Psychonomic Bulletin & Review
2000, 7 (1), 26-48

Accounts of the confidence–accuracy relation in recognition memory

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

and

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., Bothwell, Deffenbacher, & Brigham, 1987; Deffenbacher, 1980). There are at least two possible (nonexclusive) 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-subjects 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. 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 different sources of information in memory. 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/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 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 as-

---

The research reported in this article was supported in part by an NIMH grant to T.A.B. and an NIMH grant to G.R.L. We thank John Wixted and Stephan Lindsay for their helpful suggestions during the review process. Thanks also to John Dunlosky for many helpful discussions. Correspondence should be addressed to T. A. Busey, Department of Psychology, Indiana University, Bloomington, IN 47405 (e-mail: busey@indiana.edu).

sumption 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 (see, 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 on 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 has been described in detail by Bamber (1979). State-trace analysis has numerous general virtues, among the most important of which for the present research are that (1) 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 wherein some nonordinal interaction can be made to disappear—or a noninteraction can be made into an interaction—by a suitable monotonic but nonlinear transformation of the dependent variable (see, 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

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

Three Types of Correlations

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. 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. 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. Note that we will not actually compute correlations, but instead use properties of the scatterplots between two dependent measures to draw conclusions about the nature of the underlying sources of information.

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 and eyewitness identification domains.

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-subjects correlations are in the range of .25 to .40, 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. Most of these studies measure either prospective confidence ratings taken subsequent to a study period or FOK judgments that are made in response to a cued-recall failure during test.

Trace access theory (Burke, MacKay, Worthley, & Wade, 1991; Hart, 1967; King, Zechmeister, & Shaughnessy, 1980) posits a direct access to the contents of memory when confidence and recognition judgments are being made. Subjects monitor the contents of their memory and assess the different strengths of the stored items. This assessment becomes the basis for their confidence judgment. In addition, because stronger items tend to be better recalled, accuracy and confidence tend to covary. Thus, because confidence ratings and recognition rely on the same information, each predicts the other.

This view has been augmented by a variety of theories that include other sources of information that specifically affect confidence. For instance, making the test cue familiar through priming or other preexposure techniques can increase FOK judgments (Metcalfe, Schwartz, & Joaquim, 1993; Schwartz & Metcalfe, 1992), while making answers familiar through tachistoscopic preexposure increases recall of general knowledge questions without affecting FOK judgments (Jameson, Narens, Goldfarb, & Nelson, 1990). The ease of retrieval or perceptual fluency of an answer (correct or not) also contributes to retrospective confidence ratings (Kelley & D. S. Lindsay, 1993), so 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 prospective confidence ratings while leaving accuracy constant or even reduced (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Benjamin, Bjork, & Schwartz, 1998). 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 and Kelemen (1997) demonstrate that there is a real metacognitive improvement at a 5-min delay condition. One possible explanation for this improvement is that the delay eliminates transient short-term memory effects, so that items that remain in memory after a 5-min 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 theory posits both intrinsic cues (those that relate to the processing of the stimulus) and extrinsic cues (those that relate to the study conditions) that, through an analytic heuristic, can raise confidence (e.g., “I rehearsed an item so I should be able to remember it”). A second set of cues, known as mnemonic cues, are those derived from a search of memory and are more nonanalytic. A somewhat related view was proposed by Gigerenzer, Hoffrage, and Kleinbölting (1991), in which they proposed that observers learn how predictive a given cue is, and they tended to assume that the validity of a cue remains constant. However, the validity of a particular cue can change, especially in experimental settings, and as a result confidence could remain constant (because it is tied to the cue, not the match to memory) while accuracy might decrease.

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 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 implicated 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 ($G = .44$).

