Psychological Methods
2004, Vol. 9, No. 1, 53–69

1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.1.53

A Revised Methodology for Research on Metamemory: Pre-judgment Recall And Monitoring (PRAM)

Thomas O. Nelson
University of Maryland

Louis Narens
University of California at Irvine

John Dunlosky
University of North Carolina at Greensboro

A revised methodology is described for research on metacognitive monitoring, especially judgments of learning (JOLs), to investigate psychological processing that previously has been only hypothetical and unobservable. During *data collection* a new stage of recall occurs just prior to the JOL, so that during *data analysis* the items can be partitioned into subcategories to measure the degree of JOL accuracy in ways that are more analytic than was previously possible. A weighted-average combinatorial rule allows the component measures of JOL accuracy to be combined into the usual overall measure of metacognitive accuracy. An example using the revised methodology offers a new explanation for the *delayed-JOL effect,* in which delayed JOLs are more accurate than immediate JOLs for predicting recall.

Since its inception in developmental psychology (e.g., Flavell, 1979; see also Butterfield, Nelson, & Peck, 1988), *metacognition*—which focuses on people's self-monitoring and self-control of their own cognitions—has been of widespread interest in various areas of psychology (reviewed in Nelson, 1992), particularly including cognitive psychology (e.g., reviewed in Nelson & Narens, 1990; Scheck & Nelson, 2003) but also extending to social psychology (e.g., Jost, Kruglanski, & Nelson, 1998) and clinical psychology (Nelson, Stuart, Howard, & Crowley, 1999). Research on metacognition has produced many thought-provoking findings (and unsolved questions about particular empirical phenomena of metacognition), both about what various metacognitive self-monitoring judgments are based on and about the accuracy of those judgments for predicting subsequent memory performance (e.g., Benjamin, Bjork, & Schwartz, 1998; Metcalfe, 2000; Nelson & Narens, 1990; Schwartz, 1994; Weaver & Kelemen, 1997).

The metacognitive self-monitoring judgments related to learning/memory performance were initiated in developmental psychology under the heading of *metamemory* (Flavell & Wellman, 1977) and were collected into an overall cognitive framework by Nelson and Narens (1990). One kind of self-monitoring judgment is called *judgments of learning* (JOLs), which are defined as judgments that "occur during or after acquisition and are predictions about future test performance on recently studied items" (Nelson & Narens, 1994, p. 16). In a review of the literature, Schwartz (1994, p. 360) concluded that JOLs are one of the most frequently investigated self-monitoring judgments, and many empirical investigations of JOLs have occurred not only in cognitive psychology

Thomas O. Nelson, Department of Psychology, University of Maryland; Louis Narens, Department of Psychology, University of California at Irvine; John Dunlosky, Department of Psychology, University of North Carolina at Greensboro.

This research was partially supported by Grant R305H030283 from the Cognition and Student Learning (CASL) research program at the Institute of Education Sciences of the U.S. Department of Education. We reported at the annual meeting of the Psychonomic Society in Washington, DC, November 1993, some early findings using Pre-judgment Recall And Monitoring (PRAM). We thank Katherine Rawson for transcribing the raw data for the present experiment, and we thank Chuck Weaver for helpful suggestions.

Correspondence concerning this article should be addressed to Thomas O. Nelson, Psychology Department, University of Maryland, College Park, MD 20742. E-mail: tnelson@glue.umd.edu

(e.g., Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Benjamin et al., 1998; Busey, Tunnicliff, Loftus, & Loftus, 2000; Dunlosky & Matvey, 2001; Dunlosky & Nelson, 1992, 1994, 1997; Kelemen & Weaver, 1997; Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002; Leonesio & Nelson, 1990; Matvey, Dunlosky, & Guttentag, 2001; Mazzoni & Nelson, 1995; Nelson, 1993; Weaver & Kelemen, 1997) but also in the developmental psychology of youth (e.g., Schneider, Vise, Lockl, & Nelson, 2000) and aging (e.g., Connor, Dunlosky, & Hertzog, 1997; Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002), psychopharmacology (e.g., Dunlosky et al., 1998; Nelson et al., 1998), neuropsychology (e.g., Kennedy & Yorkston, 2000; Moulin, Perfect, & Jones, 2000), and educational psychology (e.g., Kelemen, 2000; Thiede, Anderson, & Therriault, 2003; Thiede & Dunlosky, 1994). Previous research on JOLs (and metacognitive monitoring in general) has been using the same data-collection and data-analysis paradigms for the past 20 years.

The present article offers a revised methodology (called PRAM for "Pre-judgment Recall And Monitoring") that is more analytic than in previous research, especially in regard to the evaluation of JOL accuracy. The article is divided into four sections: (a) a comparison of the previous technique versus the revised technique for data collection in which the learner attempts recall immediately prior to the JOL, (b) a description of a new technique for evaluating the degree of metacognitive accuracy that stems directly out of the revised collection of data and that includes the decomposition of the usual measure of metacognitive accuracy (the Goodman–Kruskal gamma coefficient; Goodman & Kruskal, 1954) into components of metacognitive accuracy that are more analytic than was previously possible, (c) a discussion of some of the scientific ramifications of the PRAM methodology, and (d) a brief example to illustrate that the PRAM methodology is feasible and potentially important for shedding new light on empirical phenomena in the literature that have defied explanation in the past but that may have a relatively straightforward interpretation now.

Although the PRAM methodology is designed primarily for researchers to attain a more analytic evaluation of the accuracy of metacognitive judgments (e.g., by isolating the locus of effects of any intervention, such as an instruction to use a particular learning strategy or such as the delay of JOLs in the empirical example described below), PRAM may directly or indirectly have implications for any applied situation in which a stage of recall can be inserted immediately prior to a metacognitive judgment about what the person claims to know. Three examples (see below for further elaboration) are as follows: (a) With a better understanding of how various interventions affect the accuracy of metacognitive judgments, educators are likely to make better decisions about which interventions to use in particular educational settings; (b) with a better understanding of what a potential witness can versus cannot recall prior to making a metacognitive judgment about his or her own eyewitness identification accuracy, law enforcement officials might make better decisions about the usefulness of potential eyewitnesses at a crime scene; and (c) students' knowledge gained from their attempts at recall immediately prior to their JOLs may facilitate more well-informed decisions about the allocation of additional study time in naturalistic learning situations such as devoting more study time to items they fail to recall and less study time to items they do recall.

Previous Techniques for Data Collection and Data Analysis

Previous Technique for Data Collection

A schematic overview of the technique for data collection for JOLs that has been used previously (e.g., from the time of Arbuckle & Cuddy, 1969, to the present)[1] is shown in the top panel of Figure 1. The person first studied the to-be-learned items (e.g., cue–target pairs such as *OCEAN–TREE*) with instructions that they should be learned so that subsequently when prompted with the cue (e.g., "OCEAN–?"), the person would recall the target (e.g., "TREE"). Then time elapsed between the termination of studying a given item and the onset of the JOL for that item. This interval could be extremely brief (e.g., as in an *immediate JOL* that occurs as close in time as possible to the offset of the studied item) or could be delayed for a more lengthy amount of time (as in a *delayed JOL*) filled with other activity and/or other to-be-learned items. Then the JOL occurred, prompted by a cue that usually consisted of only the cue from the studied

[1] There have been minor variations, such as whether the cue for JOL should be the stimulus alone or the stimulus–response pair (e.g., see Dunlosky & Nelson, 1992, for an example of the impact of this variation). However, such minor variations can easily be incorporated into the PRAM methodology.

PREVIOUS METHODOLOGY FOR DATA COLLECTION:

	Study	JOL	Final test
Example:	OCEAN – TREE	OCEAN - ? (0%…100%)	OCEAN - ?
Activity:	Study the item	Estimate the likelihood of recall 10 min from now	Recall the target (i.e., recall TREE)

PRAM METHODOLOGY FOR DATA COLLECTION:

	Study	Pre-JOL recall	JOL	Final test
Example:	OCEAN – TREE	OCEAN - ?	OCEAN - ? (0%…100%)	OCEAN - ?
Activity:	Study the item	Recall the target (i.e., recall TREE)	Estimate the likelihood of recall 10 min from now	Recall the target (i.e., recall TREE)

Figure 1. Main stages in data collection for the previous methodology (top panel) and for the PRAM (Pre-judgment Recall And Monitoring) methodology (bottom panel). JOL = judgment of learning.

item (e.g., "OCEAN–?"). The person generated a JOL by choosing the predicted likelihood (e.g., on a Likert-type rating scale or on a scale ranging from 0% to 100% in steps of 20%) of remembering the item on the eventual criterion test (e.g., 10 min later). Then other items were studied, and JOLs were made for them, until every item had been studied and had received a JOL. Finally, following an interval of perhaps 10 min (filled with other items) from the time of studying a given item, the person received the eventual memory test on that item. The memory test was self-paced, usually asking the person to recall the target when prompted by the cue (e.g., the person attempted to recall "TREE" when prompted by "OCEAN–?"), although sometimes the test was one of forced-choice associative recognition (e.g., Dunlosky & Nelson, 1997; Thiede & Dunlosky, 1994).

