Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. (2000). Accounts of the confidence-accuracy relation in recognition memory. Psychonomic Bulletin & Review, 7, 26-48. Page numbers here will not match the printed article.

Accounts of the confidence-accuracy relation in recognition memory

THOMAS A. BUSEY, JENNIFER TUNNICLIFF,
Indiana University, Bloomington, Indiana

GEOFFREY R. LOFTUS, AND ELIZABETH F. LOFTUS
University of Washington, Seattle, Washington

Confidence and accuracy, while often considered to tap the same memory representation, are often found to be only weakly correlated (e.g. Deffenbacher, 1980; Bothwell, Deffenbacher & Brigham, 1987). There are at least two possible (non-exclusive) reasons for this weak relation. First, it may be simply due to noise of one sort or another; that is, it may come about because of both within- and between-subject statistical variations that are partially uncorrelated for confidence measures on the one hand and accuracy measures on the other. Second, confidence and accuracy may be uncorrelated because they are based, at least in part, on different memory representations that are affected in different ways by different independent variables. In this article, we propose a general theory that is designed to encompass both of these possibilities and, within the context of this theory, we evaluate effects of four variables—degree of rehearsal, study duration, study luminance, and test luminance—in three face-recognition experiments. In conjunction with our theory, the results allow us to begin to identify the circumstances under which confidence and accuracy are based on the same versus on different sources of information in memory. In particular, we conclude the following. First, prospective confidence (assessed at the time of original study) and eventual accuracy are based on at least two different sources of information: Accuracy may be based upon some form of memory strength indicator that results from a recall-based mechanism, while prospective confidence may additionally be based on consideration of the conditions under which an item was studied. Second, given identical test circumstances, retrospective confidence (assessed at the time of test) and accuracy can be considered to be based on the same source of information, such as memory strength. Third, degrading a picture at test results in subjects including an analytic heuristic at the time of test in which subjects conclude (erroneously) that a brighter test face will always help performance. The results demonstrate the conditions under which subjects are quite poor at monitoring their memory performance, and are used to extend cue-utilization theories to the domain of face recognition.

Of interest in numerous circumstances is the ability to assess the degree to which a person's reported memory faithfully reflects the original, objective reality that gave rise to the memory. One such circumstance, for example, is the common legal scenario wherein a witness to a crime identifies a suspect as the person who committed the crime. Another is a laboratory setting wherein a subject claims to recognize a test stimulus in a recognition experiment.

In a controlled laboratory setting, the researcher has various tools available to assess memory. Two of the most commonly used are accuracy and confidence. Thus, to each recognition test stimulus, a subject can respond "old" or "new" and can also provide a confidence rating (say on a scale from 1 to 5) indicating his or her subjective assessment that the just-made recognition response is correct. Often, these two kinds of responses are assumed, either implicitly or explicitly, to be two measures of the same underlying psychological dimension. Thus experimenters often report both confidence and accuracy as parallel measures, or combine them into a single measure (e.g., multiplying a 1-5 point confidence rating by 1 or -1 for "old: and "new" responses respectively to arrive at a scale ranging from -5 to 5, which is assumed to reflect a continuum of internal evidence).

In the laboratory setting, a memory researcher is *able*, of course, to measure both confidence and accuracy. The measurement of confidence is straightforward: Numerical confidence ratings in some experimental condition are provided by the subject, and are taken at face value. The measurement of accuracy is also straightforward: Because the experimenter knows the "truth" for each test trial, the correctness of each test response is similarly known, and some variant of proportion correct can be computed over test trials for each experimental condition.

In an applied setting—for instance a legal setting—confidence ratings are, as in the laboratory, easily available: A police officer, for instance, asks the witness identifying a suspect to provide a "zero-to-seven" confidence rating. As in the laboratory, such ratings can be (and are) taken at face value. Accuracy, however, cannot be measured because the police officer, unlike the memory researcher, does not have the luxury of knowing the objective truth about what the witness originally saw (if such information were available, the witness's identification would not, of course, be necessary to begin with). Thus, a confidence rating is the

The research reported in this article was supported in part by an NIMH grant to Geoffrey Loftus.

only measure that is used to assess the validity of the witness's memory. Within the legal system, it is very explicitly assumed that confidence is a universally valid reflection (i.e., can be assumed to be a monotonic function) of accuracy. This assumption is, in fact, incorporated into Supreme Court decisions (e.g., *Neil v. Biggers*, 1972), and various other legal issuances and, indeed, high witness confidence appears to be a powerful variable in convincing jurors of the witness's accuracy (e.g., Cutler, Penrod, & Stuve, 1988).

Despite the frequently assumed correspondence between confidence and accuracy, there is a good deal of debate about the circumstances under which confidence and accuracy are in fact two measures of the same psychological entity. A growing body of evidence within the metacognitive literature suggests that confidence ratings may be influenced by information other than that retrieved from memory. In this article we elaborate upon this evidence using a new technique that provides a number of advantages over previous methods. This technique implies a simple dichotomization of theories within which the relation between confidence and accuracy can be assessed, along with corresponding data analyses. The combination of theory and data analysis is called *state-trace analysis*, the logic of which is described in detail by Bamber (1979). State-trace analysis has numerous general virtues, among the most importance of which for the present research are (1) that it addresses the same issues as do dissociation techniques but in a more general and more powerful manner (see Loftus & Irwin, 1998, pp. 140-145) and (2) it entirely avoids problems entailed in interpretation of scale-dependent interactions (e.g., Bogartz, 1976; Loftus, 1978).

Using state-trace analysis, we describe several findings concerning the circumstances under which confidence and accuracy can be construed to be measures of the same versus different memory representations. The results demonstrate how the sources of information that subjects use when making confidence ratings differ from those that underlie a recognition judgment.

Definitions

To avoid ambiguity, we define two types of confidence ratings and three types of correlations with which we are concerned and/or which are of concern in the literature.

Two Types of Confidence Ratings

1. A **prospective** confidence rating is one obtained at the time some stimulus is studied about how confident is the person that he or she will correctly recognize the stimulus. In the verbal learning domain, these are often called judgments of learning (JOLs).
2. A **retrospective** confidence rating is one obtained at the time of test about how confident is the person that he or she has made the correct recognition decision. In recognition, these confidence ratings differ from feeling of knowing (FOKs) ratings in that they are given after every recognition judgment, not just after recall failures.

Three Types of Correlations

1. A **within-subjects** correlation, computed for a given experimental condition, reflects the degree to which an individual subject is more accurate on trials when greater confidence is expressed.
2. A **between-subjects** correlation, also computed for a given experimental condition, reflects the degree to which subjects who are more confident also tend to be more accurate.
3. An **over-conditions** correlation reflects the degree to which confidence and accuracy are affected in equivalent ways by manipulations of experimental variables.

In the vast majority of past research on the confidence-accuracy relation, either within- or between-subjects correlations have constituted the primary measure. These correlations have been augmented by dissociation techniques in which an experimental variable is found that selectively affects confidence but not accuracy, or vice versa. In the present research, our focus is on over-conditions correlations. Here, we experimentally *induce* variation in both confidence and accuracy via manipulation of suitable independent variables, and we assess the degree to which these variables affect confidence—both prospective and retrospective confidence—and accuracy in similar fashions. It is via these assessments that we will be able to ascertain the circumstances under which confidence and accuracy are based on the same or different memory dimensions.

For comparison with previous work, we also report within- and between-subject correlations. However, we argue that there are difficulties with both measures that are addressed by state-trace analysis.

Correlations have been used in conjunction with a variety of dissociation and calibration techniques to provide a theoretical framework that describes the basis of prospective and retrospective confidence judgments. Below we discuss how confidence and accuracy measures might be related, as inferred from evidence from the verbal learning domain.

What Are Confidence Ratings Based On?

Prospective confidence ratings are generally found to be moderately good predictors of subsequent recognition (Leonesio & Nelson, 1990; Vesonder & Voss, 1985). The within-subject correlations are in the range of .25 to .4, and can improve to much higher levels (.90) if the rating is delayed several minutes after study (Nelson & Dunlosky, 1991). This suggests that confidence ratings and recognition judgments appear to be based, at least in part, on the same information. To account for these effects, a variety of theories have been proposed, which are reviewed by Schwartz (1994), and briefly summarized below.

Trace Access Theory (Burke, MacKay, Worthley, & Wade, 1991; Hart, 1967) posits a direct access to the contents of memory when making confidence and recognition judgments. Because confidence ratings and recognition rely on the same information, each predicts the other. This view has been augmented by a variety of theories which include other sources of information

that specifically affect confidence. For instance, making the test cue familiar through priming or other pre-exposure techniques can increase confidence ratings (Metcalfe, Schwartz & Joaquim, 1993; Schwartz & Metcalfe, 1992), while making answers familiar through tachistoscopic pre-exposure increases recall of general knowledge questions without affecting confidence judgments (Jameson, Narens, Goldfarb & Nelson, 1990). The ease of retrieval or perceptual fluency of an answer (correct or not) also contribute to confidence ratings (Kelley & Lindsay, 1993), such that an irrelevant dimension such as the speed of retrieval can inflate confidence beyond that warranted by an increase in accuracy. Other demonstrations show that attributes of the test item can differentially affect confidence and accuracy. For example, the retrieval fluency or ease of processing of the test cue appears to increase confidence ratings while leaving accuracy constant or even reduced (Benjamin, Bjork & Schwartz, 1998; Begg, Duft, Lalonde, Melnick, & Sanvito, 1989). If the prospective confidence ratings are delayed after the study session, predictive accuracy goes up, perhaps because the memory contents have settled (Nelson & Dunlosky, 1991; Thiede & Dunlosky, 1994), which Nelson and Dunlosky termed the Delayed JOL effect. This improvement is due in small part to a shift to the extremes of the confidence scale, but simulations by Weaver & Kelemen (1997) demonstrate that there is a real metacognitive improvement at a 5 minute delay condition. One possible explanation for this improvement is that the delay eliminates transient short-term memory effects, such that items that remain in memory after a 5 minute delay are likely to remain in memory at test.