Research looking at the confidence–accuracy relation in face identification typically has measured retrospective confidence ratings taken subsequent to a mock lineup response. Earlier studies found little or no relation between confidence and accuracy (Bothwell, Deffenbacher, & Brigham, 1987; R. C. L. Lindsay, Wells, & Rumpel, 1981), which may not be surprising given that these studies typically measured between-subjects correlations in very restricted experimental conditions. Between-subjects correlations have the potential to be contaminated by the fact the some eyewitnesses will tend to be more confident than others; within-subjects correlations that obtain confidence ratings for a variety of questions posed to a single observer have the potential to eliminate the noise associated with criterion shifts across observers. More recent research has begun to suggest a much stronger relation between confidence and accuracy in face recognition. Read, D. S. Lindsay, and Nicholls (1998) conducted a number of between- and within-subjects 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 identified a variety of possible moderators of the confidence–accuracy relation, including immediate versus 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. In related work, D. S. Lindsay, Read, and Sharma (1998) demonstrated that the confidence–accuracy relation could be further improved by introducing variability into the study conditions. This variability is designed to simulate the fact that some eyewitnesses may find themselves in favorable encoding conditions (e.g., long viewing duration, good lighting) while others will not.

A number of researchers have looked at the use of retrospective confidence ratings within the context of testing models of memory using signal detection analysis. Although basic signal detection theory (SDT) does not specify the placement of decision criteria that determine confidence ratings, some extensions of SDT such as ideal observer models do predict where and when confidence ratings should shift. These shifts are often studied in the context of the *mirror effect* (Glanzer & Adams, 1985, 1990; Wixted, 1992), which is a general finding that as conditions become more favorable for memory performance, false alarms tend to decrease and hits tend to increase. Under some circumstances this improvement results from shifts in the criterion cutoffs in a signal detection model (Stretch & Wixted, 1998a, 1998b), although differences in the locations of target and distractor distributions can also account for the mirror effect. These models will be discussed within the context of Experiment 3.

To summarize, a number of factors other than direct memory access have been identified as the basis of confidence judgments made in response to the 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. In the present experiments, we explored the possibility that confidence and recognition judgments may in fact be based on different sources of information and then 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 prevented from rehearsing it.

Three dependent variables were measured in each experiment. A prospective confidence rating was obtained after the 15-sec rehearsal/nonrehearsal 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 judgment.

THEORY

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

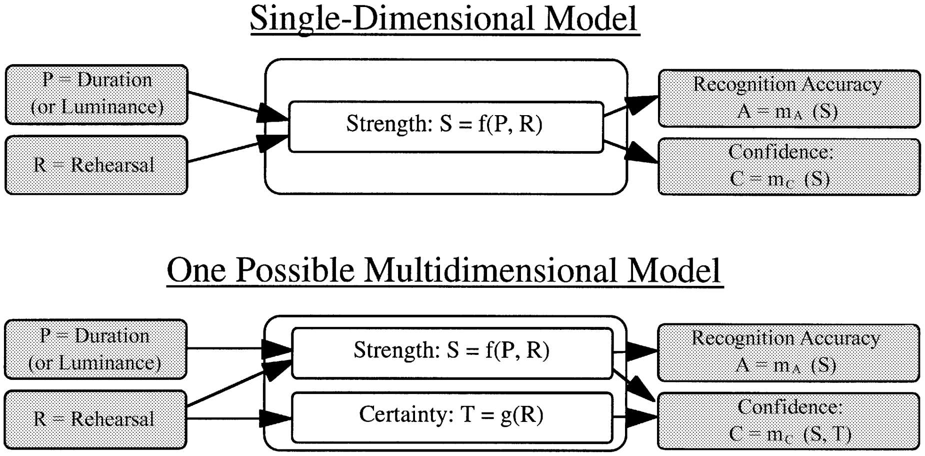

Figure 1. Two models of the confidence–accuracy relation.

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, e.g., if one rejects a null hypothesis in an analysis of variance, 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, R. C. L. Lindsay, 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 flavor of this dimension from the perspective of the metacognitive literature, T might include probe-related cues such as probe familiarity due to preexposure 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 had been used instead) and making specific, although somewhat arbitrary, choices for the functions shown in Figure 1—in particular, the monotonic function f relating S to P and R, and the monotonic functions, m_C and m_A relating confidence and accuracy to strength.[1]

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 in the top and bottom rows 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