An example of illustrative data from this technique of data collection appears in Table 1, which shows the observable outcomes for eight items. The first column indicates each of the items (where a given item might be, say, "OCEAN–TREE" for noun–noun paired associates or "ARDHI–SOIL" for foreign-language translation equivalents). The second column contains only hypothetical speculations when using the previous technique (but becomes observable when using the PRAM methodology); this column can be ignored for now and is discussed later in the present article. The third column shows the JOL rating for each item. The fourth column shows the outcome on the criterion test for each item. The first note at the bottom of the table shows the coding of the data for purposes of data analysis, as discussed next.

Previous Technique for Data Analysis to Assess the Accuracy of JOLs

Using the data generated by the previous technique, the investigator assessed the accuracy[2] of the JOLs at

[2] JOL accuracy in terms of comparing one item relative to another item (e.g., if item *i* receives a higher JOL rating than item *j,* then the likelihood of eventual memory performance should be greater for item *i* than for item *j*) is referred to as *relative* accuracy of prediction. Another kind of JOL accu-

Table 1
Illustrative Data Obtained From the Previous Technique and the Pre-judgment Recall And Monitoring (PRAM) Technique of Data Collection

Item	Target recalled during JOL rating?[a]	JOL rating (%)	Target recalled on criterion test?
F	Yes	100	Yes
G	Yes	80	Yes
H	Yes	60	No
I	Yes	40	Yes
J	No	20	No
K	No	20	No
L	No	0	No
M	No	0	No

[a] Entries in this column are not observable using the previous methodology (i.e., are only hypothetical within that methodology), but they become observable using the PRAM methodology.

Note. For the previous methodology, which evaluates only the non-tied dyads (e.g., the dyad {F, G} is excluded because it is tied on criterion-test performance insofar as both Item F and Item G were recalled), a concordance occurs whenever item *r* is greater than item *w* on both the JOL rating and criterion-test performance (e.g., Item G has a JOL rating of 80 whereas Item H has a JOL rating of 60 and Item G is recalled whereas Item H is not), and a discordance occurs whenever the JOL rating is higher for item *r* than for item *w* and criterion-test performance is lower for item *r* than for item *w* (e.g., Item H has a higher JOL rating than Item I and recall is lower for Item H than for Item I). Accordingly, the above data yield the concordances {F, H}, {F, J}, {F, K}, {F, L}, {F, M}, {G, H}, {G, J}, {G, K}, {G, L}, {G, M}, {I, J}, {I, K}, {I, L}, and {I, M} and the discordance {H, I}. Thus 14 dyads are concordances and 1 dyad is a discordance, such that $C = 14$ and $D = 1$, and therefore $\gamma = (14 - 1)/(14 + 1) = 13/15 = .87$ (see Equation 1).

For the PRAM methodology, which like the previous methodology evaluates only the non-tied dyads, the entries in the second column become observable (with no changes in the remaining entries), and then the dyads that are non-tied in JOL rating and non-tied in criterion-test performance are {F, H}, {G, H}, and {H, I} for the RR dyads and whereas the remaining 12 dyads of those listed earlier in this note are RN dyads. Therefore by Equation 3, $p_{RR} = 3/15 = .20$ and $p_{RN} = 12/15 = .80$, and $\gamma_{RR} = (2 - 1)/(2 + 1) = .33$ and $\gamma_{RN} = (12 - 0)/(12 + 0) = 1.00$, such that $\gamma.. = (.20)(.33) + (.80)(1.00) = .87$, which is the same as the overall γ of .87 above. The γ_{NN} is indeterminate because all NN dyads contain ties on criterion-test performance, which is not unusual after only one study trial (e.g., Nelson & Narens, 1990). JOL = judgment of learning.

predicting eventual memory performance under various conditions manipulated in the experiment. This assessment typically occurred by computing the degree of relatedness between (a) the magnitudes of JOLs and (b) the outcomes on eventual memory performance.

The particular measure of accuracy used in almost all articles published since the 1980s is the Goodman–Kruskal gamma correlation, designated as γ. γ is a nonparametric correlation coefficient that has several advantages including the following: (a) The γ statistic does not assume interval scales on either of the variables being correlated (which is particularly useful for metacognitive ratings, e.g., because Likert-type scales should not be assumed to be on an interval scale; Surber, 1984). (b) The γ statistic is appropriate both for one-to-one relationships and for many-to-one relationships (Freeman, 1986), which is especially relevant to metacognitive ratings because tied ratings are forced to occur by the experimental procedures (e.g., whenever a *j*-place rating scale is used to rate *k* items in which $j < k$, such as the example in Table 1 in which a six-place rating scale is used to rate eight items and therefore at least some tied ratings necessarily have to occur—for elaboration, see Gonzalez & Nelson, 1996). (c) The γ statistic is unaffected by ties either in the ratings or in the eventual memory performance (e.g., two items that are recalled are treated as tied in memory performance) insofar as the computation of γ *excludes dyads containing ties* (for related measures that include dyads tied on the predictor variable and/or criterion variable, and for the rationale for choosing γ over those measures, see Gonzalez & Nelson, 1996). (d) The γ statistic is appropriate either for metacognitive rankings derived from paired comparisons (e.g., as elaborated in Nelson & Narens, 1980) or for metacognitive ratings, which are more efficient to obtain than are metacognitive rankings (Nelson, 1984). (e) The γ statistic does not make an arbitrarily strong assumption of an underlying *linear* relationship between the metacognitive ratings and the memory performance being rated but instead assumes only a monotonic relationship; correspondingly, the interpretation of γ is not in terms of the

racy, referred to as *absolute* accuracy of prediction, entails comparing the magnitude of the JOLs with the likelihood of eventual memory performance (e.g., the subset of items receiving a JOL of "80% predicted likelihood" would be 80% correct on the eventual memory test if the JOLs for those items had perfect absolute accuracy). Relative accuracy is sometimes called *resolution* in the literature on judgment and decision making, and absolute accuracy is sometimes called *calibration* (e.g., see Lichtenstein & Fischhoff, 1977). Because Lichtenstein and Fischhoff concluded that resolution in comparison with calibration "is a more fundamental aspect of probabilistic functioning" (p. 181) and because almost all published research on JOLs has focused on relative accuracy, we focus on relative accuracy in the present article.

degree of linear relationship but rather is a probabilistic interpretation in terms of the degree of monotonic relationship.[3] (f) The γ statistic is not marginsensitive and therefore can be used appropriately when the level of eventual test performance varies in ways that are unknown by the learner at the time of the metacognitive judgments (Nelson, 1984). (For further elaboration and other advantages of γ as a measure of the degree of metacognitive accuracy, see Gonzalez & Nelson, 1996; Nelson, 1984, 1996.)

The definition of γ is as follows:

$$\gamma = (C - D)/(C + D), \quad (1)$$

where C = the number of concordant *dyads* of items, and D = the number of discordant dyads, where the dyads can be either obtained directly from the learner (e.g., by paired comparisons) or derived by the experimenter from the learner's ratings (as in the example in Table 1). Concordant dyads are dyads of items for which the person predicted greater eventual memory performance on item *i* than on item *j* (e.g., assigning a higher rating or ranking to item *i* than to item *j*) and the eventual memory performance was greater on item *i* than on item *j* (e.g., correct recall on item *i* and incorrect recall on item *j*). Discordant dyads are dyads of items for which the person predicted greater eventual memory performance on item *i* than on item *j* and the eventual memory performance was worse on item *i* than on item *j*. Dyads containing ties on either the predicted memory performance or the eventual memory performance are ignored because they are regarded as being noninformative (for the rationale, see Gonzalez & Nelson, 1996; Nelson, 1984), and accordingly those dyads are ignored throughout the remainder of the present article.

