JOURNAL OF MEMORY AND LANGUAGE 32, 320–335 (1993)

The Feeling of Another Person's Knowing

ANTHONY JAMESON

University of Nijmegen, The Netherlands

THOMAS O. NELSON AND R. JACOB LEONESIO

University of Washington

AND

LOUIS NARENS

University of California at Irvine

We investigated whether predictions about the currently unrecallable knowledge that a person possesses (a) are facilitated by that person's privileged access to nonobservable information and (b) are influenced by aspects of his or her behavior that can also be observed by another person. A total of 106 *target* subjects (a) attempted to answer general-information questions (b) predicted which of the unanswered items they would be most likely to recognize, and (c) took a multiple-choice test on the items. The recognition performance of each of these target subjects was also predicted by an *observer*, who had watched the target's recall attempts, and by a *judge*, who had virtually no information about the target. The targets predicted more accurately than the observers, who were in turn more accurate than the judges. The predictions of the targets and the observers were related to three cues in the targets' behavior: (a) type of recall failure (omission error vs commission error), (b) latency of omission errors, and (c) plausibility of commission errors.

THE FEELING OF ANOTHER PERSON'S KNOWING

How do we make predictions about what another person knows? What are the bases for such predictions, and how accurate are they relative to the predictions made by the person himself/herself?

These questions are important for several intellectual domains:

This research was partially supported by NIMH Grant MH32205 to the second author. The participation of the first author was supported by the foundation Psychon, which is financed by the Netherlands Organization for Pure Research (N.W.O.). We are indebted to A. Greenwald, E. Hoenkamp, R. W. Meertens, C. Wegman, and the reviewers for comments on earlier versions of the manuscript. Address correspondence and reprint requests to Dr. Thomas O. Nelson, Department of Psychology (NI-25), University of Washington, Seattle, WA 98195.

1. For psychology in general because of its concern—at least since the time of the radical behaviorists—with the distinction between observable and nonobservable information.

2. For cognitive psychology because of its subarea of metamemory, in which the basis and accuracy of the monitoring of what one knows is a central topic (Nelson, 1992).

3. For the psychology of language, because linguistic communication requires continual judgments about which words and facts are known to the participants in the communication.

4. For social psychology because of its interest in what one person can know about another person and also in self-perception theory (e.g., Bem, 1972).

5. For the philosophy of mind because of

0749-596X/93 $5.00

its central concern with "self-consciousness" and the "consciousness of other other minds" (Nelson, 1992, Chap. 2). Concerning the latter, answers to the above questions may yield nonmystical mechanisms by which one person can predict what another person knows, thereby helping to demystify the notion of "mind reading" (cf. Nelson, Leonesio, Landwehr, & Narens, 1986).

The Vesonder/Voss Paradigm

Vesonder and Voss (1985, Experiment 2) introduced an experimental paradigm within which a number of general issues concerning knowledge prediction can be investigated. The paradigm involves three subject roles, which we will refer to as those of *target, observer,* and *judge.*[1] In Vesonder and Voss' experiment, each target attempted to learn 48 sentences, predicting for each sentence in each of four study-test trials whether he would be able to recall the sentence when cued with its first two words. The target's future performance was also predicted by an observer, who listened to the target's recall attempts, and by a judge, who was not aware of any of the target's responses.

One of the issues investigated by Vesonder and Voss concerned the ways in which the target and the observer used information about the target's previous performance when predicting his future performance. They found that once the target had recalled an item correctly, the target and the observer almost always remembered this fact and accordingly gave optimistic predictions. This result corroborated previous research which had indicated that predictions of this sort are based to an important extent on memory for previous recall performance (see, e.g., King, Zechmeister, & Shaughnessy, 1980; Lovelace, 1984).

[1] Vesonder and Voss (1985) used different labels for these roles. For clarity in exposition, we use masculine pronouns to refer generically to target subjects and feminine pronouns for observers and judges.

A second issue concerned predictions about items that a target had *failed* to recall on the previous trial. Even the judge predictors showed significantly above-chance accuracy on such items; and although the observers and especially the targets were more accurate than the judges, the differences were not statistically significant. Vesonder and Voss accordingly concluded that a prediction of this sort is based largely on an analysis of characteristics of the item to be learned (as opposed to specific information about the knowledge of the person being predicted). They noted, however, that this latter type of information might play a greater role in an experiment involving items about which subjects differed more with respect to their prior knowledge, such as the general-information items that have often been used in experiments within the *feeling-of-knowing* paradigm (Wilkinson & Nelson, 1984).

The present study takes up this last suggestion. Our two experiments involve the attempted retrieval of information that could only have been acquired prior to the experiment—as opposed to during it, as in Vesonder and Voss' (1985) Experiment 2. The basic design of Vesonder and Voss' experiment is adapted to the feeling-of-knowing paradigm: The target (a) attempts to answer a number of general-information questions, (b) predicts which of the nonrecalled items he is most likely to recognize, and (c) attempts to recognize the answers in a multiple-choice test. (See Leonesio & Nelson, 1990, for a comparison of feeling-of-knowing judgments with judgments of learning—the type of judgment involved in Vesonder and Voss' study.) The observer, but not the judge, is able to watch the target during the first phase; both the observer and the judge predict the target's later recognition performance.

The Target's Observable Behavior

Within this design, there can be no role for the one cue in the target's observable behavior that Vesonder and Voss examined specifically, i.e., recall success on the pre-

vious trial: All of the items about which predictions were made in our experiments were items that the target had failed to answer correctly. But the target's behavior during an unsuccessful recall attempt might contain more subtle cues that could influence the predictions of the target and/or the observer, for example: (a) whether the target made an *omission error*, producing no response, or a *commission error*, giving an incorrect response; (b) the latency of the target's recall attempt; and (c) for commission errors, the plausibility of the incorrect response given. In an exploratory fashion, we check for relationships between these variables and the predictions of subjects in the three roles.

