNN, C. D. (in press). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition*. Hillsdale, NJ: Erlbaum.

SCHUNN, C. D., REDER, L. M., NHOUYVANISVONG, A., RICHARDS, D. R., & STROFFOLINO, P. J. (in press). To calculate or not to calculate: A source activation confusion (SAC) model of problem-familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

SNODGRASS, J. G., & CORWIN, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.

VAN OOSTENDORP, H., & DE MUL, S. (1990). Moses beats Adam: A semantic relatedness effect on a semantic illusion. *Acta Psychologica*, **74**, 35-46.

## NOTES

1. The standard measures of discrimination and bias, $d'$ and $\beta$, yielded similar results. The mean $d'$ scores in the three conditions were not significantly different (critical term capitalized, $d' = 3.11$; answer capitalized, $d' = 2.88$; no capitalization, $d' = 3.09$; $F < 1$). Subjects appeared biased against calling a question distorted; they were less likely to judge a question as being distorted when the study sentence had the answer term capitalized ($\beta = 13.4$) than when the target term ($\beta = 10.2$) or nothing ($\beta = 11.7$) had been capitalized.

2. The standard measures of sensitivity and bias were similar: mean sensitivity to distortions did not differ between the two tasks ($d' = 1.70$ for question answering; $d' = 1.62$ for the single task; $F < 1$), while response bias differed significantly between the two tasks [$\beta = 5.3$ in the single-task condition vs. $\beta = 25.9$ in the standard task; $F(1,50) = 7.14$, $p < .05$].

3. The standard measures of sensitivity and bias gave similar results: sensitivity was no different when the critical term was capitalized ($d' = 2.23$) than when it was not ($d' = 2.34$; $F < 1$), while bias to call a question distorted was greater when the critical term had been capitalized ($\beta = 32.6$) than when no term was capitalized [$\beta = 51.8$; $F(1,39) = 4.45$, $p < .05$].

4. The reason why Experiment 3A did not have a posttest while Experiment 2 did was because Experiments 3A and 3B were conducted before Experiment 2. For reasons of exposition, we did not present them in chronological order.

5. In this and the following experiment, the results were also analyzed without regard to posttest performance. Both analyses yielded comparable results and conclusions.

6. Looking at the standard sensitivity and bias measures gives similar results. Sensitivity did not differ between the two conditions ($d' = 2.54$ with capitalization vs. $d' = 2.43$ without capitalization; $F < 1$), while there was a significant effect on bias; subjects were significantly more likely to call a question distorted when the critical term was capitalized ($\beta = 32.6$) than when there was no capitalization [$\beta = 46.4$; $F(1,35) = 3.57$, $p < .07$].

7. The standard sensitivity and bias measures showed similar results. When the preceding question emphasized distinguishing features, $d' = 2.95$, while for the similar and irrelevant preceding questions, $d' = 2.44$ and 2.46, respectively [$F(2,86) = 2.92$, $p < .06$]. Bias to report a question as distorted did not differ among the three conditions (distinctive question, $\beta = 48.59$; similar question, $\beta = 53.19$; irrelevant question, $\beta = 56.54$; $F < 1$).

8. Power analyses indicated that, given the size of the differences obtained, in order to find significant effects of discriminability in the experiments for which we found nonsignificant differences, we would have had to run about 500 subjects. Thus, it may have been possible to get significant results in Experiments 1–3, but they would not have been very meaningful or useful for theory.

9. It is important to note that our analyses excluded any trial where the subject did not know the answer. Since subjects who received distorted questions were much less likely to know that it was Noah (to give the correct answer on the posttest) than those who received the undistorted questions, our reported results in previous research are actually underestimates of the true effect.

(Manuscript received January 17, 1995;
revision accepted for publication November 13, 1995.)