Equation 1 yields a measure of the degree of JOL accuracy that ranges from −1.0 (complete negative accuracy) to 0 (nil accuracy) to +1.0 (perfect accuracy). For the illustrative data shown in Table 1, C = 14 and D = 1 (as derived in the first paragraph in the note in Table 1), and therefore $\gamma = .87$ by Equation 1.

Notice (e.g., in the second column in Table 1) that the dyads that enter into the computation of γ for JOL accuracy can be composed of some items that were retrievable at the time of the JOL and other items that were not retrievable at the time of the JOL. However—and this is of crucial importance—the previous technique for data collection does not allow the investigator to know whether a given item was or was not retrievable when it was receiving a JOL, and investigators (e.g., Dunlosky & Nelson, 1992; Kelemen & Weaver, 1997; Nelson & Dunlosky, 1991, 1992; Spellman & Bjork, 1992) have frequently speculated about the presence/absence of retrieval (at the time of the JOL) as a hypothetical event and have attempted to explain the obtained degree of JOL accuracy by making then-untestable assumptions about hypothetical retrieval at the time of the JOL. Now, however, the investigator will no longer have to make assumptions about correct–incorrect retrieval at the time of the JOL that are untestable, because they become observable (and therefore testable) when the PRAM methodology is used, as discussed next.

PRAM Methodology: Revised Techniques for Data Collection and Data Analysis

Revised Technique for Data Collection

The PRAM technique for data collection is schematized in the lower panel of Figure 1, which may be contrasted with the upper panel. The critical revision is the insertion of a new stage—namely, pre-judgment recall—that occurs immediately prior to the JOL judgment. This pre-judgment recall provides an assessment of whether or not the item can be recalled at the time of the JOL. Although this may initially appear to be only a modest change, it has major repercussions for data analysis and allows the assessment of JOL accuracy to be more analytic, in ways that were impossible previously.

---

[3] The relationship is between γ and the probability *P,* where *P* is the probability that item *i* is greater than item *j* in magnitude of JOLs, given that item *i* is greater than item *j* on the criterion test of memory (e.g., recall), with no ties on either variable. The exact relationship (Nelson, 1984, pp. 129–130, Proof 3) is that $P = .5 + .5\gamma$, and because this relationship is linear, the conclusions of inferential statistical tests such as analyses of variance that are conducted on γ will generalize meaningfully to the probabilistic interpretation *P*. (Extensions to situations incorporating ties on either/both variables are available in Gonzalez & Nelson, 1996.) Such generalization would not be the case for other measures such as the Pearson's product–moment correlation, whose interpretation is in terms of the proportion of variance accounted for, as based on the square of the Pearson correlation; other problems making the Pearson correlation unsuitable as a measure of the degree of metacognitive accuracy have been elaborated in Nelson (1984) and Gonzalez and Nelson (1996).

Revised Technique for Data Analysis

The computations for data analysis using the PRAM methodology build on those from previous research. However, the overall measure of JOL accuracy is decomposed into particular components of JOL accuracy that previously were not observable and not computable, and those components are of great theoretical interest.

When the data are collected using the PRAM methodology, the items can be partitioned into items that were recalled at the time of the JOL versus items that were not recalled at the time of the JOL (e.g., see the second column in Table 1), and then γ can be computed separately for those partitions. Because the assessment of JOL accuracy (cf. Equation 1) is based on *dyads* of items, the partition of items in terms of the outcome on pre-judgment recall yields three distinguishable subcategories that are mutually exclusive and exhaustive (see the second paragraph of the note in Table 1 for an example): (a) dyads of items in which both of the items in a given dyad were recalled during pre-judgment recall (which are designated as RR dyads); (b) dyads of items in which neither of the items was recalled during pre-judgment recall, that is, in which both items were nonrecalled (which are designated as NN dyads); and (c) dyads of items in which one of the items was recalled whereas the other item was not recalled during pre-judgment recall (which are designated as RN dyads). Then the corresponding computation of γ occurs the same way as in previous research, except that a separate γ is computed for each of those three subcategories of dyads. The γ for only those dyads in which both of the two items in a given dyad were recalled during pre-judgment recall is designated γ_{RR}; the γ for only those dyads in which both of the two items in a given dyad were nonrecalled during pre-judgment recall is designated γ_{NN}; and the γ for only those dyads in which one of the two items in a given dyad was recalled during pre-judgment recall whereas the other item was nonrecalled is designated γ_{RN}.

This partitioning allows the aggregation of dyads for a given γ to be more homogeneous than in previous research in which the aforementioned three subcategories were combined into one heterogeneous group of dyads. Therefore, in previous research the observed value of γ could have been dominated greatly by one or another of those kinds of dyads, without the investigator's knowing the relative dominance of each subcategory of dyads.

Combining Each of the Three Component γs Into the Overall γ

Because the three subcategories of dyads are mutually exclusive and mutually exhaustive (insofar as every dyad goes into exactly one of the three subcategories), and because γ is computed on all of those dyads (without partitioning them into subcategories), it is possible to determine the overall γ (hereafter designated $\gamma..$) by combining the three component γs. The combinatorial rule is as follows:

$$\gamma.. = \frac{(f_{RR} \cdot \gamma_{RR}) + (f_{NN} \cdot \gamma_{NN}) + (f_{RN} \cdot \gamma_{RN})}{f_{RR} + f_{NN} + f_{RN}}, \quad (2)$$

where f is the frequency of occurrence of the subcategory of dyads that is indicated by the two letters following the f (e.g., f_{RN} is the frequency of dyads in which each dyad is composed of one item recalled during pre-judgment recall and one item nonrecalled during pre-judgment recall). The $\gamma..$ is the overall gamma resulting from a weighted combination of the three component gammas γ_{RR}, γ_{NN}, and γ_{RN}, where the weight for each component gamma is the frequency of all dyads that are in each subcategory divided by the sum of the dyads in all three subcategories (e.g., the weight for γ_{RR} is f_{RR} divided by the denominator $f_{RR} + f_{NN} + f_{RN}$).

Equation 2 can be simplified by expressing the weight for each component gamma as a proportion, by first dividing each frequency f in the numerator of Equation 2 by the sum of the three frequencies in the denominator, so as to express the combinatorial rule as follows:

$$\gamma.. = (p_{RR} \cdot \gamma_{RR}) + (p_{NN} \cdot \gamma_{NN}) + (p_{RN} \cdot \gamma_{RN}), \quad (3)$$

where each p is the proportion of all the dyads in the denominator of Equation 2 that belong to one of the three subcategories designated by the two letters following the p—for example, $p_{RR} = f_{RR}/(f_{RR} + f_{NN} + f_{RN})$—and where $p_{RR} + p_{NN} + p_{RN} = 1$. Equation 3 is what we use henceforth to represent the decomposition of the overall gamma $\gamma..$ into its component γs.

Notice that the $\gamma..$ from Equation 3 is arithmetically identical to the overall γ computed in Equation 1 (where all of the dyads are aggregated without regard for whether the two items in a given dyad were recalled or nonrecalled at the time of the JOL). Therefore, Equation 3 allows the investigator to determine γ directly from the three component gammas (as long as the frequency of dyads composing each component

gamma is recorded and entered into Equation 3), and hence the $\gamma..$ from Equation 3 can be compared directly with the values of γ in the previous literature.

Put differently, γ in previous research was composed (per Equation 3) of the weighted sum of the three component gammas, but without the investigators being able to identify the values of the component gammas, because the data-collection technique in previous research did not allow for the determination of whether a given dyad was composed of items that were recalled or nonrecalled at the time of the JOL. Instead, all dyads were aggregated even though they were not homogeneous. This is illustrated in Table 1, where the second paragraph of the note in the table shows both the partitioning of the dyads into subcategories and the decomposition of the overall gamma into its component gammas.

Graphical Display of the Components That Combine to Form $\gamma..$

In the previous methodology, investigators often represented in a bar graph the values of $\gamma..$ for various conditions. In the PRAM methodology, a new kind of graph can show how each component in Equation 3 contributes to overall JOL accuracy (illustrated in Figure 2).