Although it is difficult to predict in advance which specific cues might be of use to the observers, our first hypothesis is that observers can benefit from at least some of this information and can thereby outpredict the judges. As already noted, in Vesonder and Voss' experiment the observer benefited from her awareness of which items the target had recalled successfully, and even on nonrecalled items she tended (though not reliably) to show greater accuracy than the judge.

The Target's Internal Responses

Another difference that was not significant in Vesonder and Voss' experiment was the predictive superiority of the targets to the observers on items not previously recalled. Our second hypothesis is that our targets can predict more accurately than can our observers. Most of the bases for feeling-of-knowing judgments that have been proposed (see Nelson, Gerler, & Narens, 1984, pp. 295–299, for an overview) involve what may be called *internal responses* to items. This term refers to any unobservable response that a target might use as evidence specifically about his own knowledge of an item—for example, a feeling of familiarity, partial recall of the answer, or retrieval of autobiographical facts concerning previous encounters with the item.

Actuarial Information

The results of Vesonder and Voss also suggest a third hypothesis about the present experiments: Even the judges can attain above-chance accuracy in predicting the targets. This prediction finds further support in experiments by Nickerson, Baddeley, and Freeman (1987) and by Fussell and Krauss (1991): Their subjects showed above-chance accuracy in estimating the difficulty of knowledge items for a population of students. (The items used by Nickerson *et al.* were from the pool of general-information items used in the present experiments.)

As noted above, Vesonder and Voss explained the above-chance accuracy of their judge subjects by postulating that predictions of memory performance following recall failure largely involve the assessment of item characteristics. For general-information items, the potential predictive value of information about objective item properties has been demonstrated by a result of Nelson *et al.* (1986): they showed that when one predicts a target's recognition performance on nonrecalled items using only a type of *actuarial information*—namely, objective statistics on the items' normative recall difficulty—the accuracy of these predictions is greater than that usually shown by a target's own feeling-of-knowing judgments. Moreover, Calogero and Nelson (in press) showed that giving actuarial information to a person who is making feeling-of-knowing judgments will improve the predictive accuracy of those judgments. Although people do not usually possess actuarial information that is as precise and accurate as recall norms, they do presumably have a less reliable sort of actuarial information, concerning, for example, how familiar American students are with particular topics.

The Judge's and Observer's Own Responses

The experiments by Nickerson *et al.* (1987) and by Fussell and Krauss (1991)

also suggest that the accuracy of our judge subjects may not be due entirely to their use of actuarial information, but also to their use of their own internal responses to the items. In both of those studies, the subjects' predictions were strongly related to the correctness of their own answers to the items and to the confidence that they expressed in their answers. This is understandable: even when trying to use actuarial information, predictors may need to rely upon information about their own responses. Consider, for example, an item that asks for the composer of a little-known opera. Actuarial information may tell the predictors that most American students are not very interested in opera; but the predictors may have little information about how well-known the specific opera is, aside from the feeling of familiarity or unfamiliarity that its title invokes in them. An item assessment based on such a feeling can differ considerably among predictors, depending on the encounters that they happen to have had with the opera in question. The result is that assessments of item difficulty tend to be systematically related to the predictor's own knowledge.

Analogous relationships between people's own beliefs and their predictions for others have repeatedly been found with attitudinal and behavioral items (for discussions of theoretical explanations, see Dawes, 1989; Hoch, 1987; Marks & Miller, 1987). Given the strength of these relationships in previous research, they may be expected not only with our judge subjects but also with our observers. To test this fourth and final hypothesis, we have each judge and observer take the same recognition test as the target on the target's nonrecalled items.

Method for Both Experiments

Experiment 1

Design

The central aspects of the design have been described under the Introduction. Several additional aspects were included to minimize unwanted differences between conditions and to limit the number of subjects required. In addition to observing the target and making predictions for him, the observer performed the same three tasks as the target, starting with a test of her own general knowledge (on different items than those for the target). The role of the judge was not filled by a separate subject but rather by a subject in the target role in one of the later sessions. Therefore, only two subjects were present at each session. Table 1 gives an overview of their tasks, which are described in detail under Procedure.

Apparatus

Data collection, many of the statistical analyses, and the presentation of most instructions were performed by the FACT-RETRIEVAL program (Shimamura, Landwehr, & Nelson, 1981) on an Apple II microcomputer. The stimuli were displayed on a video monitor, and subjects entered their responses on the computer keyboard.

Items. Recall stimuli were taken from a pool of 240 of the 300 general-information questions from which Nelson and Narens (1980b) compiled norms. Recognition stimuli were the same questions, each paired with four randomly ordered response alternatives, of which exactly one was correct. The three distractors were chosen so as to represent plausible responses; e.g., for the item "What is the name of the brightest star in the sky excluding the Sun?" (correct answer: "Sirius") the distractors were "Rigel," "Polaris," and "Betelgeuse."

Procedure

The procedure can be divided into three discrete phases: the *recall phase*, the *prediction phase*, and the *recognition phase*, each of which is described separately.

Recall phase. Each subject received a test on a series of general-information questions. The questions were randomly chosen from the set of stimulus items with the con-

TABLE 1
TASKS OF SUBJECTS IN THE THREE PHASES OF THE EXPERIMENTS

Phase	Target (judge)[a]	Observer[b]
Recall	General-information test	Observation of target's general-information test [Own general-information test]
Prediction[c]	Ranking of own nonrecalled items (Ranking of previous target's nonrecalled items)	[Ranking of own nonrecalled items] Ranking of target's nonrecalled items
Recognition	Recognition test for own nonrecalled items (Recognition test for previous target's nonrecalled items)	Recognition test for own nonrecalled items Recognition test for target's nonrecalled items

[a] Each target also served as a judge by predicting the recognition performance of a previous target. The tasks in parentheses were performed in the role of the judge.