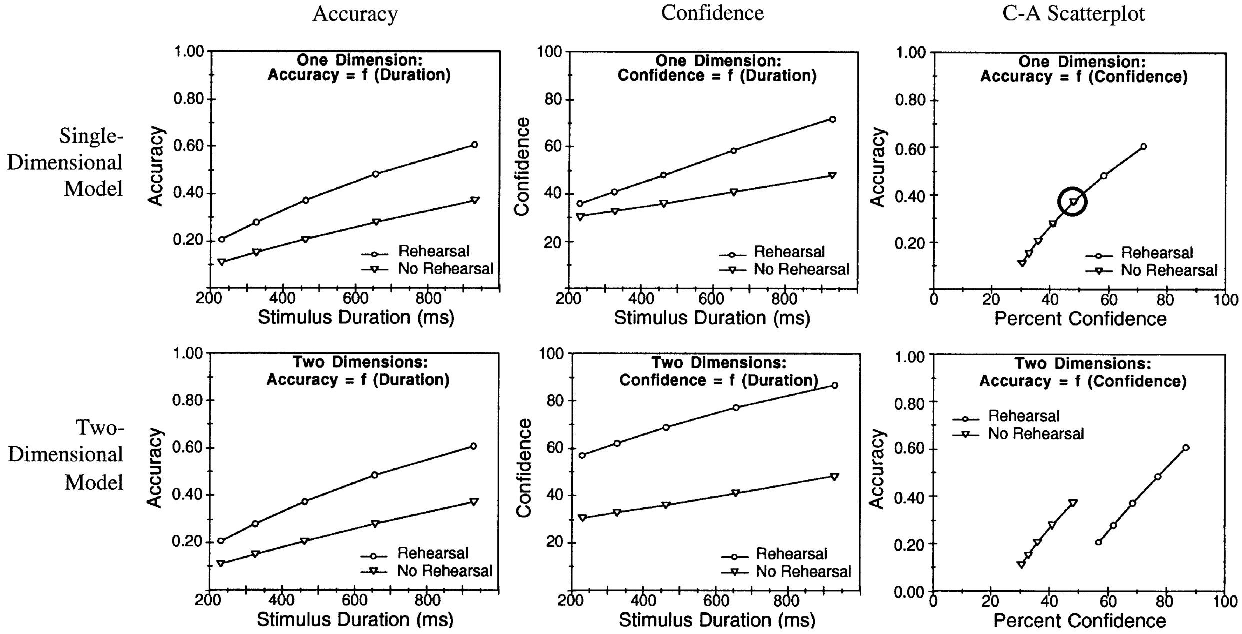

Figure 2. Predictions of the two models.

asserting them to be replications of one another—and yet one was generated by a single-dimensional model, while the other was generated by a two-dimensional model.

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. Bamber (1979) referred to these scatterplots 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 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-msec rehearsal condition, while the triangle corresponds to the 930-msec no-rehearsal condition. Because these two physically distinct conditions produce the same level of accuracy (.372), they must, by the single-dimensional model of Figure 1, have produced the same value of strength (specifically, $S = m_A^{-1}$ [.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 $m_C(S) = m_C[m_A^{-1}(.372)] = 48.1\%$. 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.

Note that there exists a special case in which two dependent measures such as confidence and accuracy might be related to a single underlying dimension (e.g., strength), but by functions that are not monotonic with that underlying dimension. This model will produce a discontinuous state-trace plot. This situation usually requires assumptions that are difficult to accept, such as "increasing strength should simultaneously increase accuracy and decrease confidence," and therefore this model tends to make less sense than one in which multiple dimensions exist. Moreover, additional information is usually required to make the nonmonotonic functions reasonable (such as the category of a particular item), and therefore this model is essentially equivalent to a model in which the two dependent measures rely on more than one source of information. For example, the two sources of information might be the underlying strength dimension and the category information about the test item, which tells the subject how the strength dimension should be interpreted when making confidence judgments for an item in this category.

Prediction Summary

A finding that the rehearsal and no-rehearsal scatterplot curves fall atop one another confirms a single-dimensional model, with its two assumptions of a unidimensional memory representation and monotonic translations of the value along this dimension to the two dependent measures. 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: first, a perceptual variable (stimulus duration in Experiment 1 and stimulus luminance in Experiments 2 and 3) and second, amount of poststimulus 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 the independent variables we have used 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 postevent 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.

Method

The methods for Experiments 1–3 are similar; we describe the general methodology here, and describe the method 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 one third of the men had facial hair. The faces were digitized and displayed on a 21-in. 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 five 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 to 930 msec in logarithmically equal steps. There were two levels of the rehearsal manipulation: For 15 sec following stimulus offset, subjects either silently rehearsed a face (without, of course, being able to see it) or performed math problems as a distractor 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 each other with new sets of faces.