The role of the cues that underlie confidence and accuracy has been summarized into an *accessibility hypothesis* proposed by Koriat (1995, 1997), in which people retrieve information from memory through a search process, and use whatever they retrieve as the basis for their confidence rating. Because this is a cue utilization theory, cues related to the target item or the item used to probe memory also influence the confidence rating. This leads to a situation in which irrelevant or even inaccurate information derived from the target item gives the illusion of expertise in the absence of any real knowledge, inflating confidence and producing a dissociation between confidence and accuracy.

The vast majority of evidence in support of the accessibility hypothesis and other cue utilization theories comes from the verbal learning domain. However, within the face recognition domain the best evidence still supports a trace access view. Sommer, Heinz, Leuthold, Matt and Schweinberger (1995) used an evoked-response-potential (ERP) analysis of judgment of learning (JOL) ratings in a picture-recognition study. This study focused on the scalp topologies of electrical activity elicited during study of a face. The resulting wave forms were segmented according to the prospective confidence rating given at the time of study and compared with the wave forms segmented according to whether the face was correctly recognized later in the test session. These two distributions were quite similar, leading the authors to conclude that the brain processes underlying the prospective confidence ratings (JOLs) and the recognition accuracy judgments were similar. They implicate facial distinctiveness as a moderating variable of both confidence and accuracy, suggesting that distinctive faces are more likely to be encoded, leading to higher JOLs at study and higher recognition at test. In support of this conclusion, the correlation between confidence and accuracy was fairly high (gamma = 0.44).

Other research supports a strong relation between confidence and accuracy in face recognition. Read, Lindsay & Nicholls (1998) conducted a number of between and within-subject correlational studies that demonstrate strong correlation coefficients. For example, the mean correlation coefficient for subjects viewing a lineup was .58, with 72% of the subjects obtaining a coefficient greater than .50. They identify a variety of possible moderators of the confidence and accuracy relation, including immediate vs delayed testing, the response options available at test (instantiated as a lineup or showup decision) and the orientation of the witness to the target. While the data contain some hints that subjects use irrelevant information when making confidence judgments, overall this work supports a view in which confidence and accuracy are highly related, perhaps because they both rely on the same information.

To summarize, a number of factors other than direct memory access have been identified as the basis of confidence judgments made to recognition or recall of verbal materials. The few studies with faces still support a direct access view, or at least one in which confidence and accuracy rely on much of the same information. The present studies explore the possibility that confidence and recognition judgments may in fact be based on different sources of information, and if so, provide a theoretical account that describes the bases of the two judgments.

Present Paradigm

In later sections, we report three experiments, all using a face-recognition paradigm. We analyze these experiments within the context of a general theory to be presented in the next section. With suitable minor modifications, the theory could be applied to virtually any memory paradigm. However, in order to have a concrete expositional basis for describing the theory, we briefly sketch our experimental paradigm here.

We used a study-test face-recognition paradigm. In a study phase, 60 target faces were presented. In an immediately following test phase, memory for the faces was tested in an old-new recognition procedure. At the time of study, two variables were factorially combined. There were five levels of a perceptual variable (stimulus duration in Experiment 1, and stimulus luminance in Experiments 2 and 3). In addition, following each studied face, subjects spent a 15-sec period during which they were either required to rehearse the just-seen face, or were preventing from rehearsing it.

Three dependent variables were measured in each experiment. A prospective confidence rating was obtained after the 15-sec rehearsal/non-rehearsal period of each study trial. An old-new recognition judgment was obtained for each test trial. Finally, a retrospective confidence rating was obtained following each old-new

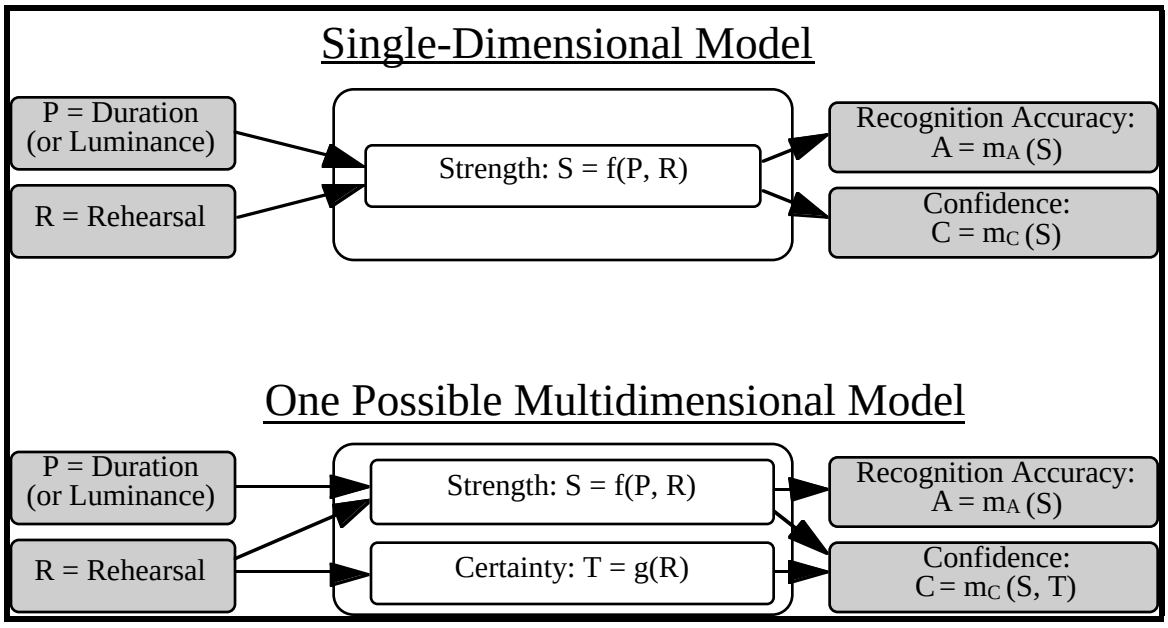

Figure 1: Two Models of the Confidence/Accuracy Relation

judgment.

Theory

Previously, confidence and accuracy have often (but not always) been dissociated by finding experimental variables that affect confidence but not accuracy, or vice versa. The resulting interactions are scale-dependent in many cases, and could be eliminated by a monotonic transformation of the dependent variables. One of the contributions of the present work is the application of an analysis technique that will demonstrate dissociations that are scale-invariant and therefore immune to any monotonic transformation. In this section, we will present a general theory within which the relation between confidence and accuracy (or the relations among any set of dependent variables) is formally and systematically conceptualized in terms of whether these variables reflect a single cognitive dimension or multiple cognitive dimensions. If they reflect a single dimension, then all independent variables observed to affect the dependent variables must do so via the "common currency" of the single dimension. If they reflect multiple dimensions, then there can be numerous configurations of the flow of effects from independent variables to the dimensions to the dependent variables, and it becomes of interest to isolate the configuration that best accounts for the data. Below we are more specific about what we mean by this.

Model Representations

The top panel of Figure 1 shows the single-dimensional model. By it, the values of both independent variables (P, the perceptual variable, and R, rehearsal) are assumed to affect a single dimension of the memory representation, which, for mnemonic convenience, we call *memory strength*, S. We should stress that this label is for expositional purposes only; in the General Discussion we explore the basis for this dimension. Until then we use this label only to denote that, under a single-dimensional model, the value of memory strength determines both confidence and accuracy. Although a single dimensional model is consistent with the trace access theory, it is also consistent with any other single dimensional model in which confidence and accuracy are based on the same information. Thus the term 'memory strength' should not be interpreted as equivalent to trace access.

The magnitude of memory strength following a study trial is

$$S = f(P, R)$$

where f is a function that is monotonic in both P and R. Confidence (C) and accuracy (A) are both assumed to be monotonic functions, m_C and m_A, of S. The exact forms of the monotonic functions are not critical to the present logic.

A single-dimensional model, (somewhat akin to a standard null hypothesis), is very specific, and makes very specific predictions, which we will describe below. If one abandons a single-dimensional model, then one must decide among the infinite number of possible multidimensional models (just as if, for example, on rejects a null hypothesis in an ANOVA, one must decide among the infinite possible alternative hypotheses). The two-dimensional model shown at the bottom of Figure 1 is designed to capture the hypothesis that rehearsal affects confidence more than it affects accuracy as found, for example, by Wells, Lindsey, and Ferguson (1979). Here, there are two dimensions in the memory representation, memory strength, S, as described above, and a second dimension, T, which (again for mnemonic convenience) we label memory *certainty*. We explore the theoretical basis for this second dimension in the General Discussion, but to give the general

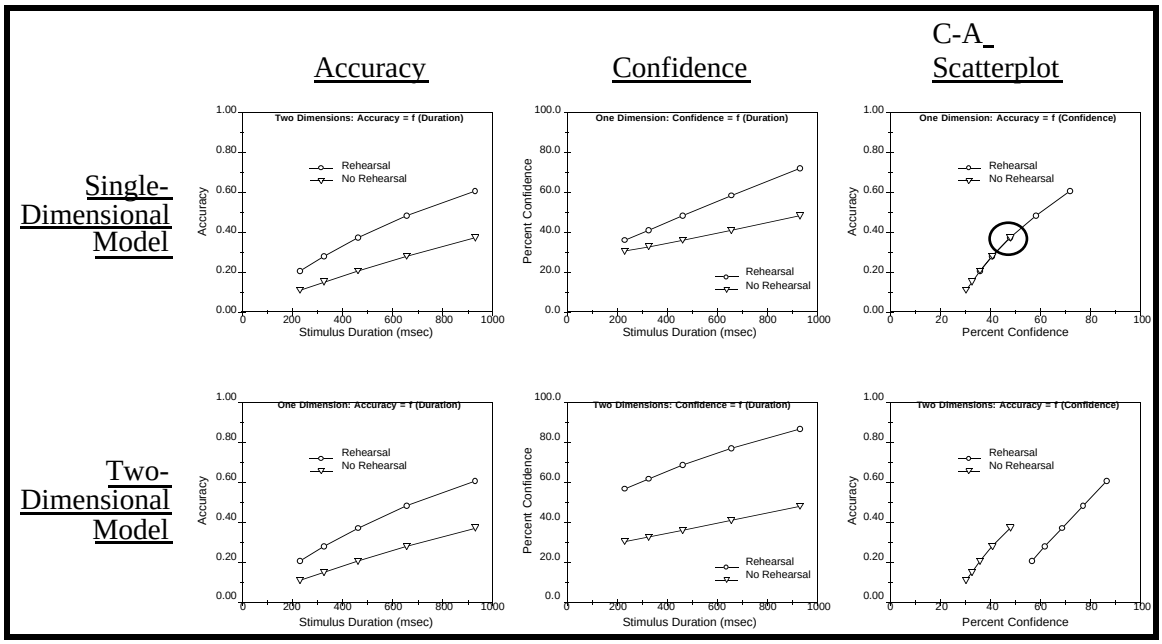

Figure 2: Predictions of the two Models

flavor of this dimension from the perspective of the metacognitive literature, certainty might include probe-related cues such as probe familiarity due to pre-exposure to the cue or analytic heuristics (this question is about U.S. Presidents and I'm an expert in this field, so I must have got it right). These are sources of information that do not or cannot influence the recognition judgment but give the illusion of accuracy and thus affect confidence.