[b] Tasks in square brackets were included only to equalize conditions across roles and were not involved in the data analyses.

[c] The order of the two tasks in the prediction phase was counterbalanced.

straint that the items presented to a target subject were not presented to the observer or the judge of the same triad. Test trials were self-paced. Subjects were instructed to guess when unsure of the answer and to type the word *next* when unable to produce any response. To keep the pool of "incorrectly answered" items as free as possible of items of which the target had merely made a spelling error, a response was scored as correct if its first two letters matched the first two letters of the correct answer. (No correct answer began with the same first two letters as any of the plausible wrong answers used as distractors in the recognition test, or with the first two letters of the word *next*. The cases in which an incorrect answer was nonetheless scored as correct do not constitute a problem, because answers scored as correct were not involved in the later phases of the experiment, or in the data analyses.) Latency of response was measured from the time the stimulus was presented until the subject had completed entry of the response. (Because subjects sometimes changed their minds while entering a response, response initiation would have been a less appropriate point at which to terminate latency measurement.) Immediately after a response had been entered, the next item was presented. The recall phase continued until there had been 15 items to which the subject had failed to produce the correct answer.

Each target took the test in the presence of an observer subject, who was seated slightly behind and to the side of the target. The observer was orally instructed to pay attention to questions that she believed that the target failed to answer correctly and to consider for each such question whether the target might "subconsciously" know the answer. The target was unaware that this was the observer's task. The two subjects were instructed not to talk to each other during the target's test, and the experimenter remained in the room to enforce this constraint. When the target completed the test, the observer was taken to another room containing a second computer, into which the experimenter loaded a diskette containing a file of the items that had been presented to the target. While the target proceeded with the remaining two phases of the experiment (cf. Table 1), the observer was given a similar test in which a new subset of items was used that excluded those presented on the target's test. The experimenter remained in the room with

the observer during the observer's test to minimize any differential effect across roles of being watched.

Prediction phase. There were two segments to this phase: in one segment the subjects were asked to rank-order the 15 nonrecalled items from their own general-information test ("questions that you answered incorrectly earlier") on the basis of their own feeling of knowing, i.e., "how well you could recognize the correct answer to a question you answered incorrectly earlier in the study." These nonrecalled items included both omission errors and commission errors. (The consequences of this fact are discussed in connection with the results concerning the role of type of recall failure as a cue; see also Krinsky & Nelson, 1985.)

In the other segment of the prediction phase the subjects ranked the nonrecalled items from the general-information test of the target to whom they had been assigned (cf. Table 1). These rankings were to refer to the likelihood that the target would recognize the correct answer, not to the subject's own likelihood of recognizing it. One-half of the subjects in each role ranked the corresponding target's items first and the other half ranked their own items first. The observer was told that she was ranking the nonrecalled items of the student she had observed shortly before; the judge was told only that "another student," about whom she was given no further information, had failed to answer these items correctly.

Within each of the two segments of the prediction phase, the 15 nonrecalled items (of the subject or the corresponding target) were presented on the screen three at a time in such a way that each possible pair of two stimulus items appeared together in an item triple exactly once. This constraint yielded a total of 35 item triples. The 35 triples were presented in 7 subseries of 5 triples, with each item occurring exactly once in each subseries [see Burton & Nerlove, 1976, solution 4.a.(3)]. Within each subseries the item triples were randomly ordered, and within each item triple the items were randomly ordered. For each item triple, the subject was instructed to choose the item for which he or she had the strongest feeling of knowing (or the item which he or she believed the target would be most likely to recognize the correct answer). The item chosen then disappeared from the display, and the subject was asked to make the same choice for the two remaining items. The first item selected in a given triple was treated as having been chosen twice (i.e., over each of the other two items), and the second item selected was treated as having been chosen once. The number of choices received by each item over all presentations could therefore range from 0 to 14. It is the rank order of the items with respect to this variable that is used to express the subjects' predictions (Nelson & Narens, 1980a). Following presentation of the 35 item triples, 6 additional triples were presented to assess reliability. Exactly 3 of these were chosen from previously presented item triples.

This procedure for eliciting predictions has two main advantages over simpler alternative procedures (e.g., asking the predictor immediately after each incorrect recall attempt by the target for a yes/no judgment as to whether the target is likely to recognize the correct answer—cf. Nelson & Narens, 1980a): (a) the procedure used here does not presuppose that any absolute criterion is applied by the predictor; such criteria can change in the course of a session. (b) The reliability assessment that it permits enables us to check whether any differences in predictive accuracy might be due to differences in the reliability of predictions. A potential disadvantage of this procedure is that several minutes elapse between the recall attempts and the predictions, during which predictors may forget some relevant information. The possible consequence of this delay will be considered at the end of the Discussion, but in any case the results will show that subjects were able to remember many details from the recall phase.

Recognition phase. In the 4-alternative

forced-choice recognition test, each of the 15 nonrecalled items was displayed individually on the screen, together with four numbered response alternatives, until the subject typed the number of the chosen alternative, whereupon the next item was displayed for recognition. All subjects were tested first on the nonrecalled items from their own test and then on those from the corresponding target's test.

Subjects

Subjects were University of Washington undergraduates who volunteered for extra credit toward their psychology course grade. Because the differences in predictive accuracy between the three types of subjects may not be large (as illustrated by the nonsignificant results of Vesonder and Voss, Experiment 2, for items not previously recalled by the target), we used a much larger number of subject triads (56 in the present experiment and 50 in Experiment 2, as opposed to 16). Almost all of the subjects in the target role also served as the judge for a previous subject triad, so 113 subjects were sufficient to fill the three roles in 56 triads. Subjects were selected and assigned to triads in such a way that each of the 8 possible permutations of gender for the 3 roles occurred in 7 triads. (The gender variable produced no reliable differences and is not discussed further.)