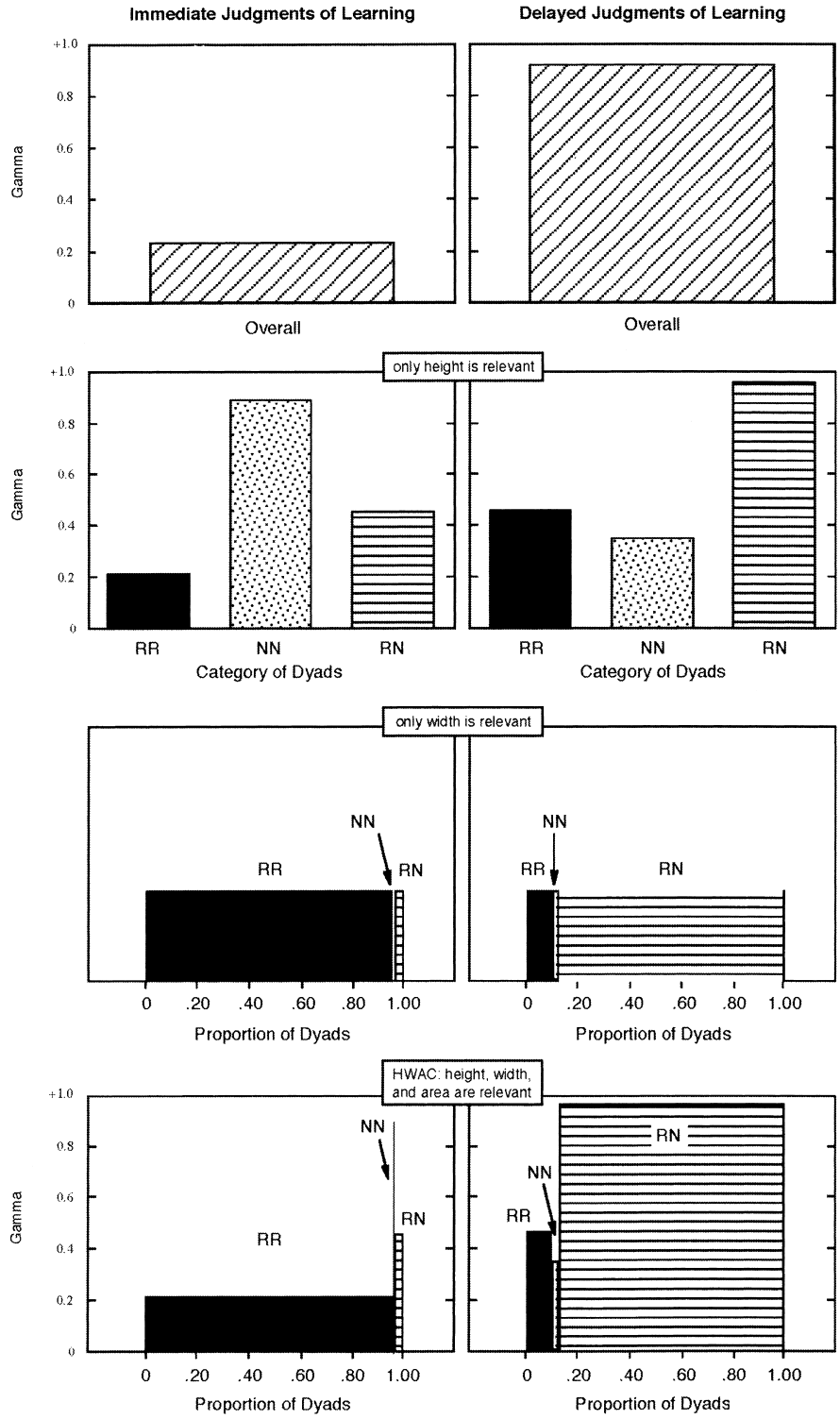
Although the three component gammas in the right side of Equation 3 (i.e., γ_{RR}, γ_{NN}, and γ_{RN}) could be presented as three separate bars for a given condition (e.g., see the second panel of Figure 2), and although the three proportions of dyads (i.e., p_{RR}, p_{NN}, and p_{RN}) could be presented in a horizontally stacked bar graph (e.g., see the third panel of Figure 2), the informative aspects of both of those graphs can be shown simultaneously in a new kind of graph that we refer to as a *height–width–area comparison* (HWAC) graph in which the height, the width, and the area of each bar are meaningful in the following way.[4] Because the decomposition of $\gamma..$ in Equation 3 is the sum of three products (e.g., one product is $p_{RR} \cdot \gamma_{RR}$), and because a product is analogous to the measurement of area (where a given area is a product of height multiplied by width), the amount of accuracy arising from a given subcategory of dyads can be represented graphically by the area of a bar in a bar graph in which the height of the bar corresponds to the value of the component γ and the width corresponds to the proportion of that kind of dyad (e.g., see the fourth panel of Figure 2). Thus in a HWAC graph, (a) each bar can vary nonarbitrarily both in height and in width, (b) the multiplicative product of the height and width of a given bar is indicated by its area, and especially important, (c) the areas of the bars can be compared meaningfully to indicate the relative contribution to overall accuracy. An HWAC graph allows the investigator to show at a glance the locus of effect (of a given independent variable) on the components underlying overall accuracy, as is illustrated below in the example using the PRAM methodology to investigate *the delayed-JOL effect.*

Several Important Ramifications of the PRAM Methodology

By computing a separate gamma for each of the aforementioned three subcategories of dyads, the assessment of the effect(s) of any independent variable of interest is more analytic, because the investigator can isolate (and observe) the separate effects of a given independent variable on γ_{RR}, γ_{NN}, and γ_{RN}. Thus, the PRAM methodology allows for better isolation of a given independent variable's locus (or loci) of effect on JOL accuracy. The independent variable might affect the participant's discrimination within the set of recalled items, discrimination within the set of nonrecalled items, discrimination between a recalled versus nonrecalled item, or some combination of those effects.

Also, and perhaps contrary to intuition, the PRAM methodology can show how a given independent variable might affect not the accuracy of any of those kinds of discrimination but instead might affect only

[4] Related to the HWAC graph is the mosaic graph proposed by Hartigan and Kleiner (1981) and popularized by Friendly (e.g., Friendly, 1994). Although the mosaic graph represents the counts in a contingency table (e.g., "brown hair and blue eyes") in a two-dimensional bar graph, the product of those counts (cf. the area inside a given bar) typically is not interpretable as a single entity (i.e., is not usually a scalar) and is not meaningful in the measurement sense. By contrast, in our HWAC graph the area does have meaning (viz., as the amount of the overall accuracy arising from a particular component of accuracy—i.e., the magnitude of that component multiplied by the weight of its role in overall accuracy, as described in the text). Also, the sum of the three areas (i.e., one area for each of the three components RR, NN, and RN) is always equal to the value of the $\gamma..$ for overall accuracy when the three component γs are non-negative, for example, the sum of the areas in the HWAC graph in the bottom panel of Figure 2 is equal to the area in the top panel of Figure 2.


Immediate Judgments of Learning
Delayed Judgments of Learning
Gamma
+1.0
0.8
0.6
0.4
0.2
0
Overall
Overall
only height is relevant
RR
NN
RN
Category of Dyads
only width is relevant
Proportion of Dyads
0
.20
.40
.60
.80
1.00
HWAC: height, width, and area are relevant

the relative frequency with which one or another of those kinds of discrimination occurs. For instance, even if both groups in an experiment have $\gamma_{RR} = .4$, $\gamma_{NN} = .3$, and $\gamma_{RN} = .8$, it is possible for the overall $\gamma..$ to be substantially higher in Group A than in Group B just because the proportion of RN dyads (i.e., p_{RN}) is higher in Group A than in Group B—per Equation 3 (also see Figure 2). Perhaps even more counterintuitive—in an instantiation of Simpson's paradox—is the possibility that all three gamma components are higher in Group D than in Group E, but the overall $\gamma..$ is higher in Group E than in Group D.[5] This could occur, for instance, if p_{RN} was markedly higher in Group E than in Group D.

Finally, by pinpointing the locus (or loci) of effect of a given independent variable, the PRAM methodology allows for more precise theorizing. The decomposition of $\gamma..$ into γ_{RR}, γ_{NN}, and γ_{RN} (and the proportions of the corresponding dyads) in Equation 3 is particularly important when separate psychological processes (whose separate effects were hitherto not directly measurable) are hypothesized to underlie judgments about items that are recalled versus nonrecalled at the time of the judgments (e.g., Kelemen & Weaver, 1997; Nelson & Dunlosky, 1991, 1992; Spellman & Bjork, 1992). An example of the above aspects of the PRAM methodology is illustrated next.