Exactly as in the single-dimensional model, S is a monotonic function of both P and R. T, however, is a function (again monotonic) only of rehearsal, R. Accuracy, as in the single-dimensional model is determined only by strength: again, $A = m_A (S)$. Confidence, however, is a function of both strength and certainty, $m_C (S, T)$, where m_C is monotonic in both arguments.

Model Predictions

Figure 2 shows predictions of the single-dimensional model (top three panels) and of the two-dimensional model (bottom three panels). The predictions were generated using study duration as the perceptual independent variable (the arguments would be identical if luminance were used instead) and selecting specific (although somewhat arbitrary) choices for the monotonic functions f relating strength to accuracy and confidence. These are shown as m_C and m_A in Figure 1, and are described in a later section.

The left and middle panels of Figure 2 show what we will refer to as "standard data." Here the two dependent variables, accuracy and confidence, are plotted as functions of the independent variables: duration and degree of rehearsal. Several comments are in order about these hypothetical data. First, the qualitative patterns are as would be anticipated by common sense, and by any reasonable model: Both confidence and accuracy increase monotonically as a function of both independent variables. Second, using just the standard data, one could not easily tell that the two data patterns issue from two quite different models. If, for instance, one observed the top and bottom data patterns in two different experiments, one would feel comfortable asserting them to be replications of one another.

The key predictions that distinguish the two models are shown in the right-hand panels, which are accuracy-confidence *scatterplots*. Thus, for each of the 12 conditions of the experiment, the accuracy value obtained from the left panel is plotted against the confidence value obtained from the middle panel. As in the left-hand and middle plots, circles correspond to the rehearsal conditions, while triangles correspond to the no-rehearsal conditions. Data points within each rehearsal condition are connected by lines. These scatterplots are referred to by Bamber (1979) as *state-trace plots*, and the reader is referred to Bamber's article for a detailed description of the formal logic underlying the relation between these plots and the kinds of models illustrated in Figure 1.

As is evident, the prediction of the single-dimensional model is that the there is a perfect rank-order correlation over the experimental conditions; in other words, the rehearsal and no-rehearsal curves completely overlap. Informally, the reason for this prediction can be illustrated as follows. Consider the circled pair of overlapping data points in the upper-right-hand scatterplot. The circle corresponds to a 462-ms rehearsal condition, while the triangle corresponds to the 930-ms

no-rehearsal condition. Because these two physically distinct conditions produce the same level of accuracy (0.372), they must, by the single-dimensional model of Figure-1 have produced the same value of strength (in particular, $S = m_A^{-1}(0.372)$ where m_A^{-1} is the inverse of m_A). This, in turn, means that these two conditions must also produce the same value of confidence, equal, in this example to $m_C(S) = m_C(m_A^{-1}(0.372)) = 48.1\%$. In other words, any two conditions producing the same value of accuracy must also produce the same value of confidence, which is why the curves must, in overlapping regions, fall on top of one another.

The prediction of the two-dimensional model is that the curves corresponding to the two rehearsal conditions are separated: As shown, the rehearsal curve falls to the right of the no-rehearsal curve. The reason for this can be illustrated as follows. Consider again two different duration-rehearsal conditions that lead to the same value of memory strength, S. Because accuracy is determined only by strength (recall that $A = m_A(S)$) these two conditions must lead to the same accuracy value. Confidence, however, is determined by both strength and certainty (recall that $C = m_C(S, T)$, where m_C is monotonic in both arguments). Thus confidence will be higher in the rehearsal condition, which produces a higher certainty value than in the no-rehearsal condition, which produces a lower certainty value. The net result is that the rehearsal curve is shifted to the right of the no-rehearsal curve. Such a situation might result if aspects of the study condition (i.e. rehearsal vs. no rehearsal) lead to an analytic process in which subjects assume that rehearsal will produce much better accuracy than no rehearsal. Rehearsal may indeed improve accuracy, but in this case the subjects overestimate the advantage given by rehearsal, which leads to the separation of the curves. At test, attributes of probe (i.e. its familiarity or ease of processing) or other conditions of testing may also affect confidence and accuracy differently.

Prediction Summary

A finding that the rehearsal and no-rehearsal scatterplot curves fall atop one another confirms a single-dimensional model. A finding that the two curves fall in different places disconfirms a single-dimensional model and confirms a multidimensional model. In the latter case, the nature of the curve separation would suggest the nature of the specific two-dimensional model. For example, a finding that the rehearsal curve is to the right of the no-rehearsal curve would suggest the two-dimensional model shown at the bottom of Figure 1 and would allow the intuitive conclusion that "rehearsal leads to an overconfidence that is not warranted by rehearsal's effect on accuracy."

EXPERIMENTS

We report three experiments. In each, two variables are factorially combined at study: a perceptual variable (stimulus duration in Experiment 1 and stimulus luminance in Experiments 2 and 3) and amount of post-stimulus rehearsal. Three dependent variables are measured: prospective confidence, accuracy, and retrospective confidence. The major question is: Can both types of confidence be accounted for by a single-dimensional model, or is a multidimensional model necessary to explain one or both? We should note that our choice of independent variables correspond to those that are important to a witness who observes a crime. The lighting might be poor or good, the criminal might be observable for a brief or longer duration, and post-event conditions might either allow or prevent rehearsal of a particular face.

Experiment 1

In Experiment 1 we used a face-recognition paradigm in which two within-subjects variables—stimulus duration and whether rehearsal was required or prevented—were factorially combined.

Methods

The methods for Experiments 1-3 are similar; we describe the general methodology here, and describe methodology particular to specific experiments in subsequent sections.

Subjects

One hundred and eight Indiana University undergraduates participated for course credit. They were run in 20 groups of at least 3 subjects per group.

Stimuli

The stimuli were 120 pictures of bald men. The pictures were all taken under similar lighting conditions and all men had similar expressions. About 1/3 of the men had facial hair. The faces were digitized and displayed on a 21" Macintosh grayscale monitor using luminance control and gamma correction provided by a Video Attenuator and the VideoToolbox software library (Pelli & Zhang, 1991). The monitor's background luminance was set to 5 cd/m^2 . The contrast of naturalistic images is not possible to define; here we simply scaled the grayscale values in the images to cover the range from 5 cd/m^2 (essentially black) to 80 cd/m^2 (essentially white).

Data were collected by a PowerMac computer using 5 numeric keypads that provided identifiable responses from each keypad.

Design

Two factors, exposure duration and rehearsal, were factorially combined. Five values of exposure duration ranged from 230-930 ms in logarithmically-equal steps. There were two levels of the rehearsal manipulation: for 15 seconds following stimulus offset, subjects either silently rehearsed a face (without, of course, being able to see it) or performed math problems as a distracter task.

Procedures

The experiment consisted of two halves, each half containing a study phase of 30 target faces, followed by an immediate test phase of 60 test faces. The two halves were merely replications of one another with new sets of faces.

During each study phase, each of the 10 distractor x rehearsal conditions occurred 3 times. The following sequence of events occurred on each study trial.

1. A 400-ms warning tone occurred beginning 500 ms prior to stimulus onset.

2. The target face was shown for the appropriate exposure duration
3. The face was replaced by either instructions to rehearse the face using elaborative strategies ("e.g. does this person look like someone you would like to meet") or by a list of math problems to complete. The math problems were displayed all at once on a slide that contained disembodied features of different faces. Both the rehearsal and the math-problem tasks continued for 15 seconds following the picture's offset.
4. Subjects then gave a prospective confidence rating on a 5-point scale ("0%, 25%, 50%, 75% or 100% certain) reflecting their confidence that they would be able to correctly identify the just-seen face later in the test session. The instructions for providing the prospective confidence rating were as follows.

> After the tone, the picture will appear. Study the picture, and try to remember it. After the picture disappears, there will be a short pause, and then we will ask you to perform one of two tasks. On some trials we will ask you to mentally rehearse the picture of the face: do this by trying to imagine the face or think about person's personality. On other trials we will ask you to perform some math problems. On these trials you will start at the top of a list of math problems and try to work the problems in your head. When you have the answer, type it into the computer keypad and go on to the next problem. After about 15 seconds of either of these two tasks, we will get a measure from you that indicates how well you think you will be able to remember the face later on. You will use the response boxes to give your answers. We want you to judge how well you think you will remember the face later on, ranging on a scale from 1, which means that you are 0% confident that you will remember the picture, to 5, which means that you are 100% confident that you will remember the picture later on. 2 means you are 25% confident, 3 means you are 50% confident and 4 means you are 75% confident.

Following the 30 study trials was a test session in which subjects viewed 60 faces—the 30 targets that they had just seen in the study session, plus 30 new (distractor) faces. The 60 test pictures were presented in random order. Each test face remained on the screen until all subjects had entered their old/new recognition response into the keypad. Following their recognition responses, subjects gave a retrospective confidence rating on the same 5 point (0% - 100%) scale that indicated their confidence in the accuracy of the just-given recognition response. Instructions for the retrospective confidence rating were analogous to those shown above for prospective confidence ratings.

As indicated, this study-test sequence was repeated twice, thereby resulting in 6 replications per condition per subject. The experimental session was preceded by a practice study session in which 3 sample study trials and 6 sample test trials were used to give subjects an idea of the nature of the procedures.