Experiment 2

This experiment was designed as a methodologically improved replication of Experiment 1. As will be seen below, the target's average feeling-of-knowing accuracy in Experiment 1 was lower than in previous studies. Because this fact made the testing of some hypotheses more difficult, several changes were made to combat possible causes of this low accuracy. The situation of the observer was also changed so as to permit more precise yet less obtrusive observation.

The basic design and the apparatus were the same.

Subjects

The 101 subjects (39 males and 62 females) were drawn from the same pool as in Experiment 1. Fifty subject triads were formed.

Procedure

The procedure was the same as in Experiment 1 except for the following modifications:

Each observer observed the target's recall attempts from a separate room, through a one-way mirror located in front of the target; this gave the observer a clear view of the target's face and upper trunk. A video monitor located in front of the observer in the observer's room presented a display identical to that on the target's own monitor. Because the target and the observer were in separate rooms, it was not necessary for the experimenter to remain with the subjects during the recall phase to prevent communication between them. The targets were not informed that they would be observed.

To ensure that the sets of 15 nonrecalled items produced by the targets in the recall phase did not contain predominantly difficult items, the items to be presented were sampled differently than in Experiment 1. Using the recall norms compiled by Nelson and Narens (1980b), the 240 items in the pool were divided into 5 blocks in such a way as to minimize a chi-squared (least squares) across blocks in terms of the number of items recalled per block if all items in the block were attempted. The items were sampled as follows: (1) A block was randomly chosen. (2) An item was randomly presented from that block. (3) If that item was not answered correctly, another block was chosen from the remaining blocks (back to Step 1); if that item was answered correctly, another item from the same block was presented (back to Step 2). (4) This process continued until 3 questions from each of the 5 blocks had been presented that the target had failed to answer correctly.

In order to reduce measurement error resulting from correct guesses in the recognition phase, the number of response alternatives presented for each item was increased from 4 to 8.

Popularity Index for Incorrect Responses

The analysis of the role of the plausibility of the target's incorrect answers requires information on the normative frequency of specific incorrect responses (excluding the special response *next*). As this information was not compiled by Nelson and Narens (1980b), it was obtained through an analysis of data from the recall phases of the present experiments and 5 other experiments conducted in the laboratory of the second author using the FACTRETRIEVAL program. These 7 experiments yielded a total of 11,209 responses, an average of 47 per item. After correction of obvious spelling errors, it was possible to define an index of the *popularity* of an incorrect response by a target: the ratio of the number of identical incorrect responses to the same item produced by other subjects in the norming sample to the total number of commission errors for that item produced by other subjects. The reliability of the popularity index is limited, especially for items that tend to evoke few commission errors; but the results to be reported show that it captures at least some major differences in the frequency of specific incorrect responses.

Statistical Tests

The results for Experiment 1 and for Experiment 2 (a methodologically improved replication of Experiment 1) will be presented and discussed together. Each analysis involves the computation, for each triad of subjects in each experiment, of a Goodman–Kruskal G (gamma) correlation between two variables, e.g., the observer's predictions and the target's recognition performance. (The latter is a dichotomous variable; see Nelson, 1984, for reasons for choosing G in contexts such as this.) Each significance test is either a t test of the hypothesis that the population mean of a particular G is zero or a paired t test comparing two Gs. Because each of the two experiments provides an independent test of each hypothesis, the two t values obtained for a hypothesis are combined to determine the statistical significance of the result in view of the evidence from both experiments. The combination method used (Winer, 1971, p. 50) yields a z statistic that approximately follows the unit normal distribution. We report the results of a statistical test by giving the t for Experiment 1, the t for Experiment 2, and the z and the p (two-tailed) for the combined results. Because the value of a G for a subject triad is sometimes indeterminate owing to ties, the value of N and the number of degrees of freedom are not always constant within each experiment.

RESULTS

Predictive Accuracy

For each subject triad, three G correlations were computed to index the accuracy of the target, observer, and judge, respectively, in predicting the target's recognition performance. In addition, for each observer and judge, G was computed between the subject's predictions and her own recognition performance. The means of these five indices are shown in Fig. 1.

The mean accuracy coefficients for subjects in the three roles predicting the target (represented by the three left-hand bars in each part of Fig. 1) show the same relative sizes as the corresponding mean Gs ($+.32$, $+.24$, and $+.15$, respectively) reported for Vesonder and Voss' (1985) Experiment 2.

The first hypothesis formulated in the Introduction that the observers are more accurate than the judges is supported, although the difference is only marginally significant ($t(55) = 1.70$ in Experiment 1 and $t(48) = 0.99$ in Experiment 2, $z = 1.89$, $p < .07$ for the combined results by a two-tailed test). Similarly, the targets are more accu-

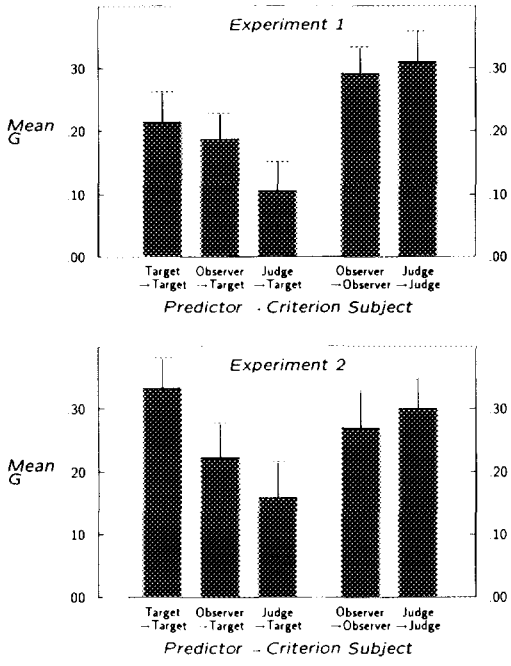

FIG. 1. Correlations between predictions and recognition performance. (In Experiment 1, $N = 56$ for all 5 mean correlations. In Experiment 2, $N = 49$ where the target's recognition performance is the criterion [first 3 bars], and $N = 50$ where the observer's and judge's predictions are related to their own recognition performance [last 2 bars]. The bracket above each bar indicates the standard error of the mean.)

