Memory & Cognition
2001, 29 (2), 222-233

Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs?

GREG MATVEY, JOHN DUNLOSKY, and ROBERT GUTTENTAG
University of North Carolina, Greensboro, North Carolina

The fluency of retrieval during a test of memory has been implicated as a cue for judgments of learning (JOLs), but little is known about how fluency affects JOLs. In three experiments, we investigated (1) whether the fluency of generation during study may be a cue for JOLs and (2) whether such fluency effects are mediated by an analytic or nonanalytic inference. To accomplish our goals, we used a learner–observer–judge method. While studying paired associates, learners generated some targets at study. For these items, their JOLs were negatively correlated with the time taken to generate targets. Observers watched learners generate targets and then predicted learners' memory performance. Judges also made JOLs but did not watch the learners generate targets. JOLs from all groups were negatively related to learners' latencies to generate targets, with the magnitude of the relationship equivalent for learners and observers and lower for judges. These and other findings are consistent with the conclusions that the fluency of generation at study is a cue for JOLs and that such fluency effects are partly mediated by an analytic inference about how fluency is related to memory.

To understand how people monitor memory, researchers have examined people's predictions of their future performance for recently studied items, a kind of judgment referred to as a *judgment of learning* (JOL). JOLs have been investigated extensively in part because of their central role in models of self-regulated learning (Nelson & Narens, 1990; Schwartz, 1994; Thiede & Dunlosky, 1999). Theory of JOLs has increasingly focused on the inferential nature of these judgments. That is, when making a JOL, people presumably draw inferences about their future performance based on a variety of cues pertaining to the task (Koriat, 1997). Researchers have discovered numerous cues that influence JOLs (e.g., Benjamin, Bjork, & Schwartz, 1998; Carroll, Nelson, & Kirwan, 1997; Dunlosky & Nelson, 1994; Koriat, 1997); our approach complements this research by investigating the nature of the inference that mediates the effects of particular cues on a person's JOLs. In particular, we evaluated how one cue, the fluency of retrieval at study, influences people's JOLs.

Understanding the nature of the inference that mediates the effects of cues on JOLs represents a critical extension of inferential-based investigations of JOLs. Previous research has primarily focused on uncovering empirical generalizations about which cues affect JOLs. Less attention has been paid to evaluating the nature of the inferences involved in making a JOL. One exception is Benjamin et al.'s (1998) analysis of retrieval fluency as a basis for JOLs. They found a negative relationship between latency of retrieval and people's confidence that the items would be recalled on a later test of memory: As items took longer to retrieve, people were less confident that they would later remember them. Their interpretation of this relationship was that people possess an implicit heuristic about retrieval fluency. This inference-based interpretation, however, was not empirically evaluated. Our primary goal was to evaluate the nature of the inference that mediates the effects of the fluency of retrieval, a cue that presumably has a substantial influence on JOLs (e.g., Benjamin et al., 1998; Koriat, 1997). More specifically, we examined whether analytic or nonanalytic processes underlie these inferences.

Analytic and nonanalytic processes draw on qualitatively different kinds of information, making the distinction useful in understanding how people make judgments (Kelley & Jacoby, 1996a). Analytic processes draw on theories or beliefs about how memory operates. An example is a belief a person has that quickly generating items at study is better for memory than is slowly generating items. Nonanalytic processes are based on the subjective experiences people have as they perform a task. An example is the subjective experience people have as they generate targets during study. Kelley and Jacoby (1996a) used the distinction between analytic and nonanalytic processes to characterize how people make judgments of anagram difficulty. In their investigation, participants solved anagrams and judged how difficult the anagrams would be for others to solve. A critical outcome was that slower solution latencies were related to higher ratings of difficulty. Their interpretation of these and other outcomes

The authors thank R. Reed Hunt and Susan L. McDonald for many productive discussions about this research. Correspondence should be addressed to J. Dunlosky, P. O. Box 26164, Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC 27402-6164 (e-mail: dunlosky@uncg.edu).

was that participants "used their subjective experience of anagram difficulty as a basis for judging for others" (Kelley & Jacoby, 1996a, p. 163). That is, judging anagram difficulty involves the use of an inference based on a nonanalytic process involving subjective experience (henceforth, a *nonanalytic inference*) rather than on the explicit use of a theory about factors that relate to the difficulty of solving anagrams (an *analytic inference*).

Although Kelley and Jacoby's (1996a) research demonstrates how analytic and nonanalytic inferences may influence subjective judgments, their conclusions regarding anagram difficulty judgments will not necessarily generalize to JOLs. Anagram difficulty judgments are assessments of how difficult a particular anagram would be for someone else to solve. In contrast, JOLs are predictions of one's own ability to remember a currently studied item on a future test. Because different kinds of judgments have different bases (e.g., Costermans, Lories, & Ansay, 1992; Leonesio & Nelson, 1990; Thiede & Dunlosky, 1994), anagram judgments and JOLs may differentially draw on analytic and nonanalytic inferences. Thus, the nature of the inferences that mediate the effects of cues on JOLs remains an open question.

To answer this question, we used a target generation task. During this task, an individual attempts to generate targets (e.g., *wing*) when presented a cue and a stem (e.g., *rhyme*: *ring–w* _ _ _). In the generation task, the fluency of retrieval is associated with the act of generating a target word at study, which is essential for investigating the relationship between fluency and JOLs. The generation task also highlights the possible contributions of analytic and nonanalytic inferences in making JOLs. Specifically, people may base JOLs on an analytic inference based on a theory about retrieval fluency. Alternatively, people may base JOLs on a nonanalytic inference based on the subjective experience of retrieving targets. Thus, the generation task provides a way to evaluate the degree to which analytic and nonanalytic inferences mediate the effects of retrieval fluency on JOLs. Note that we are less interested in the generation task itself than we are in using this task to discover the nature of the inferences underlying JOLs.

Another advantage of using the generation task is that it has already been used in conjunction with JOLs. Begg, Vinski, Frankovich, and Holgate (1991) had participants generate some items at study and read others. After studying each item, participants made a JOL, where greater values indicated more confidence in subsequent memory performance. People's JOLs were greater for items that had been generated at study than for items that had been read at study (see also Mazzoni & Nelson, 1995). We refer to this outcome as a *generation effect on JOLs* because the magnitude of the JOLs was differentially influenced by generating versus reading targets. Begg et al. concluded that this generation effect occurred because participants "expected that generating would lead to better memory than reading" (p. 493). That is, participants had a theory about the generation effect, such as that "generating targets at study is better for memory than reading targets." Although results from Begg et al. support the idea that an analytic inference mediates the generation effect on JOLs, they do not rule out an alternative explanation. In particular, given the different subjective experiences associated with generating versus reading targets, the effects of generation (vs. reading) on JOLs may be mediated by a nonanalytic inference. Even though we further explore the generation effect on JOLs, this aspect of our research was secondary to understanding the effects of fluency on JOLs. Thus, it is important not to confuse the generation effect on JOLs (which involves both generated and read targets) with the relationship between generation fluency and JOLs (which involves only generated targets).