The counterbalancing procedures were such that, over the 20 groups, each face appeared as a target for 10 groups and as a distractor for the other 10 groups. In addition, each face appeared in each of the 10 study conditions over the 10 groups for which it appeared as a target.

Dependent Measures

Subjects making both prospective and retrospective confidence ratings on a 5 point scale were encouraged to use the entire scale from 0% to 100% in an effort to discourage shifting of the confidence criteria across trials.

Accuracy is based on both the hit rate and the false-alarm rate and is computed via the equation, accuracy = (H- FA)/(1-FA), where H and FA are hit and false-alarm probabilities. The high-threshold model that implies this measure is based on dubious assumptions. However, because there is only a single false-alarm rate, any measure that is monotonically related to hit rate is sufficient for testing the models described above. The accuracy measure that we have chosen has the advantages of having a meaningful zero point, and not being uncomputable under frequently occurring situations (as, for example, happens with d' when either the hit or the false-alarm rate is either zero or 1.0).

Results and Discussion

The mean false-alarm rate across subjects was 0.266, and the mean confidence rating for distracters was 69.25%. Figure 3 shows main data. Figure 3, which is typical of other data figures to be presented in this article, contains five panels. The top three panels correspond to what we have referred to as the "standard data": They show, respectively, accuracy, prospective confidence, and retrospective confidence as functions of exposure duration, with separate curves shown for the rehearsal and no-rehearsal conditions. In this and subsequent data figures, the error bars are standard errors. Note that in some instances, there appear to be no error bars. This is because the size of the error bars are smaller than the size of the curve symbols. The bottom two panels show the accuracy-confidence scatterplots. In this and all data figures, circles represent the rehearsal conditions while triangles represent the no-rehearsal conditions. The small panel within each of the panels shows theoretical predictions. These predictions were generated using somewhat arbitrary functions[1] to replace the monotonic functions that comprise our general theory. These predictions should be taken only as an existence proof that at least one quantitative instantiation of our general theory can predict data that mirror the observed data reasonably well.

[1] Strength, S, and certainty, T, were assumed to be linear functions of P and R; accuracy was assumed to be a negative exponential function of S, and confidence was assumed to be a cumulative normal function of S+T.

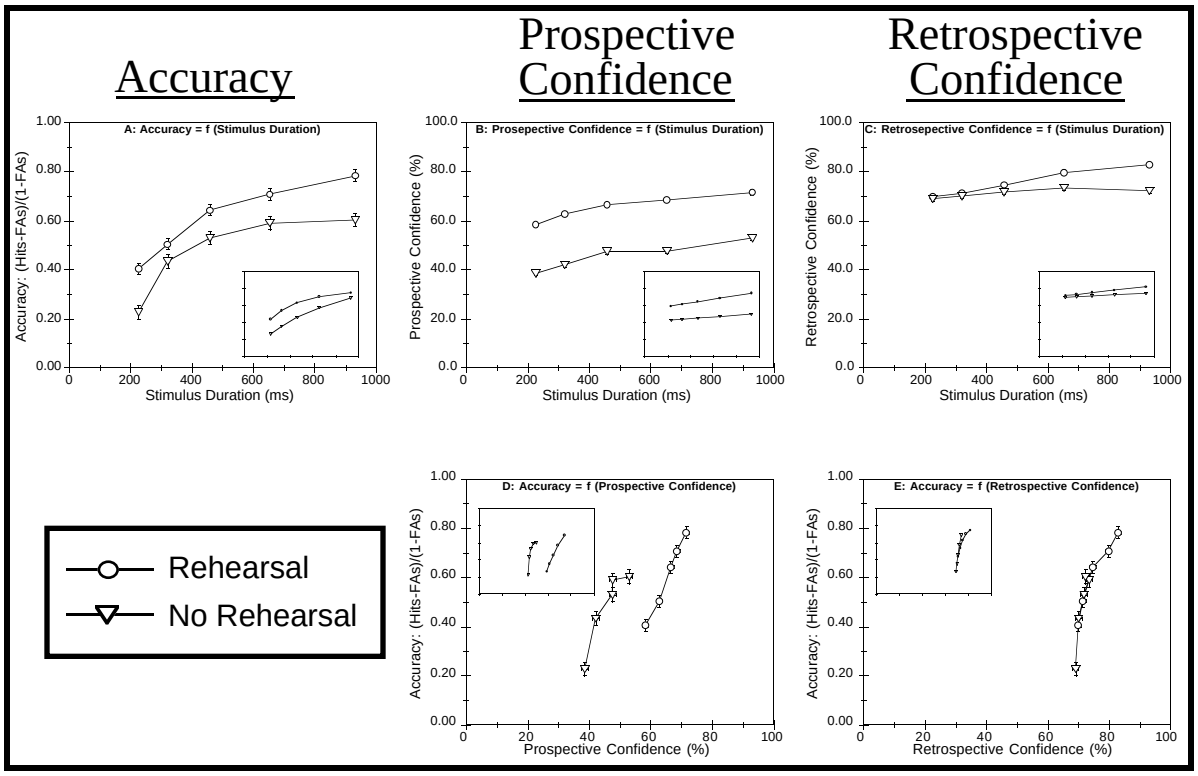

Figure 3: Experiment-1 data.

Standard data

There is little of surprise in the standard data. Both accuracy and prospective confidence increase with stimulus duration and with rehearsal. Much the same is true of retrospective confidence except that there is a relatively small effect of rehearsal at the shortest three exposure durations.

Confidence vs accuracy: Scatterplot data

The scatterplots relating accuracy to confidence are shown in the two bottom panels for prospective and retrospective confidence. For each panel, the two curves correspond to the two rehearsal conditions, and the five points within each curve correspond to the five durations within each rehearsal condition.

The results, and corresponding conclusions, could not be more clear-cut. For prospective confidence, the rehearsal curve falls to the left of the no-rehearsal curve. This disconfirms a single-dimensional model and confirms the two-dimensional model that is depicted in the bottom of Figure 1. A straightforward interpretation is as described earlier: Accuracy is determined by a single dimension (e.g., "strength") which is positively affected by both duration and rehearsal. Prospective confidence, however is determined by two dimensions: (e.g., what we have referred to as *strength* and *certainty*. Certainty is positively affected by rehearsal but is unaffected by duration. This result is consistent with Koriat's Accessibility Hypothesis, in which an analytic process is used by the subjects to provide an estimate of the benefits of rehearsal. Subjects assume that rehearsal will produce better recognition, and therefore inflate their confidence rating. However, they overestimate the benefits of rehearsal, which results in a rehearsal curve that is shifted to the right of the no rehearsal curve. These results do not support a trace access view as the only bases of confidence ratings.

For retrospective confidence, the rehearsal and no-rehearsal curves fall atop one another. This confirms a single-dimensional model and disconfirms multi-dimensional models. Thus, accuracy and retrospective confidence can be construed as being determined by a single dimension, strength, in memory. This finding is consistent with the trace access theory, although it is also consistent with any other single-dimensional model in which confidence and accuracy are based on the same information.

It thus appears that the relation between confidence ratings and recognition performance changes over time: initially confidence ratings are overly influenced by the rehearsal manipulation, while later during the test session the confidence ratings appear to be based on the same source of information as is recognition performance. This is consistent with the improvement seen in the Delayed JOL effect (Nelson & Dunlosky, 1991; Thiede & Dunlosky, 1994; Kelemen & Weaver, 1997). One important difference between the two measures is that at test, the conditions of study may no longer be in memory to affect the confidence ratings through an analytic heuristic.

Between and within-subjects correlations

We defer presentation of between- and within-subjects correlations until a later section wherein we present these results from all three experiments simultaneously.

Experiment 2

Experiment 2 was identical to Experiment 1, except that the stimulus exposure duration manipulation was replaced with a stimulus luminance manipulation. Exposure duration and luminance are both methods for limiting the rate at which information can be acquired from a scene, and hence the total amount of information that can be acquired during a given exposure duration (e.g. Loftus, 1985). The major purpose of Experiment 2 is to generalize the Experiment-1 findings by replicating them using a different environmental variable.

Methods

Experiment 2 used the same stimuli and equipment as Experiment 1, with the following exceptions:

Subjects

Subjects were 99 Indiana University undergraduate students who took part in the experiment for course credit. They were run in 20 groups of at least 3 subjects per group.

Stimuli and Design

The faces during the study session were presented at one of 5 luminance levels. The luminance of the faces was modified by reducing the luminance of the brightest white in the picture from 80 cd/m^2 (used in the Experiment 1 stimuli) down to a minimum of 10 cd/m^2. The intermediate luminance values were linearly interpolated between the minimum and maximum values. This manipulation has the effect of reducing the contrast of the image, analogous to dimming the lights in a room[2].

All stimuli were presented for 1350 ms during the study session. All test stimuli were presented in the bright (80 cd/m^2) condition.

Procedure

Subjects were expressly instructed to respond "old" to a face they thought they had seen in the study session *regardless of whether it was at a different luminance level*. All of the test faces were shown at the brightest luminance level.

Results and Discussion

The mean False-Alarm rate across subjects was 0.265, and the mean confidence rating for distracters was 68.95%. Figure 4 shows the main data. The top three panels indicate that luminance in Experiment 2 acts very much as did duration in Experiment 1. However, there are some differences between the results of the two experiments. First, the positive effect of rehearsal on accuracy is smaller and indeed is reversed for the lowest luminance level. This effect does not replicate in Experiment 3, wherein the identical condition produced the expected positive rehearsal effect; hence we believe that the reversal results from statistical error. The second difference is that the effect of rehearsal on retrospective confidence is very small.

Despite these apparent interexperiment inconsistencies, the scatterplots shown in the bottom of Figure 4 are essentially identical to their Experiment-1 counterparts. Again for prospective confidence, the rehearsal scatterplot falls to the right of the no-rehearsal scatterplot, and for retrospective confidence, the two scatterplots fall atop one another.