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

1. A 400-msec warning tone occurred beginning 500 msec 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 sec 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 the 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 indicating 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. The subjects were told that half of the pictures were old and that 0% confidence was associated with pure guessing.

As indicated, this study–test sequence was repeated twice, thereby resulting in six replications per condition per subject. The experimental session was preceded by a practice study session in which three sample study trials and six 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 (A) is based on both the hit rate and the false alarm rate and is computed via the equation, $A = (H-FA)/(1-FA)$, where H and FA are hit and false alarm probabilities. The high-threshold model that implies that 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, e.g., happens with d' when the hit or the false alarm rate is either 0 or 1.0).

Results and Discussion

The mean false alarm rate across subjects was .266, and the mean confidence rating for distractors was 69.25% (Figure 3). Figure 3, which is similar to other data figures to be presented in this article, contains seven panels. The left four panels correspond to what we have referred to as the *standard data*: They show, respectively, accuracy, prospective confidence, retrospective confidence, and retrospective confidence conditioned on an "old" response, all graphed 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 error bars are smaller than the size of the curve symbols. The right three 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 panels embedded within panels A–C, E, and F show theoretical predictions that were generated using the functions described in Note 1 to replace the monotonic functions that constitute 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.

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. This is consistent with D. S. Lindsay et al. (1998), but perhaps unexpected given the null relation between confidence and accuracy reported by R. C. L. Lindsay et al. (1981).

Confidence versus accuracy: Scatterplot data. The scatterplots relating accuracy to confidence are shown in the three right 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 (Figure 3, panel E), the rehearsal curve falls to the right 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") that is positively affected by both duration and rehearsal. Prospective confidence, however, is determined by two dimensions (e.g., what we have referred to as memory strength, S, and memory certainty, T). 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 subjects to provide an estimate of the benefits of rehearsal. Subjects assume that rehearsal will produce better recognition, and they therefore inflate their confidence ratings. 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 (Figure 3, panels F and G), the rehearsal and no-rehearsal curves fall atop each other. This confirms a single-dimensional model and disconfirms multidimensional models.[2] 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. The confidence conditioned on saying "old" data demonstrate a similar result. Note that rehearsal has only a modest (but signifi-