As in the case with the generation effect on JOLs, a nonanalytic inference may also mediate the relationship between retrieval fluency and JOLs (cf. Benjamin et al., 1998). Of course, people may also believe that generating targets more fluently at study indicates better memory, which suggests that an analytic inference may instead mediate this relationship. A critical implication here is that both analytic and nonanalytic inferences may underlie the effects of retrieval fluency on JOLs. Importantly, previous research has not evaluated the degree to which these inferences mediate the relationship between retrieval fluency and JOLs. In the remainder of this introduction, we explain our approach to exploring this issue. Our specific illustrations concern the relationship between generation (aka *retrieval*) fluency and JOLs because this relationship is most central to current hypotheses of JOLs (e.g., Benjamin et al., 1998; Koriat, 1997) and is the main focus of all three experiments described in the present research.

We investigated whether the effects of generation fluency on JOLs are mediated by analytic or nonanalytic inferences by using the learner–observer–judge method (Brennan & Williams, 1995; Jameson, Nelson, Leonesio, & Narens, 1993; Vesonder & Voss, 1985). This method includes three groups: a learner who studies and predicts his or her own performance (as described above), an observer who "observes" a learner generating and reading targets during study, and a judge who is shown a learner's study items but does not observe the learner studying.[1] Participants are assigned to triads, with each learner yoked to an observer and a judge. Both observers and judges make JOLs for the yoked learner's performance. Because analytic and nonanalytic inferences differentially contribute to JOLs made by each group, data from the groups can be used to draw conclusions about the contribution of these inferences to JOLs. To illustrate how this was accomplished, we summarize the analytic and nonanalytic inferences that may influence the JOLs (with respect to generation fluency) made by each group in Table 1.

A learner's JOLs may be influenced both by an explicit theory about generation (under "Analytic Influence" in Table 1) and by the subjective experience of generating

Table 1
Summary of the Influences of Generation Fluency on Judgments of Learning for Groups Used in the Learner–Observer–Judge Method

Group	Task	Analytic Influence	Nonanalytic Influence
Learner	Generate target (e.g., *cave–s* _ _ _)	Theory about generation fluency Theory about the study situation[d]	Subjective experience of generation fluency
Observer[a]	Observes a learner perform task Generate targets not filled in	Theory about generation fluency Theory about the study situation	No access to learner's subjective experience
Judge[b]	Items presented in same order as presented to a learner Does not observe learner perform task Generate targets filled in	Theory about the study situation	No access to learner's subjective experience
Pure observer[c]	Observes a learner perform task Generate targets filled in	Theory about the study situation Theory about generation fluency	No access to learner's subjective experience

[a]Observer group was included in Experiments 1 and 2. [b]Judge group was included in Experiments 1 and 3. [c]Pure-observer group had generate targets filled in (e.g., *cave–sa̲ v̲ e̲*) and was included in Experiments 2 and 3 to rule out the possibility that an observer bases judgments on his or her own generation fluency. [d]Theory about the study situation refers to the influence of some factor other than generation fluency that was related both to generation fluency and to judgments of learning. See text for details.

items (under "Nonanalytic Influence"), or some third factor (e.g., word frequency) that is related to generation fluency. Thus, learners' JOLs cannot typically be used in isolation to evaluate the extent to which an analytic or nonanalytic inference mediates the relationship between generation fluency and JOLs. By contrast, the contribution of these inferences to JOLs can be evaluated by examining JOLs from the observer and judge groups.

For a given observer, items were presented in the same order and in the same format (e.g., *generate*: *rhyme*: *wing–r* _ _ _) that was used for the yoked learner. Each target the yoked learner generated was presented to an observer at the rate taken by the learner to generate that word. Note that an observer was not physically present in the same room watching a learner complete the task, but instead he or she viewed (on a computer screen) a simulation of the learner's responses from an earlier session. Most important, an observer is deprived of the subjective experience of a learner but still has access to the learner's observable responses, such as the latencies to generate each target (Table 1). Thus, an observer's JOLs may be based on a theory about the learner's observable responses, which may be about how differences in the learner's generation latencies relate to memory, or a third factor (e.g., word frequency) that is related to generation fluency and JOLs. One critical outcome is the correlation between observers' JOLs and the learners' latencies to generate targets. In particular, observers do not have access to the learners' subjective experience of generating targets. Thus, a nonzero correlation between observers' JOLs and learners' latencies to generate targets is presumably mediated by the observers' theory about how generation latencies relate to memory. Similar correlations for the observers and learners would suggest that the relationship between fluency and JOLs is based primarily on an analytic inference.

Because a third factor may also contribute to any latency–JOL relationship for learners and observers, including a judge group is also informative. For judges, the items from the yoked learner were presented in the same order. Judges were presented with only a description of the conditions of presentation (e.g., *generate*: *rhyme*) along with the cue and the completed target for each trial. Thus, a judge's JOL will be driven primarily by an explicit theory about the study situation (Table 1). Because a judge is not aware of a learner's generation latencies, the correlation between judges' JOLs and learners' generation latencies is expected to be near zero. A nonzero correlation here would suggest that generation latency and JOLs may be related to some third factor, such as item characteristics. If the correlation for judges is the same magnitude as the corresponding correlation for the learners, this third factor may entirely account for the latency–JOL relationships, as if generation fluency itself does not directly influence people's JOLs. Although this possibility is plausible and may account for other fluency–judgment relationships presented in the literature, it has not yet been thoroughly evaluated. Finally, even if the judges' correlation is nonzero, given that it is smaller in magnitude than the correlations for learners and observers, the conclusion would be that some third factor cannot entirely account for the latency–JOL relationships.

In summary, the present research was motivated by two central questions. Is the fluency of generating targets at study related to JOLs? An affirmative answer is expected from previous research (e.g., Benjamin et al., 1998). Such a negative relationship between learners' latencies to generate targets and JOLs would suggest that generation fluency is a cue for people's JOLs. And, as important, is the relationship between generation fluency and JOLs mediated by an analytic inference or nonanalytic inference? Comparing outcomes from all three groups will provide evidence relevant to answering this question.

EXPERIMENT 1

Method

Participants and Design. The design of Experiment 1 was a 2 (cue: rhyme or category) × 2 (target: generate or read) × 3 (group: learner, observer, judge) mixed-model factorial, with cue and target as within-participants factors and group as a between-participants factor. Eighty-nine undergraduates from the University of North Carolina at Greensboro (UNCG) participated to fulfill a course requirement. Thirty participants were assigned to each of the three groups by order of appearance (with learners of each triad of yoked participants being run first); data from 1 judge were not accessible due to computer error.

Materials. All items in the experiment were presented on Macintosh computers. Items consisted of 60 cue–target pairs, with the target either rhyming with a cue (e.g., *ring–wing*) or belonging to the same category as a cue (e.g., *silver–gold*). Thirty items were in the rhyme condition, and thirty items were in the category condition. Rhyme items were chosen from the Paivio, Yuille, and Madigan (1968) norms. All targets were between four to six letters in length and were medium- to high-frequency words (21 to AA) (Thorndike & Lorge, 1944). No two pairs of items rhymed with each other (e.g., *ring–wing*, *king–sing*). Category items were chosen from the Battig and Montague (1969) category norms. Both the cue and the target of each item in the category condition came from the 3rd to 11th instance of a particular category. Targets were from four to six letters in length. Twenty-one of the category targets were medium- to high-frequency words (21 to AA), and the remaining 9 targets were low-frequency words (< 15). A particular category occurred once across the 30 category pairs. An additional 120 high-frequency nouns (A or AA) were chosen from the Paivio et al. (1968) norms to serve as distractors on the recognition test.