Summary of Experiments 1 and 2

The state-trace plots comparing prospective confidence with accuracy reveal that prospective confidence ratings and recognition judgments are based on different sources of information in memory. The situation can be summarized by supposing that rehearsing a face increases a subject's confidence more than is warranted by what will be the eventual increase in accuracy that rehearsing the picture actually confers. In contrast, retrospective confidence judgments and accuracy appear to be based on the same source of information in memory, perhaps because the study conditions surrounding each face are no longer preserved in memory to differentially influence retrospective confidence.

As with Experiment 1, these data are consistent with a cue utilization theory that proposes that analytic processes applied to the knowledge of the study conditions can result in an overestimation of the benefits of rehearsal when making prospective confidence judgments. This demonstrates that although a covert retrieval attempt might contribute to prospective confidence ratings (e.g. Spellman & Bjork, 1992), additional information about the study conditions also contributes to confidence judgments. Retrospective confidence judgments appear to be based on the same sources of information as the recognition judgment, which is consistent with a trace access theory, although it is also consistent with any other single-dimensional model of confidence and accuracy judgments.

Experiment 3

The findings concerning retrospective confidence judgments in Experiments 1 and 2 imply that, at the time of test, both confidence and accuracy are based on the same sources of information. This supports familiarity-based models such as signal-detection theory which assume that studied and non-studied items will generate a value on a *single* dimension (e.g., strength) whose value then determines both confidence and accuracy. By such models, confidence ratings are simply a

[2] Some comments about the display device are in order. The combination of the VideoToolbox library routines and the video attenuator provide an increase in the resolution of the grayscales available. Most computer video cards can display up to 256 gray levels, and the range of voltage values spanning the 5 to 10 cd/m^2 range might be only 4-5 gray levels. An attempt to display the grayscale images at this reduced luminance on such a monitor would introduce artificial boundaries in the faces. The video attenuator used in the current experiments combines the red, green and blue channels into a single luminance channel, which provides 4096 separate gray levels. This becomes important when the luminance is reduced: all changes in luminance that occurred at high luminance levels were present in the low luminance stimuli, albeit at proportionately lower levels. No artificial boundaries were introduced into the face by a reduction of the pixel luminance values.

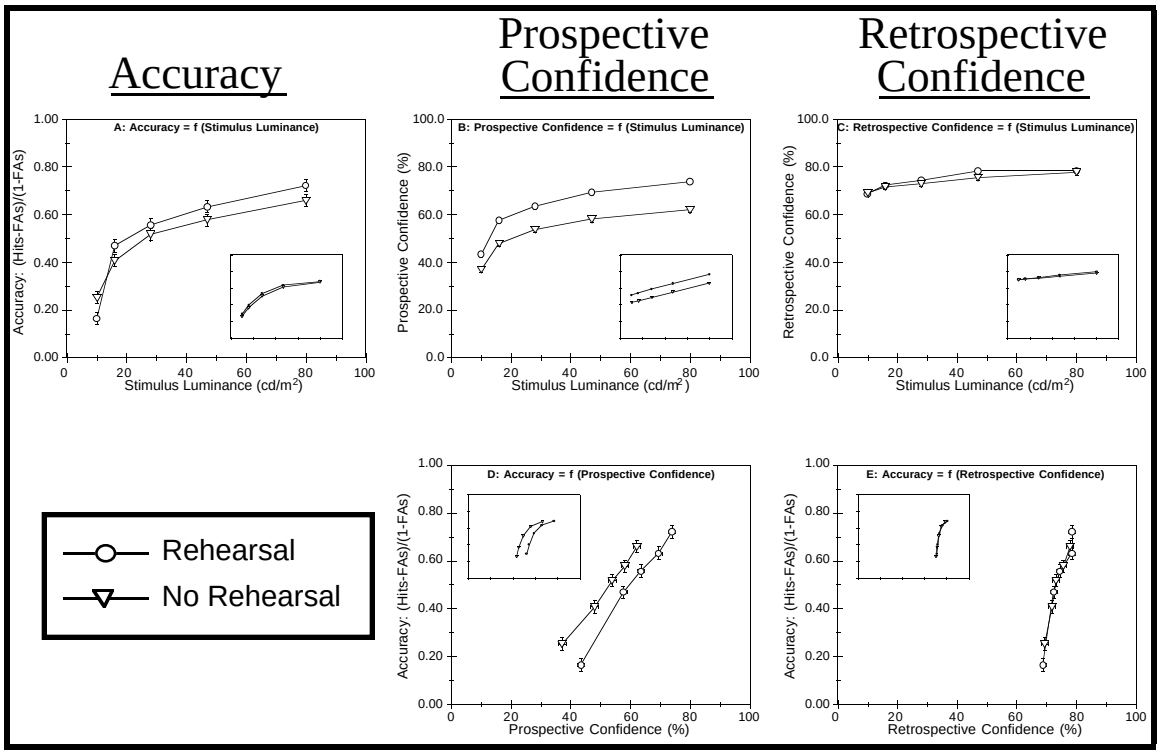

Figure 4: Experiment-2 data.

more fine-grained estimate of the value along the single dimension. However, several studies have shown that accuracy and retrospective confidence do not always covary in the identical fashion. Three examples are as follows.

Wells, Lindsay, & Ferguson (1981) carried out a simulated theft following which eyewitnesses attempted to pick out the thief from a lineup. Twenty subjects who correctly picked out the thief and 38 who incorrectly picked someone else from the lineup were selected for further study. A randomly selected half of each of these two groups was then briefed by a prosecutor about what they would say during cross-examination at trial; the other half was not briefed. Confidence was then assessed. When not briefed, the accurate subjects were more confident than the inaccurate subjects; however the reverse held true for the briefed subjects. Thus in the Wells et al. experiment, the effect of briefing on retrospective confidence was akin to the effect of rehearsal on prospective confidence in the present Experiments 1 and 2: It increased confidence more than was warranted by its effect on accuracy.

Chandler (1994) presented pictures at study, and then presented either related or unrelated pictures during an intervening phase of the experiment. She found that studying related pictures during the intervening phase increased confidence and decreased accuracy for a forced-choice task. She attributed this finding to participants using generic knowledge about a picture when making confidence judgments, without realizing that only the specific detail information was relevant to the task.

Tulving (1981) presented a series of photographs (indexed as A, B, C...) and then presented forced-choice test trials. In each test trial the two pictures contained an original photograph (denoted as A) and a foil that was either similar to the original photograph (denoted as A') or similar to another photograph in the study list (denoted as B'). Following each response, subjects made a confidence judgment on a 1 (least confident) to 4 (most confident) scale. Surprisingly, forced-choice accuracy was better in the A/A' condition than the A/B' condition, while confidence was higher in the A/B' condition.

Experiment 3 was designed generally to investigate the effect of another post-study variable, test luminance, on the retrospective confidence-accuracy relation, and was motivated by the following common legal scenario. During a crime, for example a mugging, a witness sees the mugger's face under poor environmental circumstances—for instance, it is dark or the witness has only limited duration for observing. Later the witness is asked whether s/he can identify a suspect in a photo montage. This "test stimulus" is customarily shown under optimal conditions—the witness has ample time and the lighting is good. The question is: does this test configuration affect confidence more than is warranted given its concomitant effect on accuracy?

In Experiment 2 all test stimuli were shown at the brightest luminance level. Because there were five luminance levels at study, this means that for 8 of the 10 conditions there was a mismatch between the lumi-

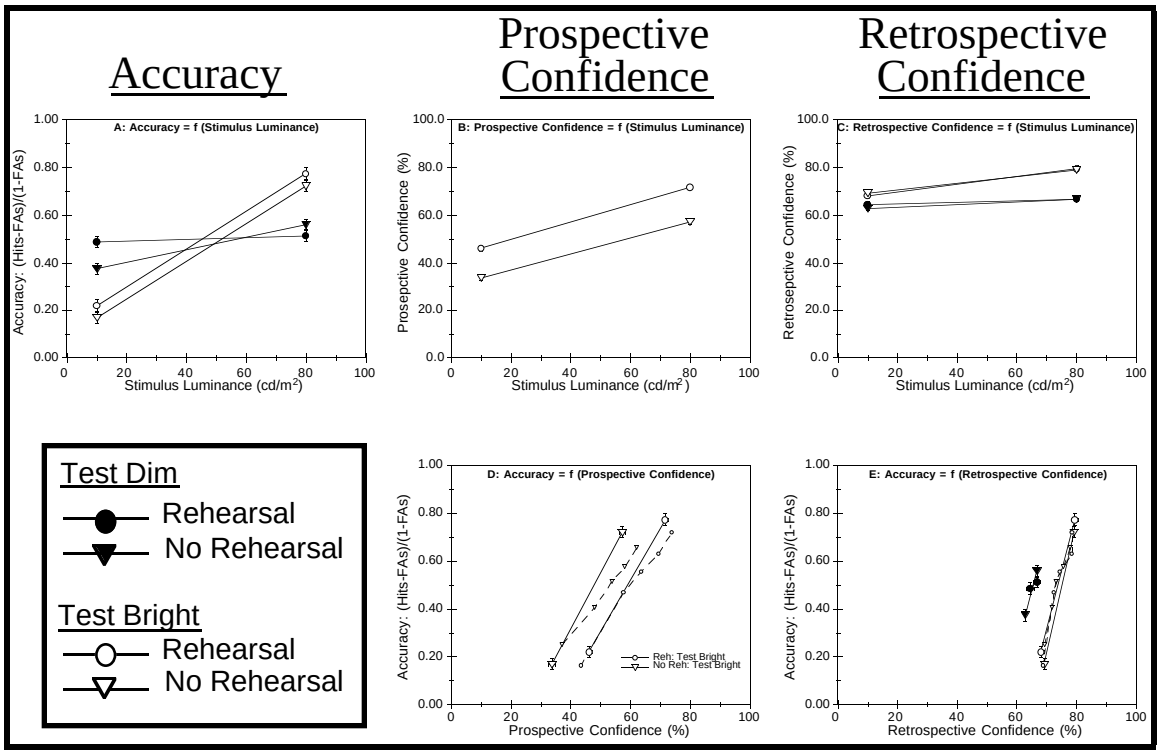

Figure 5: Experiment-3 data.

nance at study and the luminance at test. In Experiment 3 we systematically varied the study and test luminances, using the dimmest (10 cd/m^2) and brightest (80 cd/m^2) luminance conditions from Experiment 2. We created four conditions in which two study luminances (10 cd/m^2 and 80 cd/m^2) at study were crossed with the same two luminances at test. The resulting four conditions were crossed with the two rehearsal conditions to give 8 conditions in all.