rate than the observers, as predicted by the second hypothesis ($t(55) = 0.59$ and $t(48) = 1.93$, $z = 1.75$, $p = .08$).[2]

These differences in predictive accuracy might conceivably be associated with differences in the reliability of the predictions in the three roles. As noted under Method for both experiments, the retest reliability of a subject's predictions was assessed on the basis of six item triples that contained previously presented item pairs. The proportion of item pairs for which the predictor's choice was identical for both presentations serves as an index of reliability for that predictor. The mean proportions for the target, observer, and judge, respectively, are .92, .89, and .91 in Experiment 1, and .91, .93, and .87 in Experiment 2. It seems unlikely that the differences between these proportions (none of which is significant in the combined results: $p > .10$) could figure in an explanation of the accuracy differences.

The third hypothesis formulated in the Introduction, that even the judges would show above-chance accuracy, is confirmed: The judges' two mean accuracy correlations are significantly greater than zero ($t(55) = 2.37$, $t(48) = 2.77$, $z = 3.56$, $p < .001$).

The fourth and final hypothesis was that the predictions of the judges and the observers would be strongly related to their own knowledge about the items. As expected, the two right-hand bars in each part of Fig. 1 show that these subjects are substantially self-predictive ($p < .001$ for each role in each experiment). Moreover, in our samples their predictions are more strongly related to the predictor's own recognition performance than to that of the target's that they were supposed to predict; and in Experiment 1 the observer and judge actually "predict" their own performance better than the target predicts his own performance.

This relatively high self-predictiveness can be explained in part by the fact that the average observer or judge must have been able to produce the correct answer to about 3 of the target's 15 nonrecalled items (the mean normative recall probability of a target's nonrecalled items was .21 in Experiment 1 and .22 in Experiment 2). Presumably an observer or judge would tend (a) to make especially optimistic predictions for the items she could answer herself and (b) to recognize these items correctly in the recognition phase.[3]

[2] Each of the two differences reported in this paragraph is significant with $p < .05$ if one uses one-tailed tests for the specific directional hypotheses formulated in the Introduction, which were based on previous research.

[3] Even if the observers and judges did not use their own responses to the items as evidence, their predictions would have some accuracy for their own perfor-

Before looking closely at the implications of these results in terms of the sorts of knowledge used by subjects in the three roles, we must examine the results concerning specific cues in the target's behavior.

Specific Observable Cues

For each of the three aspects of the target's recall–phase behavior from which data were collected, we will examine (a) its relationships to the predictions of the observer and the target and (b) its validity as a cue to the target's recognition performance.

Type of Recall Failure

Perhaps the most obvious cue in the target's response to an item is whether he typed in some potential answer (commission error) or simply gave up by typing the word *next* (omission error). The mean proportion of commission errors among a target's 15 nonrecalled items was .48 in Experiment 1 and .52 in Experiment 2.

Krinsky and Nelson (1985) designed a noncomputerized experiment to compare these two types of recall failure for subjects in the target role. They found that subjects tended to give higher feeling-of-knowing rankings to items on which they had made a commission error, even though they were fully aware that their original answer had been incorrect. These authors did not, however, address the issue of how an observer might interpret a commission error.

TABLE 2
MEAN CORRELATIONS BETWEEN TYPE OF TARGET'S RECALL FAILURE AND SUBJECTS' PREDICTIONS

Predictor	Experiment 1[a]	Experiment 2[b]
Target	+.41**	+.46**
Observer	+.40**	+.45**
Judge	+.22**	+.18*

Note. Each value represents the mean *G* correlation between the dichotomous variable "type of recall failure" (explained in the text) and the predictions of the subject in the specified role. Standard errors are less than .063.

[a] $N = 53$.
[b] $N = 47$.
*$p < .01$.
**$p < .001$.

It is convenient to treat the variable "type of recall failure" as a dichotomous ordinal variable, with the value "commission" ranking higher than the value "omission." We can then compute a *G* correlation between this variable and a subject's predictions. Table 2 shows the mean *G*s for subjects in all three roles.

The high mean correlations for the targets reflect the tendency reported by Krinsky and Nelson for targets to be more optimistic after commission errors.[4] The corresponding mean *G*s for the observers are as high as those for the targets and much higher than those for the judges (for the observers vs the judges $t(52) = 2.88$ and $t(46) = 4.01$, $z = 4.77$, $p < .001$). This pattern suggests that the observer's predictions—and perhaps those of the target as well—were influenced by the nature of the target's recall failure.

The fact that the mean correlations for the judge are also significantly positive shows that there must be some item characteristics that tend to give rise to both commission errors and relatively optimistic predictions, even for persons who do not observe a target's recall attempt.

Validity. Krinsky and Nelson (1985) re-

mance, simply because item difficulty has some generality over persons. Using an appropriate index of ordinal partial correlation, it is possible to control for the role of information that is not specifically related to the person being predicted (Jameson, 1990, Chap. 2 and Appendix A). When this is done, the self-predictiveness of the observers and judges remains highly significant. A similar analysis shows that the superiority of the targets and observers to the judges is not due merely to the fact that the targets and observers had more time to think about the items and therefore to make use of actuarial information that was not specifically related to the target.

[4] Krinsky and Nelson also reported this tendency in an analysis of part of the data for the targets in the present two experiments.

ported that there was no reliable tendency for a target's recognition performance to be better with items on which he had made a commission error. So in spite of the strong relationship that this cue has with the predictions of targets and observers, its validity is at best very low.