Two lists were then constructed using the 60 items. Both lists were blocked by cue (rhyme or category). Each block contained 15 items with read targets and 15 items with generate targets, with the order of items randomized within a block. Accordingly, the generate and read targets were nested (and placed in random order) within the rhyme and category conditions. The block of category items appeared first and was followed by the block of rhyme items in List A, whereas this order was reversed in List B.

Procedure. The participants were tested individually. Lists were randomly assigned to the learners so that 15 learners received List A and 15 learners received List B.

Learner. The learners were instructed to study each item for a future recognition test. Each item was shown on the screen along with the word *rhyme* or a category label (e.g., *animal*) to the left of the item. A read target was indicated by the word *read*, which was presented above the item. For read targets, the learners were instructed to read the target of the item silently. A generate target was indicated by the word *generate*, which was presented above the item. For each to-be-generated target, the learners were prompted by (1) the cue word, (2) the first letter of the target word, and (3) the number of spaces corresponding to the number of letters needed to complete the target word (e.g., *cave–s* _ _ _, for *cave–save*). After the prompt was presented, the learners were to type the target word as soon as they had retrieved it. As a learner typed his or her response, it was presented in a response window placed directly below the prompt for the to-be-generated target. After reading or generating the target, the learners then made a JOL (described in detail below) for that target. As the learners performed this task, their latencies to read and generate targets were recorded. For items with a to-be-read target, we measured the time between the presentation of a to-be-read item and a learner's keypress ("Return") to proceed to the JOL prompt. For items with a to-be-generated target, we measured the time between the presentation of the item (i.e., *cave–s* _ _ _) and both a learner's first keypress to type in a response and the final keypress to proceed to the JOL prompt. As explained below, these latencies were used in the presentation of a learner's responses to the yoked observer.

Observer. Each observer was yoked to a learner. The observers were instructed that they would be presented with a recorded presentation of a learner performing a task involving generation and reading. However, the observers received no instructions about using information from this presentation to make JOLs for the yoked learner. For an observer, items were presented in the same order that the yoked learner had been presented the items. Moreover, the same prompt used to collect a learner's response was presented to the yoked observer: Each item was presented on the screen in the same format used for the learners (e.g., *cave–save* for items with a to-be-read target or *cave–s* _ _ _ for items with a to-be-generated target). The word *rhyme* or a category label (e.g., *animal*) was presented to the left of the item, and the word *read* or *generate* was presented at the top of the screen to indicate whether the yoked learner had read or generated the target of that item.

For targets the yoked learner had generated, the target was not initially presented with the cue; instead, the target was presented in the response window sometime after the prompt for the target had been presented. For a given item, the interval between the presentation of the cue and the target was equal to the amount of time that the learner had originally used to begin typing that target. The item was then presented for the amount of time that the learner had used to type the target, which was operationalized as the amount of time between the learner's first keypress to begin typing and the final keypress to proceed to the JOL prompt. Thus, an observer could observe how much time the yoked learner used to generate a target. For targets the yoked learner had read, the item appeared for the amount of time that the learner had originally taken before proceeding to the JOL prompt. After the item was presented, a JOL was made for the target.

Judge. Each judge was also yoked to a learner. The judges were instructed about the conditions of the task for the learner, but they did not receive instructions about using this information when making JOLs for the yoked learner. For a judge, items were presented in the same order that a yoked learner had been presented the items. For both read targets and generate targets, the item was presented along with the word *rhyme* or a category label (e.g., *animal*) to the left of the item. The word *read* or *generate* was also presented at the top of the screen to indicate whether the yoked learner had read or generated the target of that item. Thus, in contrast to the learners and the observers, both the cue and the target were shown at the beginning of the presentation of each item. After an item had been presented, the judges were instructed to press the "Return" key when they wanted to continue; at that time, they made a JOL for the target.

Judgment of learning. After the presentation of each item, the participants in all three groups made a self-paced JOL. For the learners, the JOL was prompted by the previously generated (or read) target at the top of the screen along with the query, "How confident are you that in about ten minutes from now you will correctly recognize the word above?": 0 (*definitely will not recognize*) to 100 (*definitely will recognize*). For each observer and judge, the JOL was prompted by the target the yoked learner had generated (or read) along with the query, "How confident are you that in about ten minutes from now the learner will correctly recognize the word above?": 0 (*definitely will not recognize*) to 100 (*definitely will recognize*).

Recognition. Following the first phase of the experiment, all participants attempted to solve a puzzle for 10 min. The participants then completed a three-alternative forced-choice (3AFC) recognition test for the targets presented in the first phase of the experiment. The test consisted of 60 sets of three alternatives. Each set included one target and two distractors that were randomly paired with each target.

Results and Discussion

Our primary goal was to discover the nature of the inference that mediates the relationship between generation fluency and JOLs. Thus, we focus first on the relationship between the latencies to generate targets and JOLs, and we then describe the relationship between latencies to read targets and JOLs. Although the magnitude of JOLs and test performance were not the primary measures of this research, we also present them to evaluate (1) whether the generation effect on JOLs (which was first reported by Begg et al., 1991) generalizes to our task and (2) whether making JOLs diminished the generation effect on test performance (cf. Begg et al., 1991). Accuracy of JOLs at predicting subsequent recognition performance was not relevant to our goals. Also, given that recognition was the criterion task, the magnitude of JOL accuracy was expected to be attenuated across groups. In particular, individuals can respond correctly on a recognition test by merely guessing even when they had no memory of the original item. In these cases, if individuals had originally (and accurately) judged that the item would not be remembered, accuracy will be attenuated (Schwartz & Metcalfe, 1994; Thiede & Dunlosky, 1994). For these reasons, we do not present JOL accuracy within the text but merely provide a brief description of these results for interested readers in the Appendix. All effects declared reliable from initial analyses of variance (ANOVAs) have *p* less than the alpha level of .05; the alpha level for *t* tests (to follow up reliable main effects and interactions) was computed via the Bonferroni correction for each family of comparisons.

Relationship between latencies to generate targets and judgments of learning. For each target generated during study, the computer recorded the latency between the presentation of an item and the total time taken by a learner to generate the target (as operationalized above). The relationship between the fluency of generation and JOLs was operationalized as an intraindividual gamma correlation between these generation latencies and JOLs. For each observer and judge, a correlation was also computed between that observer's or judge's JOLs and the yoked learner's latencies to generate targets. The means across individuals' correlations are reported in Figure 1.