Encoding specificity (Tulving & Thomson, 1973) predicts better performance when study and test luminances match, and if retrospective confidence and recognition judgments rely on the same information in memory we should find that confidence judgments are also highest when study and test luminances match. To anticipate, we find a dissociation between confidence and accuracy, such that conditions that produce decreases in accuracy also produce increases in confidence.

Methods

Experiment 3 used the same stimuli and equipment as Experiment 2, with the following exceptions:

Subjects

Subjects were 104 Indiana University undergraduate students who took part in the experiment for course credit. They were run in 24 groups of at least 3 subjects/group.

Stimuli and Design

Experiment 3 contained two levels of rehearsal, which were crossed with four levels of study/test luminance as described above.

Procedure

As in Experiment 2, Subjects were instructed to respond "old" to a face they thought they had seen in the study session *regardless of whether it was at a different luminance level*. Subjects were given several examples during the practice study and test sessions, and one example included a face shown dim in the practice study session and bright in the practice test session. Subjects who erroneously said "new" to this practice trial were informed of their mistake, and the experimenter then made sure that the subject understood that a target face shown at a different luminance level at test is still an old face.

Results and Discussion

For faces tested dim, the mean False-Alarm rate across subjects was 0.331, and the mean confidence rating for distracters was 59.60%. For faces tested bright, the mean False-Alarm rate across subjects was 0.246, and the mean confidence rating for distracters was 70.30%. The dim distracter false-alarm rate was used to correct conditions that were tested dim, and likewise the bright distracter false-alarm rate was used to correct conditions that were tested bright.

Figure 5 shows the main data for Experiment 3. As in Figures 3 and 4, the top three panels show accuracy, prospective, and retrospective confidence as functions of accuracy. The bright-tested conditions are duplicates of Experiment-2 conditions, and their data rep-

resented by open curve symbols, mimicking the curve symbols used in Figures 3-4. Data from the dim-tested conditions are represented by solid curve symbols. Because prospective confidence was given prior to manipulation of test luminance, test luminance cannot logically have any but a statistical effect on it; hence the Figure 5B data, are the average of the bright- and dim-tested pictures. For similar reasons, the prospective confidence-accuracy scatterplot is useful only as a replication of Experiment 2; hence Figure 5D shows confidence data averaged over only bright- and dim-tested pictures, while the accuracy data are for the bright-tested pictures only. Finally, for reasons to be described below, the Experiment-2 data are re-presented as dashed lines in Figures 5D and 5E. There are several noteworthy aspects of these data.

Bright test pictures: Replications

Consider the bright-tested pictures only (open curve symbols). There is close agreement between the Experiment-3 and Experiment-2 data. Study luminance has a positive effect on accuracy and on both kinds of confidence. As foreshadowed earlier, there is a small effect of rehearsal on accuracy which, in Experiment 3, occurs at both study luminance levels.

As in Experiment 2, rehearsal effect has a substantial effect on prospective confidence, but very little effect on retrospective confidence, as shown in Panels B and C. And, as in Experiments 1 and 2, the rehearsal and no-rehearsal curves fall atop one another in the accuracy-retrospective confidence scatterplot shown in Panel E.

Dim test pictures

As already noted, test luminance cannot have any but a statistical effect on prospective confidence. With respect to accuracy, a picture enjoys a clear advantage when it is tested at the same luminance in which is studied compared to a picture whose study and test luminances are different: Pictures studied dim are recognized better when tested dim, and pictures studied bright are recognized better when tested bright.

With respect to retrospective confidence, however, quite a different picture emerges: As indicated in Panel C, retrospective confidence for dim-tested pictures is decreased compared to retrospective confidence for bright-tested pictures. The accuracy-retrospective confidence scatterplot shown in Panel E confirms this: for a given level of accuracy, subjects are less confident for dim-tested than for bright-tested pictures.

Dissociations of Confidence and Accuracy

The Figure-5 data reveal a dissociation between confidence and accuracy. Consider a face that was studied *dim*. As is evident in Panels A and C, increasing the test luminance of a face studied dim *decreases* recognition accuracy (by 0.283 ± 0.047[3], averaged over rehearsal condition) but *increases* retrospective confidence (by an average of 4.75% ±1.39%). Subjects apparently believe (slightly) that a brighter test stimulus will help them, when in fact it causes a substantial decrease in accuracy.

[3]In this and similar usage, the number that follows the "±" refers to a 95% confidence interval.

The dissociation seen between retrospective confidence and accuracy for faces studied dim holds for all conditions. The two sets of state-trace curves in Figure 5E map out the state-spaces for items tested dim and tested bright. The two sets of curves do not fall on the same contour, allowing us to reject the single-dimension model. Subjects apparently pay too much attention to the nature of the test item and fail to take into account that in some cases a bright test item is actually *detrimental* to performance compared to a dim test item.

When confined to the Experiment-3 data, this analysis of the state-trace curves is somewhat limited because the state-trace curves do not overlap, and there are relatively few points along the Test Bright contours. It is for this reason that we superimposed the corresponding Figure-2 data, which more completely maps out the test-bright scatterplot. Note that the Bright-Bright condition is equivalent to the brightest study condition of Experiment 2, and that the Dim-Bright condition is equivalent to the dimmest study condition of Experiment 2. Thus Experiment 3 is a partial replication of Experiment 2. It is evident that there is a good correspondence between the replication points. It is also evident that the test bright contour does not connect the Bright-Dim or Dim-Dim points. Thus we are able to reject the single-source model for retrospective confidence judgments and accuracy. It appears that subjects inappropriately use information about the test item when making confidence ratings: they assume that a brighter face is better for recognition performance, when in some cases a bright test face actually decreases recognition performance.

Summary of Experiment 3

The state-trace plots comparing both prospective and retrospective confidence with accuracy disconfirm single-source models. When making prospective confidence judgments subjects pay too much attention to how an item was rehearsed. When making retrospective confidence judgments, subjects erroneously assume that a brighter test face will always lead to an increase in performance. This incorrect assumption leads to a dissociation between confidence and accuracy for faces studied dim and then tested bright. These data are consistent with a cue-utilization theory of metacognition in which analytic processes applied to the testing conditions can influence the retrospective confidence judgments. Thus while mnemonic processes may provide the primary basis for retrospective confidence and recognition judgments (as in Experiments 1 and 2), the additional analytic information about the testing conditions can overwhelm these processes and produce a surprising illusion of accuracy when in fact performance is quite poor. This demonstrates that In the absence of such changes in the testing conditions, the non-analytic mnemonic processes may provide the basis for much of the confidence ratings and produce strong correlations between confidence and accuracy, as described below.

Within-and Between-Subject Correlations

The traditional method of data analysis within both the metamemory and the eyewitness testimony litera-

ture has been to compute correlations between confidence and accuracy and determine the conditions under which subjects can predict their performance. A typical eyewitness testimony experiment asks each subject only a single question, which requires a between-subject correlation. Subjects in a metamemory experiment answer many questions, allowing a within-subjects analysis as well. For a variety of reasons discussed below, within-subject correlations are preferable, although consideration of between-subject correlations is often important in jury trials where two witnesses differ in their confidence levels. The confidence ratings are on a 5-point scale and accuracy is on a 2-point scale (correct or incorrect). This results in a situation in which tied scores reduce the value of the correlation, and for this reason we have chosen to compute *gamma* correlations, which ignore ties and consider only untied data. The resulting correlations are unbiased by the use of the coarse 5-point scale for confidence and 2-point scale for accuracy.

Note that there are many difficulties with computing and interpreting gamma correlations, which is one reason we propose State-Trace analysis as an alternative technique to assess the confidence/accuracy relation. For instance, to compute between subject gamma correlations, one has to assume that all subjects use the confidence scale in the same way. These and other assumptions may be unwarranted. However, for comparisons with other studies and to provide ties to the legal set-

TABLE 1

Within-Subject Gamma Correlations. Each subject gives 6 (Experiments 1 and 2) or 8 (Experiment 3) replications for each condition. The prospective confidence rating is paired with the subsequent accuracy (0 = miss, 1 = hit) and a gamma correlation is done on these 6 (or 8) pairs. These are then averaged, and here we report the Mean, Standard Error and N for each distribution of gamma correlations. In some cases a gamma correlation cannot be computed, resulting in different N's for different conditions.

	Prospective confidence vs. accuracy						**Retrospective confidence vs. accuracy**					
	Experiment 1						Experiment 1					
	Rehearsal			No Rehearsal			Rehearsal			No Rehearsal		
Duration	Mean	SE	N	Mean	SE	N	Mean	SE	N	Mean	SE	N
226 ms	0.381	0.068	97	0.259	0.069	100	0.411	0.066	98	0.170	0.099	100
321 ms	0.371	0.087	95	0.274	0.074	99	0.498	0.072	92	0.370	0.079	96
458 ms	0.223	0.098	84	0.365	0.085	93	0.619	0.052	79	0.460	0.078	89
653 ms	0.203	0.111	75	0.425	0.088	89	0.539	0.108	73	0.619	0.057	86
931 ms	0.266	0.108	66	0.276	0.084	87	0.716	0.071	67	0.660	0.074	90
	Experiment 2						Experiment 2					
	Rehearsal			No Rehearsal			Rehearsal			No Rehearsal		
Luminance	Mean	SE	N	Mean	SE	N	Mean	SE	N	Mean	SE	N
10 cd/m^2	0.266	0.068	91	0.095	0.097	94	0.097	0.092	88	0.000	0.079	91
16 cd/m^2	0.289	0.075	86	0.439	0.068	95	0.279	0.090	84	0.237	0.069	92
28 cd/m^2	0.420	0.093	86	0.247	0.064	90	0.570	0.060	86	0.512	0.060	86
47 cd/m^2	0.258	0.095	80	0.178	0.091	86	0.711	0.059	77	0.469	0.096	80
80 cd/m^2	0.249	0.109	67	0.355	0.084	76	0.761	0.064	66	0.697	0.059	70
	Experiment 3						Experiment 3					
	Rehearsal			No Rehearsal			Rehearsal			No Rehearsal		
Condtion	Mean	SE	N	Mean	SE	N	Mean	SE	N	Mean	SE	N
Dim/Brht	0.175	0.073	100	0.275	0.065	99	**0.013**	0.067	99	**0.047**	0.071	98
Dim/Dim	0.223	0.053	96	0.243	0.067	102	0.512	0.052	95	0.441	0.061	102
Brht/Dim	0.101	0.090	87	0.226	0.073	92	0.605	0.051	89	0.473	0.072	92
Brht/Brht	0.188	0.109	70	0.330	0.073	81	0.496	0.085	70	0.603	0.061	77