Latency of Recall Attempt

The observer might also take into account the length of time that the target took to respond to an item. Nelson *et al.* (1984) investigated the relationship between the feeling-of-knowing and the latency of a target's unsuccessful recall attempt. They found a clear pattern of results only when the *G* correlations were computed conditionally on the type of recall failure: within the subset of items on which a subject had made omission errors, the items that the subject had spent more time thinking about tended to elicit significantly higher feeling-of-knowing rankings; on the subset of commission-error items, there was no reliable correlation between recall latency and feeling-of-knowing ranking.

The present experiments yield corresponding conditional *G* correlations for observers and judges as well as for targets (Table 3). In the right-hand side of the Table, only one relationship between commission-error latency and predictions can be discerned: the mean correlations for the observer in both experiments are slightly negative ($t(53) = -1.79$ and $t(46) = -1.58$, $z = -2.33$, $p = .02$). This result suggests that an observer tends to be slightly more positively impressed by a commission error when it is made quickly than when it is made slowly. Further research will be required to determine whether this tendency is reliable and how it might be explained.

As in the experiment of Nelson *et al.* (1984) cited above, the results for the omission-error items are clearer. For the target, the mean correlations are very similar to those reported by Nelson *et al.* Although the means for the observers are somewhat lower than those for the targets ($t(52) = 2.11$ and $t(44) = 1.74$, $z = 2.66$, $p < .01$), they are also clearly higher than those for the judges ($t(52) = 2.89$ and $t(44) = 2.87$, $z = 3.99$, $p < .001$). So it appears that the target's omission-error latency had considerable impact on the predictions of the observer: when the target failed to produce any specific answer, the observer was relatively optimistic if he at least did not give up quickly.

TABLE 3
MEAN CORRELATIONS BETWEEN LATENCY OF TARGET'S RECALL ATTEMPT AND SUBJECTS' PREDICTIONS

	Type of recall failure			
	Omission		Commission	
Predictor	Expt 1[a]	Expt 2[b]	Expt 1[c]	Expt 2[d]
Target	+.40**	+.46**	+.04	−.04
Observer	+.26**	+.37**	−.11	−.10
Judge	+.07	+.17	+.02	+.03

Note. Each value represents the mean *G* correlation between the latency of a target's recall attempt and the predictions of the subject in the specified role, computed for the subset of items on which the target's recall failure was of the specified type. Standard errors are less than .079.

[a] $N = 53$.
[b] $N = 45$.
[c] $N = 54$.
[d] $N = 47$.
** $p < .001$.

The correlations for the judge are also slightly positive ($t(52) = 1.18$ and $t(44) = 2.69$, $z = 2.68$, $p < .01$), suggesting the existence of some item characteristics associated with both slow omission errors and generally higher predictions.

Validity. We can index the validity of omission-error latency as a predictor of recognition performance by computing, for each target, a *G* correlation between his latency on omission-error items and his later recognition performance on those same items. The resulting mean correlations are greater than zero in both experiments: +.06 (±.09) in Experiment 1 and +.26 (±.09) in Experiment 2 ($t(48) = .72$ and $t(37) = 2.89$, respectively, $z = 2.48$, $.01 <$

$p < .02$). The discrepancy between the two means may be due largely to a higher rate of successful guessing in Experiment 1. Omission-error latency may therefore have some validity as a cue to a target's recognition performance. In any case, its validity is surely limited by a variety of largely irrelevant factors that can influence the latency of a recall attempt, e.g., the complexity of the reasoning called for by an item and random fluctuations in the concentration of the target.

Plausibility of Incorrect Responses

The importance of the third cue to be considered is obvious in extreme cases: if a target offers a ridiculous incorrect answer to a question, an observer should be relatively pessimistic about his chances of recognizing the correct answer, compared to cases where she thinks his answer was in some sense almost correct. But most incorrect responses given to the FACT-RETRIEVAL items have an intermediate degree of plausibility, and it is an empirical question whether, within this naturally occurring range, the plausibility of an incorrect response is related to predictions of recognition performance.

It is difficult to define *plausibility* precisely, because it is not an objectively measurable property of a response (as are the two observable cues considered earlier in this section): an assessment of plausibility depends on the knowledge of the person making the assessment. Any objective index of plausibility can therefore correspond only roughly with the assessments made by individual predictors. For this exploratory analysis, we use the index of response popularity described under Method, on the basis of the following assumption: the greater the popularity of an incorrect response within a given population, the more plausible it will tend to seem to a subject from that population.

Table 4 shows that the popularity index is significantly correlated with the target's predictions in each experiment. The mean correlations are higher than those for the judge ($t(53) = 1.54 = t(46) = 2.61$, $z = 2.88$, $p < .01$). They confirm that the popularity index reflects some aspect of the target's specific incorrect response that is related to feeling-of-knowing judgments.

TABLE 4
MEAN CORRELATIONS BETWEEN POPULARITY OF TARGET'S RECALL RESPONSE AND SUBJECTS' PREDICTIONS

Predictor	Experiment 1[a]	Experiment 2[b]
Target	+.19*	+.25**
Observer	+.08	+.27**
Judge	+.07	+.07

Note. Each *G* correlation was computed for the subset of items on which the target made a commission error (other than a misspelling of the correct answer). Standard errors are less than .064.

[a] $N = 54$.

[b] $N = 47$.

* $p < .01$.

** $p < .001$.

The results for the observer are different in the two experiments. In Experiment 1 there is no apparent impact of the content of the target's specific incorrect answer, but in Experiment 2 the mean correlation for the observer is as high as that for the target and significantly higher than that for the judge ($t(46) = 2.87$, $p < .01$). A post hoc explanation for this discrepancy involves the major procedural difference between the two experiments: whereas in Experiment 2 the observer could follow the target's behavior on a separate monitor, in Experiment 1 she was seated slightly behind and to the side of the target. The observer may therefore have been in a less suitable position to read and think about the specific responses entered by the target. (Note that in order to use the two other cues discussed above, the observer only had to notice whether the target's response was the word *next*, and how soon he entered it.)