As the learners generated targets more rapidly, both the learners and the observers made greater JOLs for those targets. By contrast, the judges' JOLs were not as highly related to the learners' generation latencies. These observations were supported by a 2 (cue: rhyme vs. category) × 3 (group: learner, observer, judge) ANOVA. A reliable main effect occurred for group [$F(2,84) = 16.42$, $MS_e = 0.08$], whereas the main effect of cue and the interaction were not reliable (Fs < 0.50, $MS_e = 0.02$). The correlations for both rhyme and category targets for the learners and the observers were reliably different from zero (all ts > 7.0). For the judges, the correlation for rhyme targets was not greater than zero [$t(27) = 1.72$, standard error of the difference (SE_D) $= 0.06$], whereas the correlation for category targets was reliably greater than zero [$t(28) = 3.06$, $SE_D = 0.04$]. Because the judges did not have access to the learners' latencies to generate targets, the latter outcome suggests that some of the relationship between the learners' generation latency and JOLs for category targets may be mediated by a third factor. Most important, the correlations for both the learners and the observers were reliably greater than the corresponding correlations for the judges (ts > 5.30), whereas differences between the correlations for the learners and the

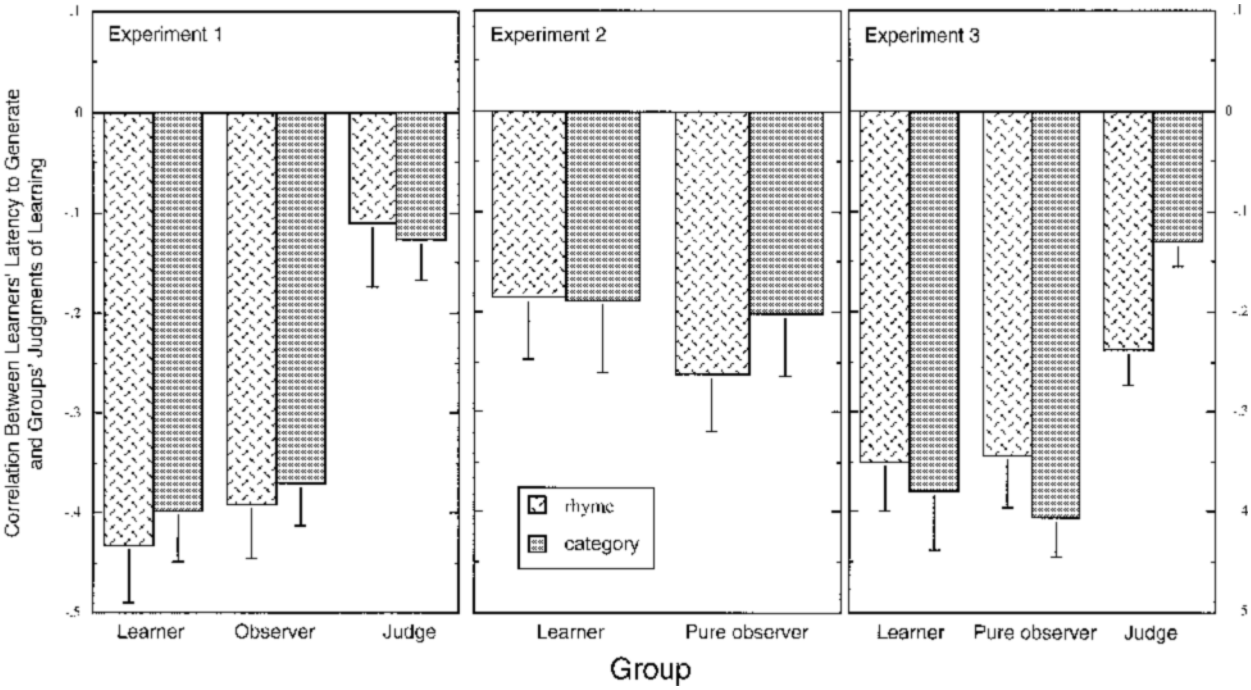

Figure 1. The relationship between JOLs and latencies to generate targets illustrated as a function of group and the kind of cue for targets (category or rhyme). For the learner group, an intraindividual gamma correlation was computed between a learner's JOLs and his or her latency to generate targets. For each participant in the observer, judge, and pure-observer groups, an intraindividual gamma correlation was computed between each individual's JOLs and his or her yoked learner's latencies to generate targets. Bars represent means across individual correlations within each group and include the corresponding standard errors.

observers were negligible [$t(58) = 0.42, SE_D = 0.07$, for category targets; and $t(58) = 0.52$, $SE_D = 0.07$, for rhyme targets]. These outcomes suggest the following conclusions: The fluency of retrieving targets at study is a cue for JOLs (cf. Benjamin et al., 1998), and the relationship between generation fluency and JOLs is partly mediated by a theory about how the fluency of generating targets relates to memory performance.

Relationship between latencies to read targets and judgments of learning. For each target read at study, the computer recorded the latency between the presentation of an item and the total time taken by a learner to read the target. The relationship between the fluency of reading and JOLs was operationalized as a gamma correlation between each learner's reading latencies and his or her JOLs. For each observer and judge, a correlation was computed between that observer's or judge's JOLs and the yoked learner's latencies to read targets.

For read targets, we did not expect to find a reliable relationship between reading fluency and JOLs because the variability in reading latencies would be minimal, which in turn would attenuate any relationship between reading fluency and JOLs. As expected, for rhyme items, the correlations were $-.12$, $-.13$, and $-.07$ for the learner, observer, and judge groups, respectively. For category items, the correlations were .01, $-.06$, and $-.05$ for the learner, observer, and judge groups, respectively. These correlations were not reliably different from 0 (all ts < 2, all SE_Ds < 0.09), and an ANOVA did not reveal reliable main effects or a reliable interaction (Fs < 2.50).

Magnitude of the judgments of learning. To examine whether a generation effect on JOLs occurred, we calculated the median across each participant's JOLs for read targets and for generate targets. For each group, a mean of the medians was then calculated. These means are reported in the top third of Table 2.

A 2 (cue) × 2 (target) × 3 (group) mixed-model ANOVA was conducted. Main effects occurred for cue [$F(1,86) = 9.75, MS_e = 193.0$] and for target [$F(1,86) = 8.53$, $MS_e = 160.0$]. Collapsed across cue, JOLs for all three groups tended to be greater for generate targets than for read targets for the learners [$t(29) = 1.45, SE_D = 2.35$, $p = .06$], for the observers [$t(29) = 2.13, SE_D = 2.94, p = .04$], and for the judges [$t(28) = 2.48$, $SE_D = 1.92$]. Although these differences were not statistically reliable given the Bonferroni correction, the trends are consistent with Begg et al.'s (1991) conclusion that people believe generating is better for memory than is reading. As is evident from inspection of Table 2, however, this main effect was qualified by several interactions: The cue × group interaction [$F(2,86) = 5.78, MS_e = 193.0$], the cue × target interaction [$F(1,86) = 5.91$, $MS_e = 144.5$], and the three-way interaction [$F(2,86) = 3.22$] were all reliable.

Table 2
Magnitude of the Judgments of Learning

	Kind of Cue			
	Rhyme		Category	
Group	Generate	Read	Generate	Read
	Experiment 1			
Learner	72 (4)	63 (4)	67 (4)	65 (4)
Observer	85 (3)	76 (3)	67 (4)	70 (3)
Judge	68 (4)	65 (4)	68 (4)	63 (4)
	Experiment 2			
Learner	67 (5)	61 (4)	68 (5)	65 (5)
Pure observer	80 (3)	78 (3)	67 (5)	66 (4)
	Experiment 3			
Learner	74 (5)	—[a]	69 (5)	—
Pure observer	82 (5)	—	71 (4)	—
Judge	81 (4)	—	70 (5)	—

Note—Cell entries are means of the participants' median judgments of learning. Standard errors of the means are in parentheses. [a]All targets were generated in Experiment 3.