TABLE 2

Table 2. Between-Subject Gamma Correlations. The mean confidence rating and mean accuracy score for each subject is computed within each condition. This gives one pair per subject. A single gamma correlation is then computed for each condition

Prospective confidence vs. accuracy			Retrospective confidence vs. accuracy		
Experiment 1			Experiment 1		
Duration	Rehearsal	No Rehearsal	Duration	Rehearsal	No Rehearsal
226 ms	0.122	0.104	226 ms	0.047	0.111
321 ms	0.150	0.091	321 ms	0.170	0.069
458 ms	0.088	0.148	458 ms	0.177	0.257
653 ms	0.079	0.096	653 ms	0.278	0.233
931 ms	0.159	0.061	931 ms	0.418	0.256
Experiment 2			Experiment 2		
Luminance	Rehearsal	No Rehearsal	Luminance	Rehearsal	No Rehearsal
10 cd/m2	0.207	0.180	10 cd/m2	0.063	0.180
16 cd/m2	0.163	0.117	16 cd/m2	0.258	0.111
28 cd/m2	0.205	0.144	28 cd/m2	0.243	0.197
47 cd/m2	0.225	0.097	47 cd/m2	0.374	0.296
80 cd/m2	0.191	0.122	80 cd/m2	0.370	0.295
Experiment 3			Experiment 3		
Luminance	Rehearsal	No Rehearsal	Luminance	Rehearsal	No Rehearsal
Dim/Brht	0.174	0.278	Dim/Brht	**0.038**	**-0.044**
Dim/Dim	0.236	0.272	Dim/Dim	0.240	0.279
Brht/Dim	0.250	0.248	Brht/Dim	0.366	0.291
Brht/Brht	0.229	0.242	Brht/Brht	0.417	0.293

ting, we include these correlations below.

Within-Subject Correlations

Within-subject correlations for both prospective and retrospective confidence are shown in Table 1 for all three experiments. Of particular importance is consideration of the computability of the gamma correlation for longer stimulus durations (or brighter study conditions in experiments 2 and 3). As accuracy improves, a situation may exist that for a given condition, a subject may make no errors. As a result, gamma becomes uncomputable. The number of subjects that have computable gamma correlations is listed under the column headed by N. To see if the loss of these subjects systematically biased the gamma correlations, we combined the data from the two longest stimulus durations for Experiment 1. Performance in these conditions was close to asymptote for both confidence and accuracy, suggesting that we will not inflate the gamma correlations due to performance differences. This combination greatly reduced the number of missing subjects, but left the gamma correlations essentially unchanged. Thus we believe that the computable gamma correlations represent an unbiased estimate of the true gamma correlations.

For prospective confidence, the gamma correlations are around 0.2 - 0.4, and do not increase with increasing stimulus duration (or luminance) at study. We found that rehearsal did not affect the gamma correlation, which is surprising given the Delay-JOL effect described by Nelson and Dunlosky (1991). Our rehearsal condition might allow the face to persist in short-term memory, while the math-problem (no rehearsal) condition might eliminate the face from STM. Thus our rehearsal condition might be analogous to the immediate JOL, while the no rehearsal condition is similar to the delayed JOL condition. However, across all three experiments, no systematic difference between the two conditions is found. There are several possible reasons for this. First, our conditions probably do not map onto the original delays, in which the delayed JOL was made some 10 items later. Even 15 seconds of math problems may not dramatically affect the memory for a face. In addition, Keleman & Weaver (1997) found only modest increases in increase in the predictability of prospective confidence ratings for distracted vs con-

trol conditions. Thus the contents of short-term memory may not be all that detrimental to prospective judgments of future accuracy.

Retrospective confidence gamma correlations systematically increase with exposure duration or study luminance. There is also a significant effect of rehearsal, with rehearsal producing higher gamma correlations on average (M_D = 0.0764 ± 0.0595). One explanation for the increase in gamma with increasing exposure duration and rehearsal is the *Optimality Hypothesis*. We describe the Optimality Hypothesis below, but first we provide a description of the between-subjects correlations.

Between-Subject Correlations

Their are several issues that need to be taken into consideration when interpreting between-subject correlations. First, all subjects need to use the scale in an identical fashion. Thus one subject's 50% confidence rating, must be equivalent to all other subjects' 50% confidence ratings. We attempted to make this scale absolute by asking subjects to make ratings that indicated their confidence in how likely they would be to later recall the item. Despite this attempt to place the subjects on an absolute scale, subjects may have differed in their overall confidence level or their optimism in their memory abilities, making between subjects gamma correlations somewhat problematic to interpret. Nevertheless, to provide consistency with other eyewitness testimony work and links to the legal setting in which between-subject correlations are important, we present the between-subject gamma correlations as well.

The overall patterns in the between-subject gamma correlations mirror those from the within-subject gammas, except that the overall magnitudes are reduced. Prospective confidence correlations were all quite low, and did not improve with increasing exposure duration or rehearsal. Retrospective confidence ratings did improve slightly with exposure duration but not with rehearsal (M_D = 0.0454 ± 0.0499). The effects of exposure duration (or luminance) are consistent with Deffenbacher's Optimality Hypothesis, described below.

The Optimality Hypothesis

The best known existing hypothesis on which these between-subjects correlation data bear issues from a well known article by Deffenbacher (1980). Based on a meta-analysis of 45 experiments, Deffenbacher proposed what he referred to as the "optimality hypothesis," according to which the confidence-accuracy relationship increases with the quality of the circumstances surrounding the formation and retention of a memory trace. Deffenbacher offered evidence for the optimality hypothesis in the form of an observation that in experiments involving "non-optimal" conditions, the relation between confidence and accuracy tended to be small and nonsignificant, whereas in experiments involving "optimal" conditions, the relation tended to be larger and statistically significant. Although he did not so state explicitly, Deffenbacher appears to have been reporting between-subjects correlations only.

The Table 1 and 2 data provide no support for this hypothesis with respect to prospective confidence, but some limited support with respect to retrospective confidence: As circumstances improve in the form of increasing duration or luminance, retrospective confidence becomes a better predictor of accuracy, although prospective confidence does not. Improved circumstances in the form of greater rehearsal certainly do not appear to have a dramatic effect on any kind of confidence-accuracy correlation

CONCLUSIONS

The principle goal of the present work was to examine whether confidence ratings and accuracy judgments are based on the same information, and if not, to determine how different sources of information contribute to performance in the different measures. The data from Experiments 1-3 demonstrate that prospective confidence ratings and accuracy judgments *are* based on different sources of information. It is reasonable to suppose that, as depicted in the bottom panel of Figure 1, when making prospective ratings, subjects assume that rehearsal will help them more than it actually does. The data from Experiments 1 and 2 are consistent with a single-dimensional model for retrospective confidence and accuracy, although the data from Experiment 3 disconfirmed this model demonstrating at least one variable (test luminance) that affected retrospective confidence ratings and accuracy in different ways. In particular, subjects assumed that a bright test face would improve accuracy and thus they gave bright test faces higher confidence ratings overall. This misconception leads to a dissociation between retrospective confidence and accuracy: for faces studied dim, testing with a bright face lowers accuracy and increases confidence overall testing with a dim face.

Mechanisms of Prospective and Retrospective Confidence Judgments

As reviewed in the Introduction, a variety of mechanisms have been proposed for judgments of learning, feelings of knowing and other related metamemory judgments. The vast majority of data relevant to these mechanisms have used paired-associates, general knowledge questions, or other verbal materials. This approach has the advantage of allowing a cue to be associated with the target, to assess the degree to which the characteristics of the cue selectively influence a confidence rating while having no (or a detrimental) effect on recall. This approach has fairly clearly demonstrated the insufficiency of a trace access model in which the contents of memory are directly accessed. The question then becomes: what other information influences confidence ratings?

A variety of other factors have been shown to influence confidence and accuracy separately, and Koriat's Accessibility Hypothesis has been recently extended to include several different divisions of cues that are used when making metacognitive judgments (Koriat, 1997). Cues such as ease of processing are thought to be intimately tied to the stimulus, and are therefore described as intrinsic cues. Cue relating to the study conditions are thought of as extrinsic cues. Both of these are analytic in nature, in that they involve heuristics that sub-

jects overtly use to make their confidence rating (i.e. "I had longer to study that item, therefore I must have a better memory for it"). There is also a non-analytic, mnemonic set of cues that relate to information extracted from memory. The current state of the literature emphasizes how cues derived from the test item influence the confidence rating while having little or no influence on memory performance. For example, intrinsic cues are thought to have a greater influence on prospective confidence ratings than extrinsic cues

Face recognition introduces a number of complexities into this process. First, unlike cued-recall, no cues are associated with each face, although the testing conditions can be altered as in Experiment 3 to manipulate the probe used to access memory. Second, subjects must take into consideration that this is a recognition task with distractors and the possibility of an appreciable guessing rate. Thus the scale of the confidence ratings is somewhat difficult to interpret, making traditional calibration plots difficult to construct. Despite these limitations, the state-trace analysis of the present data provides for a number of conclusions about the mechanisms underlying metamemeric judgments of faces. Below we describe the information that we believe underlies prospective and retrospective confidence ratings.