Illustration Using the PRAM Methodology to Investigate the Delayed-JOL Effect

The delayed-JOL effect (Nelson & Dunlosky, 1991) is the empirical finding that the accuracy of JOLs made after a relatively brief delay following study is substantially greater than the accuracy of JOLs made immediately after study. This finding has been replicated often (e.g., Connor et al., 1997; Dunlosky & Nelson, 1992, 1994, 1997; Kelemen & Weaver, 1997; Nelson & Dunlosky, 1991; Thiede & Dunlosky, 1994; Weaver & Kelemen, 1997), and other kinds of delays (e.g., in generating keywords either immediately or sometime after reading text as encountered in educational situations) can also facilitate the accuracy of metacognitive monitoring as well as the regulation of study and eventual test performance (Thiede et al., 2003). The delayed-JOL effect is potent and robust. For instance, in the experiment by Nelson and Dunlosky (1991), the mean $\gamma..$ was only .38 for immediate JOLs but was .90 for delayed JOLs, and every one of the 30 participants had greater accuracy for delayed JOLs than for immediate JOLs. However, in spite of the potency and robustness of the finding, no consensus has emerged about the psychological mechanisms that produce it. Several researchers have commented on this problem for current theory. For instance, Weaver and Kelemen remarked, "Though the underlying cause of the delayed-JOL effect is still a matter of some debate . . . there can be no doubt that delaying judgments improves JOL accuracy" (p. 318), and Benjamin and Bjork (1996) described the delayed-JOL effect as "a phenomenon that has been somewhat of an enigma in the JOL literature" (p. 321).

Our goal here is not to review the literature about

[5] As one example, suppose that the obtained values for Groups D and E are .4 and .3 for γ_{RR}, .3 and .2 for γ_{NN}, and .9 and .8 for γ_{RN}, respectively (i.e., the values are greater for Group D than Group E on all three component γs); however, if the proportions of non-tied dyads for Groups D and E are .8 and .1 for p_{RR}, .1 and .1 for p_{NN}, and .1 and .8 for p_{RN}, respectively, then the overall $\gamma..$ is (by Equation 3) .44 for Group D and .69 for Group E (i.e., greater for Group E than Group D). *Simpson's paradox,* wherein a different pattern across conditions occurs in the subcomponents than in the aggregation of the subcomponents, is elaborated by Simpson (1951) and Samuels (1993); see Hintzman (1980) for examples from research on memory.

Figure 2 (opposite). A graphical representation of the primary data from the PRAM methodology for the experiment described in the text, in which the two conditions were immediate judgments of learning (JOLs) versus delayed JOLs. The top panel shows the mean overall JOL accuracy ($\gamma..$). The second panel shows the mean for each component gamma (γ_{RR}, γ_{NN}, and γ_{RN}). The third panel shows the mean for each proportion of dyads (p_{RR}, p_{NN}, and p_{RN}). The bottom panel is a height–width–area comparison (HWAC) graph of the mean gammas and the mean proportions of dyads (note that the height, width, and area of a given bar are each relevant). Also, the scaling of the figure is such that the area is the same in the top panel as in the bottom panel (i.e., the total area of the three bars in the bottom panel is the same as the area of the one bar in the top panel, consistent with Equation 3 and with the values reported in Equations 4 and 5 for immediate JOLs and delayed JOLs, respectively). RR = dyads in which both items were recalled during prejudgment recall; NN = dyads in which both items were nonrecalled during prejudgment recall; RN = dyads in which one item was recalled and the other item was nonrecalled during prejudgment recall.

those mechanisms (they were reviewed in Schwartz, 1994) but rather to emphasize that most of the proposed mechanisms assume *hypothetical* aspects of retrieval—either successful or unsuccessful retrieval—occur at the time of the JOL. For instance, Nelson and Dunlosky (1991, 1992) hypothesized that retrieval from short-term memory occurs with a high probability at the time of immediate JOLs but not at the time of delayed JOLs; by contrast, Spellman and Bjork (1992) hypothesized that successful retrieval at the time of the JOLs gives a beneficial potentiating effect for future performance after delayed JOLs but not after immediate JOLs. Unfortunately, with the previous methodology, direct empirical tests of those hypotheses were impossible because retrieval at the time of the JOL was only speculative and was not assessed.

The PRAM methodology allows for an assessment of retrieval at the time of JOLs and thereby offers the possibility of making observable some of the hypothetical mechanisms that previously were only speculative. Next we report data from the PRAM methodology that illustrate the assessment of retrieval during JOLs.[6]

Method

Items and apparatus. The items were 126 pairs of unrelated, concrete nouns (e.g., OCEAN–TREE) that were displayed on Apple computers, which also recorded the participants' responses.

Design and participants. The interval between study and the JOL (immediate JOL vs. delayed JOL) was a within-subject manipulation (described below), and all JOLs were preceded by pre-JOL recall. The 45 university undergraduates received extra course credit for participating. Participants were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association [APA], 1992).

Procedure. Prior to study, the participants were instructed to learn the items and were informed about the recall tests and JOLs. During study, the items were presented in random order for 10 s per item. The first 6 items served as a primacy buffer and did not receive any JOLs or recall trials. The remaining 120 items composed four blocks of 30 items per block. Each item was randomly slated for immediate or delayed JOL, with the only two restrictions being that (a) 15 items in each block were slated for each kind of JOL and (b) the final 5 items studied in each block were slated for immediate JOLs, so that at least 5 items intervened between the study and JOL of every item having delayed JOLs. The randomization occurred anew for each participant.

Pre-JOL recall consisted of self-paced paired-associate recall that immediately preceded every JOL. For example, sometime after studying "OCEAN–TREE," the participant saw "OCEAN–?" and attempted to recall "TREE" by typing a guess (or typing *next* if no guess).

The prompt for JOLs was the stimulus alone (e.g., "OCEAN–?") and the query, "How confident are you that in about ten minutes from now you will be able to recall the second word when prompted with the first word? (0 = definitely won't recall, 20 = 20% sure, 40 . . . , 60 . . . , 80 . . . , and 100 = definitely will recall)." JOLs were self-paced. Each immediate JOL (and its pre-JOL recall) occurred immediately after the offset of the item, and each of the delayed JOLs (and its pre-JOL recall) of a given block occurred in random order after the final immediate JOL of that block.

The order of items was randomized anew for final recall, and the method of testing was identical to that for pre-JOL recall. To minimize errors due to incorrect spelling, we scored a response as correct if the first 3 letters were correct or if it was obviously misspelled, such as responses that included typing errors (e.g., *sstar* for *star*), excluded one letter (e.g., *boquet* for *bouquet*), or transposed two letters (e.g., *brarel* for *barrel*). To assess the reliability of scoring a response as correct–incorrect, two raters scored 9 participants' recall data consisting of 1,080 recall trials. Most of the incorrect responses were errors of omissions, which gave trivially perfect inter-rater scoring reliability (as did the correct responses that were spelled perfectly), but 64 responses were not identical to the objectively correct response; the two raters' scoring was in 100% agreement concerning whether each of those 64 was incorrect or was a correct response that had been misspelled.

Results and Discussion

Here we report only some basic statistics that correspond mostly to the components of Equation 3 and

[6] Here we present only the highlights. Elsewhere a more thorough report will be published about the details of the method that would be of interest mostly to specialists. Our present goal is primarily to illustrate how the PRAM methodology can make observable some hitherto untestable and speculative theoretical mechanisms.

that illustrate a few of the many findings that can be obtained from the PRAM methodology.[7] For all differences reported as significant, $p < .05$. Also, to be in accord with the recommendation incorporated into the APA publication manual (APA, 2001), we report a measure of effect size (viz., Cohen's *d* corrected for the correlation between dependent measures being compared in a repeated-measures design, as given in Equation 3 from Dunlap, Cortina, Vaslow, & Burke, 1996) for each significant difference, as well as an estimated effect size (referred to as ES_{est}) required to obtain a significant difference (with power = .80) using parametric statistics (estimated from tables in Bausell & Li, 2002) for nonsignificant differences.