If the participants had a theory about the generation effect in general, we would expect JOLs to be greater for generated targets than for read targets regardless of the kind of cue. In contrast to this possibility, follow-up analyses indicated that the kind of cue for the target moderated the generation effect on JOLs. For rhyme items, the magnitude of JOLs was reliably greater for generate targets than for read targets for the learners [$t(29) = 4.42, SE_D = 1.93$] and for the observers [$t(29) = 3.35, SE_D = 2.84$]. But, for category items, no group showed a reliable difference in the magnitude of JOLs for generate targets than for read targets. Of course, any effects involving the kind of cue may partly be due to the use of different words under the category cues versus the rhyme cues, a point we consider more fully in the General Discussion. Regardless of the specific factor contributing to these effects, however, these results suggest that a theory about the generation effect does not solely account for the generation effect on JOLs.

Test performance. Obtaining a generation effect on memory was not critical to the interpretation of the central results of this research, which pertain to JOLs and their relation to generation fluency. However, given that Begg et al. (1991) found that the generation effect on memory did not occur when participants made JOLs, we report recognition performance to assess whether this unexpected result would replicate.

To evaluate recognition performance, we computed the proportion of items that the learners answered correctly on the 3AFC recognition test. For rhyme items, mean performance was .84 for generated targets and was .80 for read targets; for category items, mean performance was .83 for generated targets and .90 for read targets (all four SEMs = .02). A 2 (cue) × 2 (target) ANOVA resulted in a main effect for cue [$F(1,29) = 4.59, MS_e = 0.01$], whereas the main effect for target was not reliable [$F(1,29) = 0.53$]. A reliable cue × target interaction [$F(1,29) = 15.8$, $MS_e < .01$] revealed that performance for category items was greater for read targets than for generated targets [$t(29) = 4.21$, $SE_D = 0.02$] but that this difference was not reliable for rhyme items [$t(29) = 1.56, SE_D = 0.03$]. These results, particularly the lack of a main effect for target, replicate those reported by Begg et al. (1991, Experiment 3). That is, the generation effect was negligible when the participants made JOLs (but see Mazzoni & Nelson, 1995), presumably because the extra processing associated with making JOLs for read targets diminishes the generation effect.[2]

EXPERIMENT 2

A primary purpose of Experiment 2 was to further investigate whether analytic or nonanalytic inferences mediate the relationship between generation fluency and JOLs. The import of Experiment 2 can be understood by considering a key outcome from Experiment 1: The correlation between the learners' latencies to generate targets and JOLs was not statistically different for the learners and for the observers (Figure 1, left panel), suggesting that the effect of generation fluency on JOLs was partly mediated by an analytic inference. The rationale was that the observers did not have access to the learners' subjective experience of generating targets at study, indicating that the use of fluency as a cue for JOLs was not mediated by the subjective experience of generating per se. One potential problem with this interpretation is that the observers may have covertly generated targets. That is, the observers may have generated targets themselves and then based their JOLs on their own subjective experience of generation. For example, as a learner attempted to generate a target (e.g., *bear* for *animal*:*horse–b*_ _ _), the yoked observer may have also attempted to generate the target. An observer's subjective experience of generating targets may have then served as a basis for a nonanalytic inference that mediated the effect of generation fluency on JOLs.

In Experiment 2, we controlled for this possibility by examining JOLs made by another group. The participants in this group observed learners generate targets but were not given an opportunity to covertly generate targets themselves. This *pure-observer* group was identical to the observer group with the exception that a yoked learner's to-be-generated target was shown in its entirety immediately as the item was presented for the generation task. "Pure" in pure observers refers to their observation of learners' generation of targets. Namely, pure observers may observe learners generate targets but are not given the opportunity to generate targets themselves. Because the entire target is presented, pure observers will not have the subjective experience of generating targets. As argued by Kelley and Jacoby (1996a), depriving participants of the subjective experience of performing a task forces them to use a theory when making the judgments—that is, to use an analytic inference about generation when making their JOLs. Accordingly, if the influence of generation fluency on JOLs is primarily mediated by a nonanalytic inference, the magnitude of the negative correlation between learners' generation latencies and JOLs will be greater for learners than for pure observers.

Method

Participants and Design. Ninety undergraduates from UNCG participated to fulfill a course requirement. The design of Experiment 2 was a 2 (cue: rhyme or category) × 2 (target: generate or read) × 2 (group: learner or pure observer) mixed-model factorial, with cue and target as within-participants factors and group as a between-participants factor.

Materials. The same items from Experiment 1 were used to construct two 60 item lists. Both lists were blocked by cue (rhyme or category). Each block contained 15 items with read targets and 15 items with generate targets, with the order of items randomized within each block. The block of category items appeared first and the block of rhyme items appeared second in List A, whereas the order of blocks was reversed in List B.

Procedure. The procedure of Experiment 2 was identical to that of Experiment 1 except that a pure-observer group was used as the comparison group. The pure-observer group was identical to the observer group described in Experiment 1 except that the yoked learner's to-be-generated target was filled in (e.g., *animal*: *horse–bear* rather than *animal*: *horse–b*_ _ _). For targets the yoked learner had generated, the target was presented with the cue, and an empty response window was presented below the item (as for the learners and the observers in Experiment 1). The pure observers were instructed that the yoked learner's response to the item would be presented in the response window. For a given item, the interval between the presentation of the item and the presentation of the learner's response to the item (which was typed into the response window by the computer) was equal to the amount of time that the learner had originally used to begin typing that target. The item was presented for the amount of time that the learner had used to type the target, as operationalized for the observers in Experiment 1. Thus, a pure observer could observe how much time the yoked learner took to generate each target but was not given an opportunity to generate the target. For targets the yoked learner had read, the item was presented for the amount of time that the learner had originally taken before proceeding to the JOL prompt. After the item was presented, a JOL was made for the target.

Results and Discussion

Relationship between latencies to generate targets and judgments of learning. As in Experiment 1, a gamma correlation was computed for each learner between generation latencies and JOLs. A gamma correlation was also computed between each pure observer's JOLs and the corresponding generation latencies of the yoked learner. The means across individuals' correlations are reported in Figure 1.

Examination of the middle panel of Figure 1 reveals that as the learners generated targets more quickly, both the learners and the pure observers made greater JOLs for those targets. A 2 (cue: rhyme or category) × 2 (group: learner or pure observer) ANOVA was conducted. Both main effects and the interaction were not reliable (Fs < 1). Although the magnitude of these correlations are lower than those reported in Experiment 1, for the learners and the pure observers the correlations were reliably different from zero (ts > 2.50, SE_Ds < 0.08), suggesting that the fluency of generation at study is a cue for people's JOLs.

Relationship between latencies to read targets and judgments of learning. A gamma correlation was computed for each learner between reading latencies and JOLs. A correlation was also computed between each pure observer's JOLs and the yoked learner's latencies to read targets. For the rhyme items (and for the category items), means across individuals' correlations were .00 (.07) for the learner group and −.04 (.04) for the pure-observer group. The main effects and interaction were not reliable (all Fs < 2.5). These correlations were not statistically different from zero either for category items [$t(28)$s < 1.16, SE_Ds < 0.07] or for rhyme items [$t(27)$s < 1.0, SE_Ds < 0.07].

Magnitudes of the judgments of learning. The means across individuals' median JOLs were computed and are reported in the middle of Table 1. A 2 (cue) × 2 (target) × 2 (group) ANOVA revealed a main effect for cue [$F(1,56) = 4.69$, $MS_e = 342.9$] and a reliable cue × group interaction [$F(1,56) = 9.64$]. All other main effects and interactions were not reliable (Fs < 2.50).