Although the results for the observers in Experiment 2 must await replication in future research, the overall pattern of results suggests that the popularity index is related

to normative plausibility, which is in turn related to the predictions of both targets and observers.

Validity. There is no positive relationship between the popularity index and the target's later recognition performance. The mean *G* correlations are not significantly different from zero, even when the results from the two experiments are combined, and they are even slightly negative: −.12 (±.09) and −.10 (±.09), respectively.

Discussion

We first discuss the results concerning the predictive accuracy of subjects in the three roles. We then consider the results of the exploratory analyses concerning the cues in the target's behavior. Figure 2 gives an overview of the variables that may have influenced the predictions of the subjects in the three roles.

Sources of Predictive Accuracy

In terms of Fig. 2, a conclusion drawn by Vesonder and Voss (1985, pp. 375–376) can be paraphrased as follows: predictions for items not previously recalled are based mainly on actuarial information—which does not differ systematically for predictors in the three roles—and on the predictor's internal responses, which also tend to be similar. By contrast, our findings of accuracy differences between the roles suggest that there were in fact differences among the three subject roles with respect to the predictive value of the types of information available to them, as discussed next.

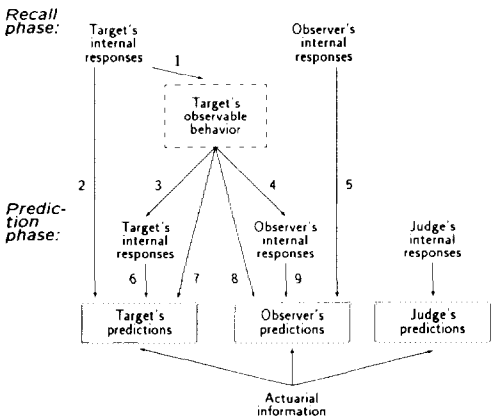

FIG. 2. Theoretical integration of the variables influencing the subjects' predictions. (The predictions of the target, observer, and judge were directly measured. The target's observable behavior consists of several variables, three of which were directly measured. Arrows represent the hypothesized influences.)

The Target's Observable Behavior

The main advantage for the observer over the judge is the availability of information about the target's observable behavior. Figure 2 shows that this information could influence the observer in two ways: (a) she could use it directly as evidence about the target's knowledge (Arrow 8); (b) the target's behavior could change the observer's own beliefs about the item, thereby affecting her internal responses in the prediction phase (Arrow 4), which could in turn affect her predictions (Arrow 9). For example, an observer who initially believed that the capital of New York State was New York City might change her belief if the target himself unsuccessfully tried "New York City."

The fact that the observers outpredicted the judges shows that in one or both of these ways the observers benefited from their access to the target's behavior. The results on specific cues take only a first step toward explaining just what information can be of use: the mean validity coefficient of +.26 (±.09) for omission-error latency in Experiment 2 suggests that this cue may have some validity. The observers may also have made use of information in the targets' behavior that was neither recorded in our experiments nor strongly correlated with the variables that were recorded (e.g., a facial expression suggesting unfamiliarity with the concepts mentioned in a question).

The Target's Internal Responses

Figure 2 suggests that the target and the observer have similar types of information available: actuarial information, the target's observable behavior, and the predictor's own internal responses (during both

the recall and the prediction phases). The superior accuracy of the target's predictions therefore does not seem to be due to an *additional* type of information that the observer completely lacks, but rather to the target's internal responses being more valuable as cues to the target's future recognition performance than are the observer's internal responses.

Information Not Specific to the Target

The results for the judges confirm that a predictor can attain above-chance—though very modest—accuracy while lacking virtually any specific information about the target subject. Figure 2 suggests that this accuracy is due to the judge's use of some combination of actuarial information and her own internal responses. Although we cannot distinguish quantitatively the relative contributions of these two types of information, the results on the self-predictiveness of the judges suggest that the judges' own responses were an important source of evidence for them.

Moreover, although the observers had access to a considerable amount of specific information about the target (and they made use of this information, as shown by the results concerning specific observable cues), their predictions were still strongly related to their own knowledge. Put differently, information about one's own responses is not used only in the absence of more specifically relevant information.

Use of Specific Observable Cues

Our results concerning specific cues in the target's behavior parallel those obtained by Vesonder and Voss (1985, Experiment 2) concerning their cue of previous recall success: each of the cues was correlated with the target's own predictions and (sometimes to a lesser degree) with those of the observer. Vesonder and Voss interpreted their results essentially in terms of *self-perception* on the part of the target: like the observer, he made inferences about his knowledge of the items on the basis of his previous behavior. A similar account is possible for the present experiments. The target may have remembered and interpreted his previous behavior (Arrow 7), just as the observer evidently did (Arrow 8). And the target's unsuccessful attempt may sometimes have changed the target's beliefs about the item (Arrows 3 and 6), as discussed for the observer (Arrows 4 and 9).

However, Fig. 2 shows that there are also other possible causal relationships between the target's observable behavior and the target's later predictions. The simplest type of explanation involves Arrows 1 and 2: an item gives rise to particular internal responses in the target, which in turn can influence both the target's observable behavior and his later predictions. For example, if the target finds that plausible answers to an item readily come to mind, these internal responses may encourage him to venture a guess (Arrow 1). In the prediction phase, after his guess turns out to have been incorrect, his recollection that he had previously found it easy to generate plausible answers may make him confident of his ability to recognize the correct answer (Arrow 2; cf. Krinsky & Nelson, 1985, pp. 152–157). Similarly, if the target is unable to produce a response immediately, he may spend more time trying to find an answer if the concepts mentioned in the question seem familiar to him, or if he can retrieve some indirectly relevant knowledge (Arrow 1); and these initial internal responses may also make him more confident that he can recognize the correct answer (Arrow 2; cf. Glucksberg & McCloskey, 1981; Nelson *et al.* 1984; Nelson & Narens, 1980a,b; Reder, 1987, pp. 120–122).