Recall performance. The mean proportion of correct pre-JOL recall was significantly greater on items having immediate JOLs ($M = .97$) than on items having delayed JOLs ($M = .53$), $t(44) = 13.2$, $d = 1.56$. The mean proportion of correct final recall was significantly greater after delayed JOLs ($M = .49$) than after immediate JOLs ($M = .39$), $t(44) = 6.86$, $d = .40$.

Relationship between judgments of learning and final recall. Equation 3 decomposes the overall accuracy of JOLs (i.e., $\gamma..$) into the components composing that overall accuracy. The mean performance on each of those components is shown in Figure 2.

The top panel of Figure 2 shows the mean $\gamma..$ for immediate versus delayed JOLs. The mean $\gamma..$ for delayed JOLs was significantly greater than the mean $\gamma..$ for immediate JOLs, $t(43) = 11.95$, $d = 2.44$, replicating the usual delayed-JOL effect on JOL accuracy. However, in the previous methodology, this molar effect is the only aspect of JOL accuracy that is examined. Next consider the additional effects that are available from the PRAM methodology.

The first finding of interest (see the second or fourth panel in Figure 2) is that the mean γ_{RR} was significantly greater for delayed JOLs than for immediate JOLs, $t(37) = 2.59$, $d = 0.57$. This is relevant for theory because the self-fulfilling-prophecy (SFP) hypothesis (Spellman & Bjork, 1992), whose explanation of the delayed-JOL effect is limited to the hypothetical mechanism of greater potentiating effects from recall during delayed JOLs than during immediate JOLs, cannot account for this greater accuracy of delayed JOLs (relative to immediate JOLs) *within* the subset of items recalled during pre-JOL recall. However, the monitoring-dual memories (MDM) hypothesis (Nelson & Dunlosky, 1991, 1992), which proposes greater discrimination between items after delayed JOLs than after immediate JOLs (because of less contamination from the to-be-judged item's still being in short-term memory), is supported by this finding. Without the PRAM methodology, we would not have known about this finding, because the previous methodology did not allow for the determination of which items are recalled versus nonrecalled during the JOLs.

The second finding of interest (see the second or fourth panel of Figure 2) is that the mean γ_{RN} was significantly greater for delayed JOLs than for immediate JOLs, $t(30) = 4.33$, $d = 0.96$. This finding supports both the SFP hypothesis (Spellman & Bjork, 1992) and the MDM hypothesis (Nelson & Dunlosky, 1991, 1992), although those hypotheses explain the finding in different, non-exclusive ways. The SFP hypothesis explains it by proposing that the potentiating benefits of a correct retrieval increase with delay; by contrast, the MDM hypothesis explains it by proposing that JOLs monitor both short-term memory and long-term memory, and by emphasizing that only the latter is diagnostic for predicting performance on the eventual long-delayed retention test (and delayed JOLs are based on whatever information about the item is retrieved from long-term memory, whereas immediate JOLs are based on whatever information about the item is retrieved from short-term memory or long-term memory). Further research is needed to tease apart the contributions from those two hypotheses for why γ_{RN} is greater for delayed JOLs than for immediate JOLs.

The third finding of interest (see the second or fourth panel of Figure 2) is that the mean γ_{NN} was lower for delayed JOLs than for immediate JOLs. However, this difference played an almost negligible role in overall JOL accuracy here because it occurred so infrequently (e.g., only 3 of the 45 participants had computable estimates of γ_{NN} for immediate JOLs, and

[7] Other statistics (e.g., conditional probabilities, frequency distributions of judgments conditionalized on correct vs. incorrect during pre-JOL recall, latencies of pre-JOL recall and their correlations with other aspects of performance) will be reported elsewhere in an article that will focus specifically on substantive issues about the delayed-JOL effect and that will include inferential statistics pertaining to other conditions we investigated (e.g., a control group that did not have pre-JOL recall; at the end of the present *Results and Discussion* section, a few illustrative comparisons are mentioned).

only 1 participant had computable estimates of γ_{NN} both for immediate JOLs and for delayed JOLs).

Especially important for theory, the fourth finding of interest (see the second or fourth panel of Figure 2) is a major factor affecting overall JOL accuracy, namely, $\gamma_{RN} >> \gamma_{RR}$. This sizable effect occurs both for immediate JOLs and for delayed JOLs; also it is significant even using a sign test ($p < .02$ for each of those sign tests, $d = 0.43$ for immediate JOLs and $d = 1.26$ for delayed JOLs) and is unusually robust (e.g., occurring for 53 out of the 69 available comparisons from participants who contributed both a γ_{RN} and a γ_{RR}). These outcomes demonstrate how much more accurate people's forecast of their future performance is on one recalled item versus one nonrecalled item than on one recalled item versus another recalled item, and without PRAM this effect could not have been discovered. This effect becomes especially important when taken in conjunction with the effect of the relative frequencies of dyads upon which those two γs are based, as described next.

The pattern of weights for the component γs (see the third or fourth panel of Figure 2) varies as a function of immediate versus delayed JOLs. One way this appears is in the crossover interaction wherein for immediate JOLs the mean p_{RR} is significantly greater than the mean p_{RN}, $t(44) = 54.57$, $d = 16.30$, whereas the qualitatively opposite pattern occurs for delayed JOLs insofar as the mean p_{RR} is extremely less than the mean p_{RN} and yields a statistically significant difference, $t(43) = 27.08$, $d = 8.00$. We mention in passing that (a) the mean p_{RR} is significantly greater for immediate JOLs than for delayed JOLs, $t(43) = 55.89$, $d = 10.27$, whereas the mean p_{RN} is significantly greater for delayed JOLs than for immediate JOLs, $t(43) = 56.09$, $d = 10.16$—the interaction effect is significant, $F(1, 43) = 3{,}191.99$, $d = 16.68$—and (b) the magnitudes of the mean p_{NN} were extremely small, both for immediate JOLs and for delayed JOLs.

The potent way in which the crossover interaction (involving immediate vs. delayed JOLs and involving p_{RR} and p_{RN}) affects overall JOL accuracy becomes evident when considered in conjunction with the above-mentioned sizable difference between γ_{RN} and γ_{RR} wherein $\gamma_{RN} >> \gamma_{RR}$. One way to see this is by filling in the obtained values of the parameters in Equation 3 (with tolerance for extremely small errors due to rounding) separately for immediate JOLs versus delayed JOLs such that for immediate JOLs Equation 3 becomes

$$\begin{aligned} .23 &= (\underline{.954} \cdot \underline{.21}) + (.0004 \cdot .87) + (.045 \cdot .45) \\ &= \underline{.2003} + .0003 + .0203 = .22, \end{aligned} \quad (4)$$

whereas for delayed JOLs, Equation 3 becomes

$$\begin{aligned} .92 &= (.088 \cdot .46) + (.017 \cdot .35) + (\underline{.895} \cdot \underline{.96}) \\ &= .0405 + .006 + \underline{.8592} = .91 \end{aligned} \quad (5)$$

In words, the major difference between Equation 4 versus Equation 5 (see underlined entries in the two equations above) is that for immediate JOLs the overall accuracy of discrimination is dominated by the extremely frequent discriminations between items recalled at the time of the JOLs (and such discriminations are of only modest accuracy, $\gamma_{RR} = .21$) whereas for delayed JOLs the overall accuracy of discrimination is dominated by the extremely frequent discriminations between one recalled item versus one nonrecalled item at the time of the JOLs (and such discriminations are of extremely high accuracy, $\gamma_{RN} = .96$).

The second way to illustrate this pattern is via a HWAC graph, as shown in the fourth panel of Figure 2. The HWAC graph divides the overall area (of each bar shown in the top panel of Figure 2) into subcomponents, each of whose height represents the degree of discriminative accuracy on a given kind of dyad (i.e., γ_{RR}, γ_{NN}, or γ_{RN}), each of whose width represents the proportion of that kind of dyad (i.e., p_{RR}, p_{NN}, or p_{RN}), and each of whose area represents the amount of discriminative accuracy contributed by that kind of dyad (e.g., the product of $p_{RR} \cdot \gamma_{RR}$).