Even though the three-way interaction was not statistically reliable, the outcomes are largely consistent with those reported in Experiment 1. Namely, for rhyme items, the learner group showed an increase in JOL magnitude for targets that were generated relative to those that were read [$t(29) = 2.14$, $SE_D = 2.91$, $p = .04$ (not reliable after Bonferroni correction)], whereas this generation effect on JOLs was not evident for category items [$t(29) = 0.80$, $SE_D = 4.19$]. As mentioned in Experiment 1, however, note that the apparent interaction with the kind of cue may have resulted from the use of different words for category cues versus for rhyme cues. Finally, the pure-observer group showed a negligible generation effect on JOLs regardless of the kind of cue. Thus, in contrast to the relationships between generation fluency and JOLs that occurred regardless of the condition, the generation effect on JOLs was less consistent.

Test performance. To evaluate recognition performance, we computed the proportion of items that the learners answered correctly on the 3AFC recognition test. For rhyme items, mean performance was .88 for generated targets and .85 for read targets; for category items, mean performance was .82 for generated targets and .91 for read targets (all four SEMs ≤ .03). A 2 (cue) × 2 (target) ANOVA resulted in a main effect for target [$F(1,29) = 5.66$, $MS_e = 0.01$], whereas the main effect for cue was not reliable [$F(1,29) < 1.0$]. A reliable cue × target interaction [$F(1,29) = 13.1$, $MS_e < 0.01$] revealed that performance for items cued with categories was greater for read targets than for generated targets [$t(29) = 4.15$, $SE_D = 0.02$] but that this difference for items cued with rhymes was not reliable [$t(29) = 1.51$, $SE_D = 0.02$].

EXPERIMENT 3

Results from Experiment 2 suggest that the subjective experience of generating targets during study does not entirely mediate the relationship between generation fluency and JOLs. Evidence for this conclusion was the negligible differences in the relationships between generation latencies and JOLs for the learners and the pure observers. Because the magnitude of these relationships was relatively low, we wanted to replicate them using a list that consisted solely of generated targets, which may provide a better estimate of the relationship between generation fluency and JOLs. Because of this change in the experimental design, we also included a judge group to evaluate again the degree to which the latency–JOL relationship could be accounted for by a factor other than generation latency.

We also examined another aspect of JOLs: the latency to make a JOL. Kelley and Jacoby (1996a) hypothesized that a person makes judgments more quickly when the judgments are made using a nonanalytic inference than when they are made using an analytic inference. Accordingly, collecting latencies to make JOLs will allow us to evaluate a possible reinterpretation of our previous data. Namely, even though the relationship between generation latencies and JOLs was the same for the learners and the pure observers, the effect of generation fluency may have been mediated by a nonanalytic inference for the learners and by an analytic inference for the pure observers. If so, latencies to make JOLs will be faster for the learners than for the pure observers. By contrast, if the learners and the pure observers are using fluency as a cue for JOLs in the same manner, latencies to make JOLs will not differ for these groups. Accordingly, the key prediction here focuses on differences in latencies to make JOLs by the learner group and by the pure-observer group.

Method

Participants and Design. Ninety-three undergraduates from UNCG participated to fulfill a course requirement. The design was a 2 (cue: rhyme or category) × 2 (target: generate or read) × 3 (group: learner, pure observer, or judge) mixed-model factorial, with cue and target as within-participants factors and group as a between-participants factor. Thirty-one participants were assigned to each group.

Materials and Procedure. The same items from the previous experiments were used to construct two 60-item lists. Both lists were blocked by cue (rhyme or category). Each block contained 30 items with generate targets, and the order of items was randomized within each block. Category items were presented first and rhyme items were second in List A; order of blocks was reversed in List B.

The procedure of Experiment 3 was identical to the procedure of Experiment 2 in all respects with the exceptions that the learners generated all targets at study, a judge group was included in the experimental design, and the latency to make each JOL was recorded.

Results and Discussion

Relationship between latencies to generate targets and judgments of learning. As in the first two experiments, a gamma correlation was computed between each learner's generation latencies and their JOLs. For each pure observer and judge, a gamma correlation was computed between the yoked learner's latencies to generate targets and the pure observer's or judge's JOLs. The means across individuals' correlations are reported in the rightmost panel of Figure 1. A 2 (cue: rhyme or category) × 3 (group: learner, pure observer, or judge) ANOVA revealed a reliable main effect for group [$F(2,88) = 7.90$, $MS_e = 0.09$] and a reliable cue × group interaction [$F(2,88) = 3.56$, $MS_e = 0.04$]. The main effect of cue was not statistically reliable [$F(1,88) = 0.0$, $MS_e = 0.04$].

Follow-up analyses revealed several outcomes. The correlations were not statistically different for the learner group and the pure-observer group (ts < 1.0, SE_Ds < 0.09). Moreover, as in Experiment 1, the magnitudes of the correlations were less for the judge group than for the other two groups. The reliable cue × group interaction indicated that these differences were moderated by the kind of cue. In particular, for category items, correlations were reliably less for the judge group than for the learner group [$t(59) = 4.04$, $SE_D = 0.06$] and the pure-observer group [$t(59) = 6.05$, $SE_D = 0.04$]. For rhyme items, given that the judge group had a relatively higher correlation, the differences between correlations here only approached significance for the learner group [$t(60) = 1.77$, $SE_D = 0.06$, $p = .08$] and for the pure-observer group [$t(60) = 1.66$, $SE_D = 0.06$, p = .10]. Finally, all of the correlations were reliably different from zero, even those for the judge group (all ts > 5.0, SE_Ds < 0.07). These outcomes suggest that the relationship between generation latency and JOLs may be partly accounted for by a third factor. Whatever the nature of this factor, however, it cannot account for the entire relationship between generation latencies and JOLs for either the learners or the pure observers.

Latency of the judgments of learning. As described above, the critical comparison involving the latencies to make JOLs was between the learners and the pure observers, with latencies of the judges' JOLs being less relevant given that the bases of the judges' JOLs were presumably different from those of the other groups (see Table 1 and Figure 1). For completeness, however, we also present JOL latencies for the judge group.

For each participant, the median latency to make JOLs was computed separately for rhyme targets and for category targets. A mean of individuals' medians was then computed. For rhyme items, the mean latency to make JOLs was 3.07 sec ($SEM = 0.17$) for the learners, 2.86 sec ($SEM = 0.19$) for the pure observers, and 3.05 sec ($SEM = 0.14$) for the judges. For category items, the mean latency to make JOLs was 3.16 sec ($SEM = 0.18$) for the learners, 3.21 sec ($SEM = 0.25$) for the pure observers, and 3.43 sec ($SEM = 0.18$) for the judges. A 2 (cue) × 3 (group) ANOVA revealed a main effect of cue [$F(1,90) = 6.40$, $MS_e = 0.54$], whereas the main effect of group and the interaction were not reliable (Fs < 1.0). Most important, the mean latencies to make JOLs showed negligible differences between the learner group and the pure-observer group, again converging on the conclusion that the participants in these two groups were using the same inference about generation fluency to make their JOLs. Of course, latencies provide only indirect evidence for this conclusion, which is complicated by the outcome from the judges. Namely, JOL latencies for the judge group were not different from those for the other two groups. These data do not decisively rule out two alternative conclusions. The first is the one just offered above, which is based on the notion that latencies for the judge group are not comparable because judges presumably use different cues for making JOLs than do the other groups. Alternatively, given that the latencies were statistically equivalent for all three groups, one may conclude that because the judge group presumably is basing JOLs on different sources than are the other two groups, perhaps the other two groups are also utilizing different sources and doing so in different ways. Although we currently prefer the former alternative, further research is needed to identify the underlying sources of JOL latencies before these alternatives can be competitively evaluated.