Prospective Confidence Ratings

The state-trace analyses clearly demonstrate that prospective confidence ratings are based on information different from that used to make a recognition judgment. In particular, it appears that subjects believe that rehearsal will provide much more benefit than it actually does. This is perhaps not surprising, because when making prospective confidence ratings the subjects have just finished 15 seconds of either rehearsal (without the face being present) or arduous math problems. This was true whether stimulus duration or luminance was manipulated. This implies that subjects overestimate the benefits of rehearsal and underestimate the effects of either exposure duration or luminance. Rehearsal and exposure duration would be considered extrinsic cues by Koriat (1997), while luminance might be seen as an intrinsic cue. If this is the case, this would be surprising, since intrinsic cues are thought to have more effect on prospective confidence judgments than extrinsic cues while in Experiment 2 the reverse is true. This overestimation of the benefits of rehearsal with visual images suggests that subjects have a very poor ability to monitor the contents of their memory, and instead must rely on analytic strategies based on the study conditions.

Retrospective Confidence Ratings

The retrospective confidence ratings appear to track accuracy quite well, unless some variable (such as luminance) is manipulated at test. The dissociation between confidence and accuracy that results from faces studied dim and then tested bright demonstrates that subjects have an extremely poor ability to monitor the output of the memory process in that condition. Instead their confidence ratings reflect the belief that a brighter test face will always produce better accuracy, and this analytic analysis leads to an unjustified shift in their retrospective confidence ratings.

Overall these data support the view of metacognition that both prospective and retrospective confidence judgments are based on more information than simply the information that determines accuracy. In particular, the data support a model in which confidence ratings are computed not only on the basis of a direct access to information in memory, but through the analytic consideration of aspects of the study and test conditions (Begg et al., 1989; Koriat, 1993, 1995, 1997; Metcalfe et al., 1993; Reder & Ritter, 1992). Over time, these study conditions fade from memory, which enables retrospective confidence to accurately track accuracy in Experiments 1 and 2. This is consistent with Koriat's Accessibility Hypothesis (1995), in which subjects move from the use of analytic heuristics applied to intrinsic and extrinsic cues at study to a non-analytic process applied to mnemonic cues at test. However, analytic considerations still may play a role at test, when subjects believe (in some cases mistakenly) that a bright test face will always lead to improved performance. With regard to the Figure 1 two-process model, the strength dimension may correspond to what Koriat describes as mnemonic cues, or perhaps a combination of mnemonic and intrinsic cues. The certainty dimension is likely to correspond to the analytic mechanisms by which the study conditions are used to adjust the prospective confidence ratings. This results in a situation where subjects believe that rehearsal will help them much more than it does. What is so surprising in these data is how much the analytic operations can overwhelm the output of the recall mechanisms at test under poor memory conditions (Experiment 3). In addition, the large, unwarranted increase in prospective confidence caused by rehearsal at study demonstrates a lack of monitoring on the part of subjects of the contents of their own memories.

We have no good explanation for the pattern of data seen in the prospective and retrospective confidence gamma correlations. The optimality hypothesis can account for the increasing gamma correlation with increasing duration or rehearsal, but it is not clear why this would not translate to prospective confidence. Perhaps subjects, in overestimating the contributions of extra rehearsal time, neglect (or are unable) to monitor the true contents of their memory and therefore cannot make an accurate prediction for the subsequent recognition judgment.

Although we have proposed a two-state model to account for the dissociations of confidence and accuracy seen in Experiment 3, Clark (1995) has successfully fit confidence-accuracy inversions described by Chandler (1994) and Tulving (1981) with a single-process strength-based vector memory model (MINERVA 2, Hintzman, 1986). Clark assumed that accuracy in a force-choice task is based on the proportion of trials in which the match of the target to an item in memory is greater than the match of the distracter to an item in memory. This assumption is implemented by subtracting the distracter strength from the target strength on each trial: a positive number implies a correct choice. Confidence is related to the unsigned difference between the two strengths; a larger separation between the two strengths implies more discriminability between targets and distracters. Predictions on each trial can be captured by subtracting the distracter strength from the target

strength. As the variability of this target-strength-minus-distracter-strength distribution increases (as a result of the intervening pictures), accuracy goes down (more distracter strengths exceed target strengths due to the increased variability) and confidence goes up (more variability gives larger absolute differences and thus larger confidence values). Note of course that two dimensions are still required: Accuracy depends entirely on one dimension (strength difference) while confidence depends on both strength and the probability that the strength difference is positive.

While Clark's model is not a complete model of confidence judgments, it does explain the confidence and accuracy inversion. Clark was also able to demonstrate how similar formulations could account for Tulving's results: greater test-item similarity in the A/A' test produces lower variability, which increases accuracy but decreases confidence. Although this is a nice application of existing memory models to confidence judgments, it is not clear how such a formulation would apply to the Experiment 3 data without assuming metacognitive effects such as the assumption on the part of subjects that a brighter test stimulus will always lead to better performance.

Implications of Confidence and Accuracy Dissociations

The present work provides evidence dissociating both prospective and retrospective confidence judgments from recognition accuracy. Below we discuss both theoretical and applied implication of these findings.

At a theoretical level, the dissociations between confidence and accuracy extend support for a cue-utilization theory such as Koriat's (1997) Accessibility Hypothesis into the domain of face recognition. It is clear that while information from memory may contribute to both prospective and retrospective confidence ratings, manipulations that duplicate real-world situations such as changes in duration, luminance or rehearsal result in the use of extraneous information when making confidence judgments. The dissociation of retrospective confidence and accuracy demonstrates that subjects have a very poor ability to monitor the output of their memory processes when conditions at test differ from those at study.

In the applied domain, we might speculate on the implications of the confidence and accuracy inversion observed in Experiment 3. When a face is viewed first in a dark setting and then again in a bright setting, what does that change in luminance do to accuracy and confidence? Clearly the news is grim on both counts: Accuracy goes down and confidence goes up. However, we are hesitant to offer prescriptive advice to members of the legal community. After all, based on Experiment 3 we would have to recommend that eyewitnesses who perceives a crime at night should view a lineup in the dark! Clearly this is a solution that only a defense attorney could love.

This difficulty suggests a current research line. Aficionados of encoding specificity (Tulving & Thomson, 1973) will certainly not be surprised by the Experiment 3 accuracy findings, although the finding of Study Bright/Test Dim performance above Study Dim/Test Dim performance rules out encoding specificity as the only property underlying these data. It might be possible to find a moderate test luminance such that accuracy is unaffected and confidence does not suffer from the inflation seen with a bright test luminance. This hypothesis is currently undergoing rather intense scrutiny in our laboratory.

References

Bamber, D. (1979). State-trace analysis: a method of testing simple theories of causation. *Journal of Mathematical Psychology, 19*, 137-181.

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., & 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.

Bogartz, R.S. (1976). On the meaning of statistical interactions. *Journal of Experimental Child Psychology, 22*, 178-183.

Burke, D.M., MacKay, D.G., Worthley, JS., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and old adults? *Journal of Verbal Learning & Behavior, 6,* 325-337.

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory and Cognition, 3*, 273-280.

Clark, S. E. (1997). A familiarity-based account of confidence-accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 23*, 232-238.

Cutler, B. L. & Penrod, S. D. (1989). Forensically relevant moderators of the relation between eyewitness identification accuracy and confidence. *Journal of Applied Psychology, 74*, 650-652.

Deffenbacher, K. & Loftus, E. (1982). Do jurors share a common understanding concerning eyewitness behavior? *Law and Human Behavior, 6*, 15-30.

Deffenbacher, K. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relation? *Law and Human Behavior, 4*, 243-260.

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

Hintzman, D. (1986). "Schema Abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Jameson, K. A., Narens, L., Goldfarb, K., & Nelson, T. O. (1990). The influence of near-threshold priming on metamemory and recall. *Acta Psychologica*, (73), 55-68.

Kelemen, W. L., & Weaver, C. A. III. (1997). Enhanced metamemory at delays: Why do judgments of learning improve over time? *Journal of Experi-*

mental Psychology: Learning, Memory and Cognition, ,

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

Koriat, A. (1993). How do we know what we know? The accessibility model f the feeling of knowing. *Psychological Review, 100*, 609-639.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*, 311-333.

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.

Loftus, G. R. (1985). Picture perception: Effects of luminance on available information and information-extraction rate. *Journal of Experimental Psychology: General, 114*, 342-356.

Loftus, G.R. & Irwin, D.E. (1998). On the relations among different measures of visible and informational persistence. *Cognitive Psychology, 35*, 135-199.

Loftus, G.R. (1978). On interpretation of interactions. *Memory and Cognition, 6*, 312-319.

Metcalfe, J., Schwartz, B.L., & Joaquim, S.G. (1993). The cue familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 851-861.

Neil v. Biggers, 409 U.S. 188, 93, S. Ct. 375; 34 L. Ed. 2d 401 (1972).

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed JOL effect". *Psychological Science, 2*, 267-270.

Nelson, T., Gerler, D. & Narens, L. (1984). Accuracy of feeling-of-knowing judgments for predicting perceptual identification and re-learning. *Journal of Experimental Psychology: General, 113*, 282-300.

Pelli, D. G. and Zhang, L. (1991) Accurate control of contrast on microcomputer displays. *Vision Research, 31*, 1337-1350.

Read, J. D., Lindsay, D. S., & Nicholls, T. (1998). The relationship between confidence and accuracy in eyewitness identification studies: Is the conclusion changing? In C. P. Thompson Herrmann, D. J., Read, J. D., Bruce, D., Payne, D. G., & Toglia, M. P. (Ed.), *Eyewitness memory: Theoretical and applied perspectives* (pp. 107-130). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Reder, L. M., &. Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18*, 435-451.

Schwartz, B. & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1074-1083.

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

Sommer, W., Heinz, A., Leuthold, H., Matt, J. & Schweinberger, S. R. (1995). Metamemory, distinctiveness, and event-related potentials in recognition memory for faces. *Memory and Cognition, 23*, 1-11.

Spellman, B. A., & Bjork, R.A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*, 315-316.

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.

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352-373.

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior, 20*, 479-496.

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

Weaver, C. A. III, & Kelemen, W.L. (1997). Shifts in response patterns or increased metamemory accuracy? *Psychological Science, 8*, 318-321.

Wells, G., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence and juror perceptions in eyewitness identification. *Journal of Applied Psychology, 64*, 440-448.