Two extreme possibilities can become obvious in HWAC graphs when one compares the effects of any conditions. One extreme possibility is that the pattern of the height of the bars (representing the pattern of the component γs) is identical across conditions but the pattern of the width of the bars (representing the pattern of the proportions of different kinds of dyads) varies across those conditions. The other extreme possibility is the opposite (i.e., the pattern of the widths of the bars is constant across conditions but the pattern of the height of the bars varies across conditions). What makes each of those possibilities extreme is that a given independent variable affects only the γs or only the proportions of relevant dyads, but not both. However, another possibility is that *both* the pattern of the height of the bars *and* the pattern of the width of the bars vary across conditions, as is obvious in the HWAC graphs depicting discriminative accuracy in the condition of immediate JOLs (left-most HWAC graph) versus delayed JOLs (right-most HWAC graph) in Figure 2.

For immediate JOLs, almost all of the discriminative accuracy arises from the RR dyads (i.e., the RR dyads—as indicated by the black bar in HWAC graph in the fourth panel of Figure 2—account for .20 of the overall area of .22 in the left-most bar in the top panel of Figure 2); because the accuracy of discriminating between items is only modest ($\gamma_{RR} = .21$), the overall accuracy of immediate JOLs is severely limited. By contrast, for delayed JOLs almost all of the discriminative accuracy arises from the RN dyads (i.e., the RN dyads—as indicated by the horizontal-lined bar in the HWAC graph in the fourth panel—account for .86 of the overall area of .91 in the right-most bar in the top panel); because the accuracy of discriminating between a recalled versus nonrecalled item is nearly perfect ($\gamma_{RN} = .96$), the overall accuracy for delayed JOLs is greatly enhanced.

Thus, as applied to the example of the delayed-JOL effect, an important and novel finding from the PRAM methodology is that the bulk of the delayed JOL effect on overall accuracy appears to be due to the conjunction of two factors: (a) $\gamma_{RN} >> \gamma_{RR}$ for both kinds of JOLs, but (b) $p_{RN} >> p_{RR}$ for delayed JOLs whereas $p_{RN} << p_{RR}$ for immediate JOLs. In words, PRAM shows that the bulk of the delayed-JOL effect arises from these two factors:

1. People who are predicting their future recall discriminate between items more accurately when discriminating between a recalled item versus a nonrecalled item (i.e., relatively easy discrimination) than when discriminating between two recalled items (i.e., relatively difficult discrimination).

2. For delayed JOLs, most of the relevant discriminations are between a recalled item versus a nonrecalled item (the relatively easy discrimination) and relatively few are between recalled items (the relatively difficult discrimination), but vice versa for immediate JOLs (i.e., most of the relevant discriminations are between recalled items).

This explanation for the bulk of the delayed JOL effect arises directly from the decomposition (of overall accuracy into its components) that is a central part of the PRAM methodology and that does not require hypothetical assumptions about unobservable processes. Elsewhere we will speculate on several kinds of processing that might underlie those components of metacognitive monitoring accuracy (e.g., different psychological processing may underlie discriminative accuracy at forecasting subsequent performance on nonrecalled items vs. recalled items, as suggested by Leonesio & Nelson, 1990).

Brief remarks about corresponding outcomes from a control group without PRAM. A no-PRAM control group of 45 participants from the same population as the PRAM group went through the same procedure as the PRAM group except without having any pre-JOL recall. We report here several[8] outcomes showing that the presence of pre-JOL recall does not lead to important changes in the general pattern of results that is obtained without pre-JOL recall. First, the mean proportion of correct final recall in the control group was similar to that in the PRAM group (.35 and .39, respectively, after immediate JOLs, and .41 and .49, respectively, after delayed JOLs). Although the main effect of immediate versus delayed JOL was statistically significant, $F(1, 88) = 61.0$, $MSE = 0.01$, the main effect of group was not significant, $F(1, 88) = 1.46$, $MSE = 0.10$, $ES_{est} = 0.43$. The Group × Judgment interaction was also not significant, $F(1, 88) = 3.54$, $MSE = 0.01$, $ES_{est} = 0.19$.

Second, the mean $\gamma..$ for JOL accuracy in the control group was similar to that in the PRAM group, with both groups showing facilitative effects of delayed JOLs over immediate JOLs (.85 and .92 for the no-PRAM group and PRAM group, respectively, for delayed JOLs, and .39 and .23 for the no-PRAM group and PRAM group, respectively, for immediate JOLs). Although the main effect of immediate versus delayed JOL was statistically significant, $F(1, 86) = 236.5$, $MSE = 0.07$, the main effect of group was not significant, $F(1, 86) = 1.77$, $MSE = 0.07$, $ES_{est} = 0.43$. The Group × Judgment interaction was also not significant, $F(1, 88) = 3.54$, $MSE = 0.01$, $ES_{est} = 0.19$. The interaction was statistically significant, $F(1, 86) = 8.01$, $MSE = 0.07$, with the magnitude of the delayed JOL effect being somewhat smaller in the no-PRAM group ($d = 2.05$) than in the PRAM group ($d = 2.44$); of primary importance, however, the simple effect of immediate versus delayed JOLs was significant for both the no-PRAM group, $t(43) = 9.69$, and the PRAM group, $t(43) = 11.95$.

Third, as shown in Figure 3, the distribution of JOLs from the control group was similar to that from the PRAM group, both for immediate JOLs (showing the usual inverted-U-shaped pattern in which JOLs of

[8] See footnote 6.

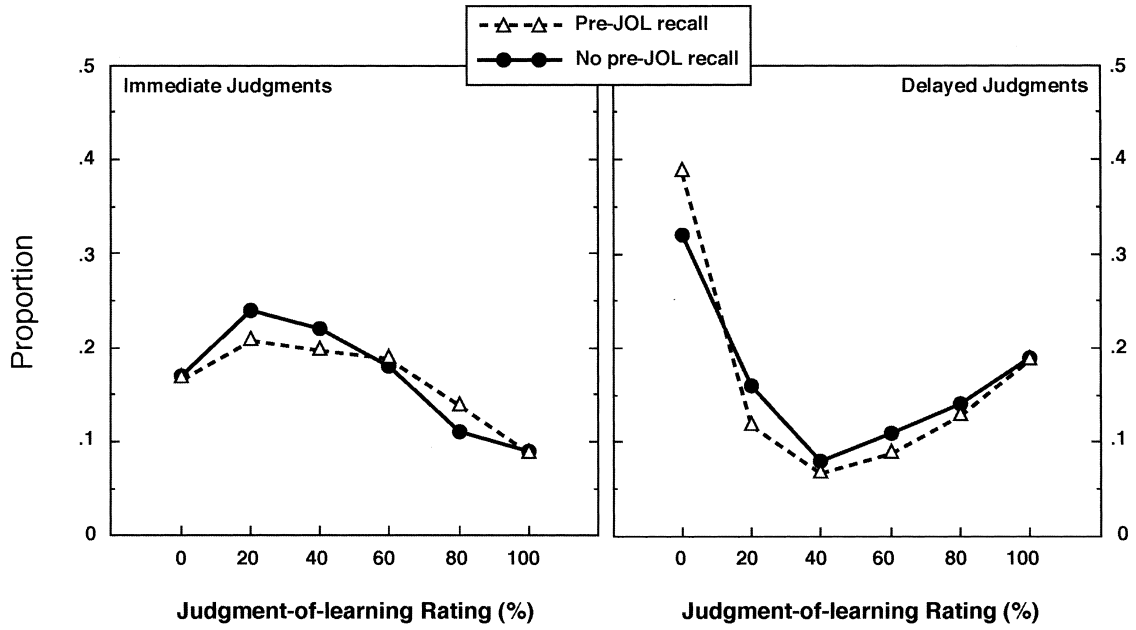


Figure 3. Mean proportion of items receiving each judgment-of-learning (JOL) rating. The inverted-U-shaped pattern (left) and the U-shaped pattern (right) are similar for the no-PRAM group and the PRAM group and are typical of the patterns reported in previous research.

20, 40, or 60 tend to occur more frequently than JOLs of 0, 80, or 100, as reported in previous research—e.g., Dunlosky & Nelson, 1994, 1997; Weaver & Kelemen, 1997) and for delayed JOLs (showing the usual U-shaped pattern in which JOLs of 20, 40, or 60 tend to occur less frequently than JOLs of 0, 80, or 100 as also reported in the above-cited research).