Magnitude of the judgments of learning and test performance. Although JOL magnitude and test performance were not relevant to accomplishing the main goals of Experiment 3, we provide these values for consistency with the previous experiments. For JOL magnitude, we computed the means across individuals' median JOLs, which are reported in the bottom third of Table 2. A 2 (cue) × 3 (group) ANOVA revealed a reliable main effect for cue [$F(1,90) = 14.18$, $MS_e = 277.5$]. Thus, JOLs were greater for rhyme items than for category items, which may be due either to the kind of cue or to the fact that different words were used for the two kinds of cue. The main effect for group and the interaction were not reliable (Fs < 1).

For test performance, we computed the proportion of items that the learners answered correctly on the 3AFC recognition test. Mean performance was .86 for rhyme items and .84 for category items (SEMs = .02, $t < 1.0$, $SE_D = .02$), indicating that the kind of cue for generating targets did not reliably influence recognition performance.

GENERAL DISCUSSION

The present research was motivated by two central questions: (1) Does the fluency resulting from generating targets at study affect JOLs? (2) Are any effects of generation fluency on JOLs mediated by an analytic or a nonanalytic inference? These questions pertain specifically to targets that individuals generate at study, a task that was central to all three experiments of the present research. Evidence from these experiments provides an affirmative answer to the first question, which extends previous research on the effects of retrieval fluency on JOLs (Benjamin & Bjork, 1996; Benjamin et al., 1998; Nelson & Dunlosky, 1996). Of course, the present evidence for this conclusion was correlational and hence does not establish the causal role of fluency on JOLs. Causal conclusions have also been limited in all other research in this area, given that differences in fluency can covary either with item characteristics or with other independent variables. Although judges (Experiments 1 and 3) showed a reliable correlation between generation latency and JOLs, this correlation for judges did not account for the entire magnitude of the corresponding correlations for learners, for observers, or for pure observers (Figure 1). Such evidence indicates that although a factor other than fluency partly contributes to the relationship between generation latency and JOLs, this factor cannot account for the entire relationship and, under some conditions, cannot account for any of it. The next part of our discussion pertains to understanding the relationship between generation latency and JOLs that is not attributable to some third factor. We then end with a brief consideration of the generation effect on JOLs.

Fluency of Generation and Judgments of Learning

Concerning our second question, fluency can be used as a cue for JOLs in two ways. A conscious inference based on a theory about memory can be made that "I ought to forget" a particular target because it was not generated fluently (cf. Koriat's, 1997, instantiation of an analytic inference for JOLs, p. 366). In this case, a theory about the deficits for memory of generating targets slowly is used to make the inference for the JOL. Alternatively, a more implicit, nonanalytic inference can be made that is driven by the subjective experience a person has while generating targets. The findings we have reported provide support for the first alternative, in which people use an analytic inference to make their JOLs. Evidence for this conclusion comes from two sources: the relationship between generation latencies and JOLs, and the latencies to make JOLs.

Consider the relationship between generation latencies and JOLs for learners and pure observers. A pure observer did not have access to a learner's subjective experience of generating targets at study. Therefore, a pure observer will be forced to make an analytic inference about how a learner's speed of generation will be related to subsequent memory performance (for a similar rationale, see Kelley & Jacoby, 1996a). As presented in Figure 1, the magnitude and direction of the relations between generation latencies and JOLs were statistically equivalent for learners and pure observers, suggesting that both groups were using a similar analytic inference to make JOLs.

Analytic inferences are often characterized as more deliberate and slower than nonanalytic inferences (Jacoby & Brooks, 1984; Kelley & Jacoby, 1996a, 1996b). Accordingly, in Experiment 3, we measured the latencies to make JOLs, with the idea that perhaps the magnitude of the relationship between generation latencies and JOLs for the learners and the pure observers resulted from a qualitatively different kind of inference. In contrast to this possibility, the latencies to make JOLs for the learners and the pure observers were statistically equivalent, which is consistent with the conclusion that both groups were using the same kind of inference to make their JOLs. However, any conclusion based on the present JOL latencies must be viewed as tentative, given that it involves interpreting null results and that the JOL latencies for judges did not differ from the other groups. Accordingly, although these outcomes do not indicate that the various groups were making JOLs differently, they should be interpreted with caution.

Even with this caveat in mind, however, the overall pattern of findings are consistent with the conclusion that, when making JOLs, learners were partly using an analytic inference with respect to generation fluency. The idea is that people appear to use a theory about the perceived deficits for memory of slowly generating targets. At one level, we are arguing that JOLs are analytic with respect to generation fluency, in the sense that people explicitly analyze differences in the ease of generation across targets. On another level, however, JOLs may also be considered nonanalytic judgments, in the sense that people likely do not analyze all of the dimensions that are relevant to fluency or memory performance (Jacoby & Brooks, 1984). Moreover, generation fluency is only one dimension that can be (but is not always) diagnostic of memory performance; other dimensions are likely to be diagnostic as well. Perhaps ironically, people's analysis of generation fluency may hinder them from a more thorough analysis of other relevant dimensions that actually affect memory.

Generation Effect on Judgment-of-Learning Magnitude

The effect of generating versus reading items on JOL magnitude does not bear on our conclusions about the effect of generation fluency on JOLs. To understand why, consider that we did not even need to include read targets in our procedure to explore the effects of generation fluency on JOLs. Indeed, only generated targets were included in Experiment 3, which provided empirical support for our earlier conclusions about the relationship between generation fluency and JOLs. Although we were most concerned with such fluency effects, we briefly discuss our results concerning the generation effect on JOLs.

A generation effect on JOL magnitude occurred under some conditions (see the two leftmost columns of JOLs in Table 2). Such an outcome has led to the conclusion that people have an a priori theory about the generation effect (Begg et al., 1991). A general theory about the generation effect, however, cannot account for some of our findings. First, the generation effect on judges' JOLs was negligible, suggesting that an a priori theory about the generation effect per se plays a minimal role in making JOLs. Second, the generation effects on learners' JOLs were reliable for rhyme items but not for category items. Note that this interaction may result from using different target words in the rhyme condition and the category condition—that is, the generation effect on JOLs may be moderated more by the kinds of item used for the rhyme and category conditions than by whether generate targets had rhyme cues or category cues. Regardless of which aspect of the stimuli produced this interaction, the theoretical implication of the interaction described above is not compromised: The generation effect on JOLs (from previous research) does not likely result from a general theory about the generation effect, else it would be expected to occur under all conditions, whether for different kinds of item or different kinds of cues.