Even if the target is unable during the prediction phase to remember his earlier internal responses (Arrow 2), he may think about the item again, thereby experiencing internal responses much like those in the recall phase; and these later internal re-

sponses may influence his predictions (Arrow 6) in much the same way as remembering the earlier responses would.

Still further possible explanations could be given in terms of Fig. 2 for the correlations between a target's observable behavior and his predictions. In short, our analysis does not so much help to narrow down the set of possible explanations as to demonstrate the number of possibilities that must be taken into account. But our empirical results do show that a self-perception account for the target requires no questionable assumptions about what a subject might remember or infer. After all, the observer was often able in the prediction phase to remember the relevant aspects of the target's recall-phase behavior; and the observer made inferences on the basis of this behavior that influenced her predictions.

The relative contribution of the various paths in Fig. 2 may depend on the length of the delay between the target's recall attempts and the making of predictions by the target and observer. Future research should examine situations in which there is no distinction between a recall and a prediction phase, e.g., where predictions are made immediately after each incorrect recall attempt (cf., e.g., Fussell & Krauss, 1991). Then the relative contributions might change for initial internal responses (cf. Arrows 2 and 5) versus the target's observable behavior (Arrows 3, 4, 7, and 8). It is not obvious whether in such a situation the impact of the target's observable behavior on predictions would be relatively weaker or stronger.

Conclusion

One way to summarize our findings is to compare them with the findings from Experiment 2 of Vesonder and Voss (1985).

First, Vesonder and Voss concluded that on previously nonrecalled items the target had "essentially no distinct advantage" relative to the observer and the judge (p. 374). In our experiments, however, the target outpredicted the observer, who in turn outpredicted the judge; the target's accuracy was about twice as great (in terms of G) as that of the judge in each experiment. These differences in accuracy suggest the following about predictions of a target's recognition performance following recall failure on general-information items: (a) cues in the target's observable behavior can enhance accuracy beyond what can be attained using only information that is not specifically related to the target; and (b) the target also benefits from information about aspects of his own knowledge that is not available to an observer.

Second, while replicating Vesonder and Voss' result that even a predictor without specific knowledge about a target can predict the latter's performance with above-chance accuracy, the results for our judges show that such predictions are not based entirely on actuarial information: the judges made use of their own internal responses as evidence (as did the observers). So although the potential value of actuarial information has been demonstrated by Nelson *et al.* (1986) and by Calogero and Nelson (in press), future research will be required to determine the extent to which predictors possess and use such information.

Third, our results concerning three specific cues in our targets' observable behavior parallel the corresponding results of Vesonder and Voss in two respects: they show that these cues influenced the predictions of the observers; and they demonstrate the plausibility of a self-perception mechanism for the targets. But in contrast to Vesonder and Voss' cue of previous recall success, the cues investigated here were not previously known to have any impact on predictions, and their impact does not appear to have much justification in terms of their validity.

One goal of future research should be to focus more closely on how the various

types of information distinguished here are combined into one overall prediction concerning the knowledge either of oneself or of another person.

REFERENCES

BURTON, M. L., & NERLOVE, S. B. (1976). Balanced designs for triads tests: Two examples from English. *Social Science Research*, **5**, 247–267.

CALOGERO, M., & NELSON, T. O. (in press). Utilization of base-rate information during feeling-of-knowing judgments. *American Journal of Psychology*.

DAWES, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, **25**, 1–17.

FUSSELL, S. R., & KRAUSS, R. M. (1991). Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology*, **21**, 445–454.

GLUCKSBERG, S., & MCCLOSKEY, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory*, **7**, 311–325.

HOCH, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology*, **53**, 221–234.

JAMESON, A. (1990). *Knowing what others know: Studies in intuitive psychometrics*. Unpublished doctoral dissertation, University of Amsterdam, The Netherlands.

KING, J. F., ZECHMEISTER, E. B., & SHAUGHNESSY, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology*, **93**, 329–343.

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

LEONESIO, R. J., & NELSON, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**, 464–470.

LOVELACE, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**, 756–766.

MARKS, G., & MILLER, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, **102**, 72–90.

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

NELSON, T. O. (1992). *Metacognition: Core readings*. Boston: Allyn & Bacon.

NELSON, T. O., GERLER, D., & NARENS, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and relearning. *Journal of Experimental Psychology: General*, **113**, 282–300.

NELSON, T. O., LEONESIO, R. J., LANDWEHR, R. S., & NARENS, L. (1986). A comparison of three predictors of an individual's memory performance: The individual's feeling of knowing vs. the normative feeling of knowing vs. base-rate item difficulty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **12**, 279–287.

NELSON, T. O., & NARENS, L. (1980a). A new technique for investigating the feeling of knowing. *Acta Psychologica*, **46**, 69–80.

NELSON, T. O., & NARENS, L. (1980b). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, **19**, 338–368.

NICKERSON, R. S., BADDELEY, A., & FREEMAN, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? *Acta Psychologica*, **64**, 245–259.

REDER, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology*, **19**, 90–138.

SHIMAMURA, A. P., LANDWEHR, R. F., & NELSON, T. O. (1981). FACTRETRIEVAL: A program for assessing someone's recall of general-information facts, feeling-of-knowing judgments for nonrecalled facts, and recognition of nonrecalled facts. *Behavior Research Methods & Instrumentation*, **13**, 691–692.

VESONDER, G. T., & VOSS, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language*, **24**, 363–376.

WILKINSON, T. S., & NELSON, T. O. (1984). FACTRETRIEVAL2: A PASCAL program for assessing someone's recall of general-information facts, confidence about recall correctness, feeling-of-knowing judgments for nonrecalled facts, and recognition of nonrecalled facts. *Behavior Research Methods, Instruments, & Computers*, **16**, 486–488.

WINER, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

(Received December 31, 1991)

(Revision received July 2, 1992)