Concluding Remarks About the Implications of the PRAM Methodology

A given independent variable might produce no (or some) effect on the overall observed $\gamma..$, but this could be due entirely to a trade-off of the components in Equation 3. It is important, therefore, to know how or if a given independent variable affects each of those component processes. In regard to the usual overall $\gamma..$ for JOL accuracy that has been reported in the previous literature, we suppose that those components were occurring but were not being observed because the data-collection procedure at the time of the JOLs did not allow for analysis of the overall $\gamma..$ into its components (e.g., as in Equation 3).

Because research from the PRAM methodology is more analytic than that from the methodology that was used in previous research, the former is better than the latter for specifying the locus of any particular independent variable's effect on JOL accuracy. For instance, the effect of a given independent variable (e.g., the amount of study time or the number or spacing of repetitions) on overall accuracy (i.e., on $\gamma..$) could be due to an effect only on γ_{RR}, only on γ_{RN}, or only on γ_{NN} (or to some combination of those effects).

It is possible for an independent variable to increase the value of one of the components (e.g., γ_{RR}) and to decrease the value of another component (e.g., γ_{RN}) such that the overall effect on $\gamma..$ cannot be adequately explained without first analyzing the component gammas. In the extreme, one overall $\gamma..$ could be higher than another overall $\gamma..$, whereas a particular component γ (or even all three component γs) could show exactly the opposite pattern! One of the major goals of the present article is to stimulate a fruitful revision in the way in which researchers investigate the accuracy of metacognitive monitoring.

The PRAM methodology can be used not only for JOLs but also in the assessment of other kinds of metacognitive monitoring. The PRAM methodology could be incorporated into judgments of confidence about recognition in eyewitness identification. For instance, prior to a recognition test for the names of people who might have committed a particular crime,

the participant could be asked to recall the alleged criminals' names (or prior to a recognition test of people in a lineup, the participant could be asked to try to retrieve the faces and describe characteristics of the retrieved images—although caution is warranted to avoid negative effects of verbalization on visual identification; Meissner & Brigham, 2001), followed by the recognition test and a confidence judgment about the subjective accuracy of each chosen recognition alternative. Another kind of metacognitive judgment for which the PRAM methodology could be used is ease-of-learning judgments about upcoming acquisition. For instance, prior to making an ease-of-learning judgment on the to-be-acquired cue–target pair of *The capital of Australia–Canberra,* the learner could be asked to recall the capital of Australia and then would be shown the cue–target pair for an ease-of-learning judgment; then, prior to making an ease-of-learning judgment on the to-be-acquired cue–target pair of *The name of the highest mountain in South America–Aconcagua,* the learner could be asked to recall the name of the highest mountain in South America; and so on.

Thus, the application of the PRAM methodology to a given kind of metacognitive judgment is limited only by the researcher's creativity (including extensions to situations investigating multitrial learning, e.g., to determine how each of the parameter values of Equation 3 changes across trials). Regardless of the kind of metacognitive judgment to which the PRAM methodology is applied, the use of this more analytic investigation of metacognitive accuracy should help both in the development of empirical generalizations and in the testing of theories about metacognition.

References

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597–1611.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81,* 126–131.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4,* 316–321.

Bausell, R. B., & Li, Y. (2002). *Power analysis for experimental research: A practical guide for the biological, medical, and social sciences.* Cambridge, England: Cambridge University Press.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28,* 610–632.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. 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 metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence–accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7,* 26–48.

Butterfield, E. C., Nelson, T. O., & Peck, G. (1988). Developmental aspects of the feeling of knowing. *Developmental Psychology, 24,* 654–663.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997, November). *Aging and metamemory: Performance-level dependence of memory predictions.* Poster session presented at the meeting of the Psychonomic Society, St. Louis, MO.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups of repeated measures designs. *Psychological Methods, 1,* 170–177.

Dunlosky, J., Domoto, P., Wang, M., Ishikawa, T., Roberson, I., Nelson, T. O., & Ramsay, D. S. (1998). Inhalation of 30% nitrous oxide impairs people's learning without impairing people's judgments of what will be learned. *Experimental and Clinical Psychopharmacology, 6,* 77–86.

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic–extrinsic distinctions of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1180–1191.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20,* 374–380.

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 and Language, 33,* 545–565.

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory and Language, 36,* 34–49.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologist, 34,* 906–911.

Flavell, J., & Wellman, H. (1977). Metamemory. In R. V. Kail Jr. & J. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–33). Hillsdale, NJ: Erlbaum.

Freeman, L. (1986). Order-based statistics and monotonicity: A family of ordinal measures of association. *Journal of Mathematical Sociology, 12,* 49–69.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association, 89,* 190–200.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119,* 159–165.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Associations, 49,* 732–764.

Hartigan, J., & Kleiner, B. (1981). Mosaics for contingency tables. In W. Eddy (Ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (pp. 268–273). New York: Springer-Verlag.

Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? *Psychology and Aging, 17,* 209–225.

Hintzman, D. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review, 87,* 398–410.

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social metacognition: An expansionist review. *Personality and Social Psychology Review, 2,* 137–154.

Kelemen, W. L. (2000). Metamemory cues and monitoring accuracy: Judging what you know and what you will know. *Journal of Educational Psychology, 92,* 800–810.

Kelemen, W. L., & Weaver, C. A. (1997). Enhanced metamemory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1394–1409.

Kennedy, M., & Yorkston, K. M. (2000). Accuracy of metamemory after traumatic brain injury: Predictions during verbal learning. *Journal of Speech, Language, and Hearing Research, 43,* 1072–1086.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 1–22.

Koriat, A., Sheffer. L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131,* 147–162.

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

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20,* 159–183.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29,* 222–233.

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, and Cognition, 21,* 1263–1274.

Meissner, C., & Brigham, J. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15,* 603–616.

Metcalfe, J. (2000). Metamemory: Theory and data. In E. Tulving & F. Craik (Eds.), *Oxford handbook of memory* (pp. 197–211). Oxford, England: Oxford University Press.

Moulin, C. J. A., Perfect, T. J., & Jones, R. W. (2000). Evidence for intact memory monitoring in Alzheimer's disease: Metamemory sensitivity at encoding. *Neuropsychologia, 38,* 1242–1250.

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

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

Nelson, T. O. (1993). Judgments of learning and the allocation of study time. *Journal of Experimental Psychology: General, 122,* 269–273.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10,* 257–260.

Nelson, T. O., & Dunlosky, J. (1991). The delayed-JOL effect: When delaying your judgments of learning can improve the accuracy of your metacognitive monitoring. *Psychological Science, 2,* 267–270.

Nelson, T. O., & Dunlosky, J. (1992). How shall we explain the delayed-judgments-of-learning effect? *Psychological Science, 3,* 317–318.

Nelson, T. O., Graf, A., Dunlosky, J., Marlatt, A., Walker, D., & Luce, K. (1998). Effect of acute alcohol intoxica-

tion on recall and on judgments of learning during the acquisition of new information. In G. Mazzoni & T. O. Nelson (Eds.), *Metacognition and cognitive neuropsychology* (pp. 161–180). Hillsdale, NJ: Erlbaum.

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

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

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: Bradford Books.

Nelson, T. O., Stuart, R. B., Howard, C., & Crowley, M. (1999). Metacognition and clinical psychology: A preliminary framework for research and practice. *Clinical Psychology and Psychotherapy, 6,* 73–79.

Samuels, M. L. (1993). Simpson's paradox and related phenomena. *Journal of the American Statistical Association, 88,* 81–88.

Scheck, P., & Nelson, T. O. (2003). Metacognition. In L. Nadel, D. Chalmers, P. Culicover, B. French, & R. Goldstone (Eds.), *Encyclopedia of cognitive science* (Vol. 3, pp. 11–15). London: Macmillan.

Schneider, W., Vise, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring: Evidence from a judgment-of-learning (JOL) task. *Cognitive Development, 15,* 115–134.

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

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society,* Series B, *13,* 238–241.

Spellman, B. A., & Bjork, R. A. (1992). People's judgments of learning are extremely accurate at predicting subsequent recall when retrieval practice mediates both tasks. *Psychological Science, 3,* 315–316.

Surber, C. (1984). Issues in using quantitative rating scales in developmental research. *Psychological Bulletin, 95,* 226–246.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66–73.

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.

Weaver, C. A., & Kelemen, W. L. (1997). Judgments of learning at delays: Shifts in response patterns or increased metamemory accuracy? *Psychological Science, 8,* 318–321.

Received December 10, 2002
Revision received August 18, 2003
Accepted August 25, 2003 ■