Why might this interaction occur? One alternative, albeit speculative, is based on the empirical relation between generation latencies and JOLs. As items took more time to generate, the magnitude of JOLs decreased. This outcome was evident regardless of whether targets were cued by a rhyme cue or by a category cue (Figure 1). Moreover, the participants used substantially more time to generate targets for category cues than for rhyme cues: For category cues and rhyme cues, respectively, the means across individual's median latencies were 15.6 and 10.1 in Experiment 1, 11.4 and 5.9 in Experiment 2, and 8.5 and 4.8 in Experiment 3 (all *SEM*s < 0.90). That is, targets took longer to generate when prompted with category cues, which may have resulted from differences in the difficulty of generating targets from a rhyme cue and a category cue, from differences in the targets for the two conditions, or from a combination of both. Regardless of why the participants took longer to generate targets with category cues, longer generation latencies would result in lower JOL magnitudes for these generated targets, which in turn would diminish the generation effect on JOLs for category items.

Although plausible, some outcomes from the present research are inconsistent with this explanation. Because learners generated targets more quickly for rhyme cues than for category cues, JOLs are expected to be greater for rhyme cues than for category cues. This outcome, however, was not apparent under some conditions. Moreover, in Experiment 3 (but not in Experiment 1), judges' JOLs were greater for targets generated by rhyme cues than those generated by category cues. Thus, the corresponding effects for learners' JOLs may be mediated partly by beliefs about how the two kinds of cue influence memory. In sum, although our data indicate that the generation effect on JOLs does not result from an a priori, general theory about the generation effect on recognition performance, we do not yet have a complete understanding of why generation (vs. reading) influences people's JOLs. We leave the solution of this puzzle for future research.

Summary

A major basis of people's JOLs is the fluency of retrieving targets. In the present research, we demonstrated that the latency of generating targets at study was related to JOLs, suggesting that generation fluency is also a basis of people's JOLs. By using a learner–observer–judge method, we investigated the degree to which the generation latency–JOL relationships were mediated by a nonanalytic inference or an analytic inference about generation fluency. Our findings converge on the conclusion that such fluency effects are partly mediated by an analytic inference about how fluency is related to memory.

REFERENCES

Battig, W. F., & Montague, W. F. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs*, **80**(3, Pt. 2).

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition*, **19**, 487-497.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 309-338). Hillsdale, NJ: Erlbaum.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metacognitive index. *Journal of Experimental Psychology: General*, **127**, 55-68.

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory & Language*, **34**, 383-398.

Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Psychologica*, **95**, 239-253.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 142-150.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory & Language*, **33**, 545-565.

Jacoby, L. R., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 18, pp. 1-47). New York: Academic Press.

Jameson, A., Nelson, T. O., Leonesio, R. J., & Narens, L. (1993). The feeling of another person's knowing. *Journal of Memory & Language*, **32**, 320-335.

Kelley, C. M., & Jacoby, L. L. (1996a). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory & Language*, **35**, 157-175.

Kelley, C. M., & Jacoby, L. L. (1996b). Memory attributions: Remembering, knowing, and feeling of knowing. In L. M. Reder (Ed.), *Implicit memory and metacognition* (pp. 287-307). Hillsdale, NJ: Erlbaum.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, **126**, 349-370.

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

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1263-1274.

Nelson, T. O., & Dunlosky, J. (1996, November). *Toward the theoretical mechanisms underlying immediate versus delayed judgments of learning.* Paper presented at the annual meeting of the Psychonomic Society, Chicago.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125-173). San Diego: Academic Press.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs*, **76**(1, Pt. 2).

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, **1**, 357-375.

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93-114). Cambridge, MA: MIT Press.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology*, **86**, 290-302.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1024-1037.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Columbia University, Teachers College, Bureau of Publications.

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

NOTES

1. In the literature, a variety of labels have been used to describe these groups, and sometimes the same name has been used across studies to refer to different groups. We adopted the present label names from Vesonder and Voss (1985) and Jameson et al. (1993) because they best characterize our groups in the context of a memory task in which the target group attempts to learn items.

2. Whether these conditions influenced recognition performance is less relevant for understanding their effects on JOLs (assuming that, when making JOLs, people do not have direct access to the underlying representation of the items in memory, ala Koriat, 1997). For instance, in Experiment 1, the generation effect on recognition was not apparent for either kind of cue and was even reversed for targets generated by category cues. Although potentially interesting for theories of the generation effect, such outcomes are less relevant to understanding whether people's JOLs are based on a theory that memory is generally better when targets are generated than when they are read. If people use this theory about the generation effect when making JOLs, their JOLs will show a generation effect regardless of whether this effect is manifest in memory performance. Accordingly, although we provide some speculation for why we did not obtain a generation effect on recognition performance (e.g., Begg et al., 1991), such outcomes do not directly bear on our conclusions about the underlying bases of JOLs.

APPENDIX

Our primary interest was in the bases of JOLs and not in their accuracy. However, to connect with the previous literature, we briefly describe JOL accuracy for all three experiments. In each experiment, we operationalized JOL accuracy for learners as the gamma correlation between a learner's JOLs and his/her recognition performance. For all other groups, JOL accuracy was operationalized as the gamma correlation between that participant's JOLs and his/her yoked learner's recognition performance. Table A1 presents the mean accuracy values across participants for all groups in the three experiments.

Table A1
Relative Accuracy of the Judgments of Learning

	Kind of Cue			
	Rhyme		Category	
Group	Generate	Read	Generate	Read
	Experiment 1			
Learner	.26 (.14)	.28 (.11)	.37 (.12)	.28 (.13)
Observer	.19 (.11)	−.07 (.11)	.31 (.11)	−.10 (.11)
Judge	.25 (.10)	.04 (.09)	.25 (.11)	.07 (.13)
	Experiment 2			
Learner	.10 (.14)	−.02 (.15)	.35 (.13)	.20 (.19)
Observer	.20 (.11)	.00 (.12)	.30 (.12)	−.06 (.17)
Pure observer	.37 (.11)	−.01 (.12)	.37 (.10)	.06 (.14)
	Experiment 3			
Learner	.06 (.13)	—[a]	.47 (.11)	—
Pure observer	.41 (.10)	—	.69 (.07)	—
Judge	.09 (.09)	—	.26 (.07)	—

Note—Cell entries are means of the participants' gamma correlation between judgments of learning and subsequent recall performance for the learner. Standard errors of the means are in parentheses. [a]All targets in Experiment 3 were generated.

The accuracy values from Experiment 1 were analyzed by a 2 (cue: rhyme or category) × 2 (target: generate or read) × 3 (group: learner, observer, judge) ANOVA. Reliable main effects were found for both cue [$F(1,41) = 4.17$, $MS_e = 0.24$] and target [$F(1,41) = 4.50$, $MS_e = 0.31$]. The participants in all three groups were more accurate for category items than for rhyme items and were more accurate for generate targets than for read targets. No other effects or interactions were reliable (all Fs < 2). Although the pattern of outcomes appeared similar across Experiments 1 and 2, the outcomes from Experiment 1 did not reliably replicate in Experiment 2, with a 2 (cue) × 2 (target) × 2 (group) ANOVA resulting in no reliable effects or interactions (all Fs < 3). Finally, in Experiment 3, a 2 (cue) × 3 (group) ANOVA revealed main effects of cue [$F(1,88) = 7.68$, $MS_e = 0.22$] and group [$F(2,88) = 10.27$, $MS_e = 0.21$]. The interaction was not statistically reliable [$F(2,88) = 0.21$, $MS_e = 0.22$]. In sum, accuracy tended to be lower (1) for read targets than for generate targets, (2) for rhyme items than for category items, and (3) for the judge group than for the other groups.

(Manuscript received February 23, 1999;
revision accepted for publication August 14, 2000.)