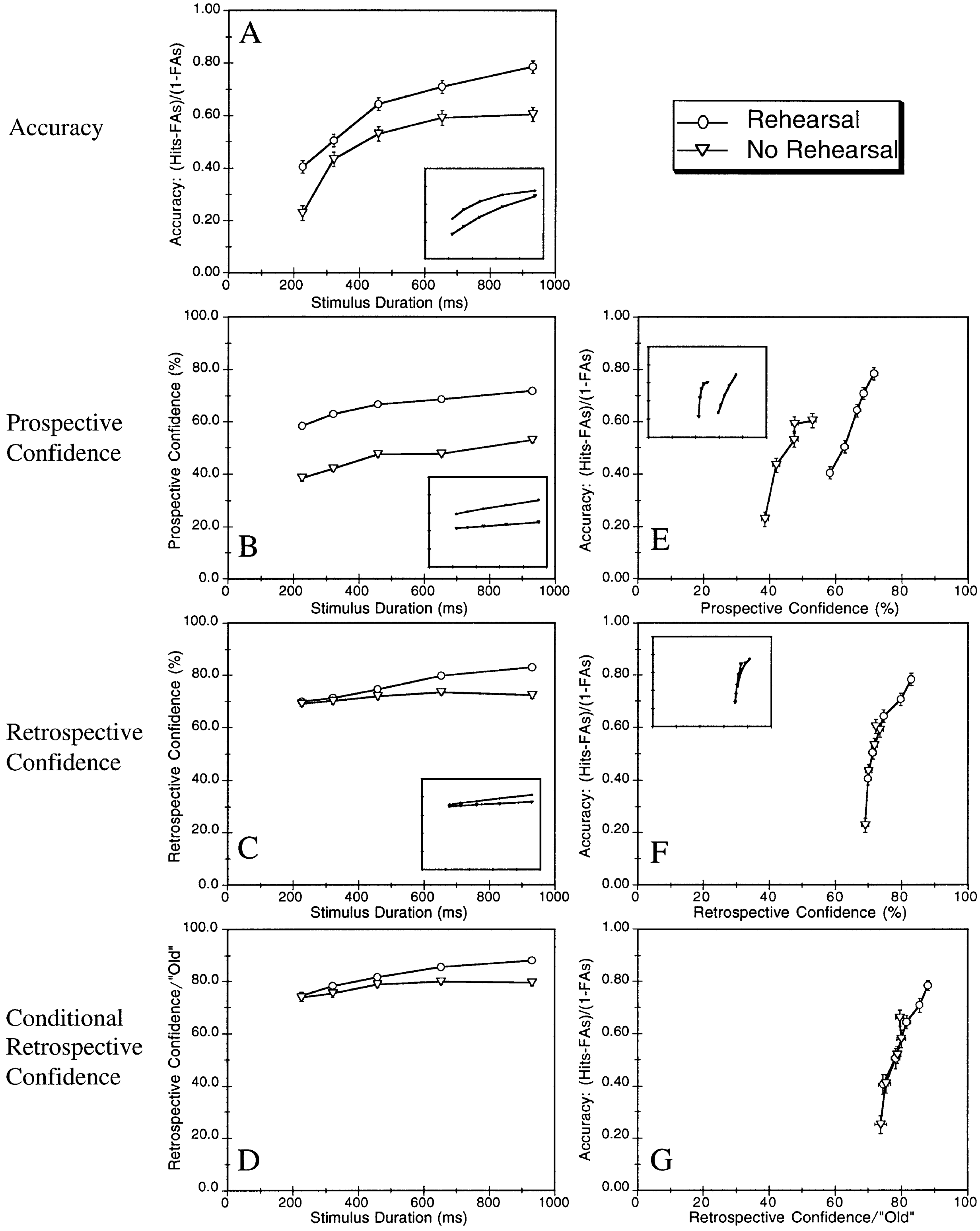

Figure 3. Experiment 1 data. (A–D) Accuracy, prospective confidence, retrospective confidence, and retrospective confidence conditioned on saying "old" as a function of stimulus duration. (E–G) State-trace plots of prospective confidence, retrospective confidence, and conditional retrospective confidence against accuracy.

cant) effect on accuracy, which will tend to force the two state-trace plots together. However, we could have detected a dissociation in the state-trace plot given our extremely high statistical power (reflected by the extremely small standard error bars). In addition, what is remarkable is the large dissociation between confidence and accuracy seen in the prospective confidence ratings and the lack of a dissociation seen in retrospective confidence ratings.

One issue that arises in the application of state-trace analysis is that specification of a single-dimensional model, while precise, may not correspond to what in the literature is typically viewed as a single-dimensional model. For example, accuracy is typically thought of as based on components such as familiarity combined with perceptual fluency to give a sense of prior occurrence. This may be viewed as a single-dimensional model despite the two components, because the two components combine into a single dimension prior to determining accuracy, resulting in the loss of information about the original values of the individual components. However, if these components have different decay rates, and if they affect confidence and accuracy differently, then a multidimensional model is implied. This is because in order for the components to have different decay rates, separate information must have been retained in memory about each component. An extended discussion of those models that are and are not multidimensional by the state-trace definition can be found in the Appendix.

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, whereas 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 (Kelemen & Weaver, 1997; Nelson & Dunlosky, 1991; Thiede & Dunlosky, 1994). 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.

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 (see, e.g., Loftus & Ruthruff, 1994). The major purpose of Experiment 2 was to generalize the Experiment 1 findings by replicating them using a different environmental variable.

Method

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 five 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.[3]

All stimuli were presented for 1,350 msec 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 .265, and the mean confidence rating for distractors was 68.95% (Figure 4). The left four panels indicate that luminance in Experiment 2 acted very much like duration did 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 is not replicated 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 right side 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 both retrospective confidence and conditional retrospective confidence, the two scatterplots fall atop each other. As with Experiment 1, the effects of rehearsal were modest but significant, indicating that we could have detected a dissociation between retrospective confidence and accuracy.

Summary of Experiments 1 and 2

The state-trace plots comparing prospective confidence with accuracy reveal that prospective confidence ratings and recognition judgments in this task are based on different sources of information in memory, or are based on nonmonotonic functions of a common source of information. 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

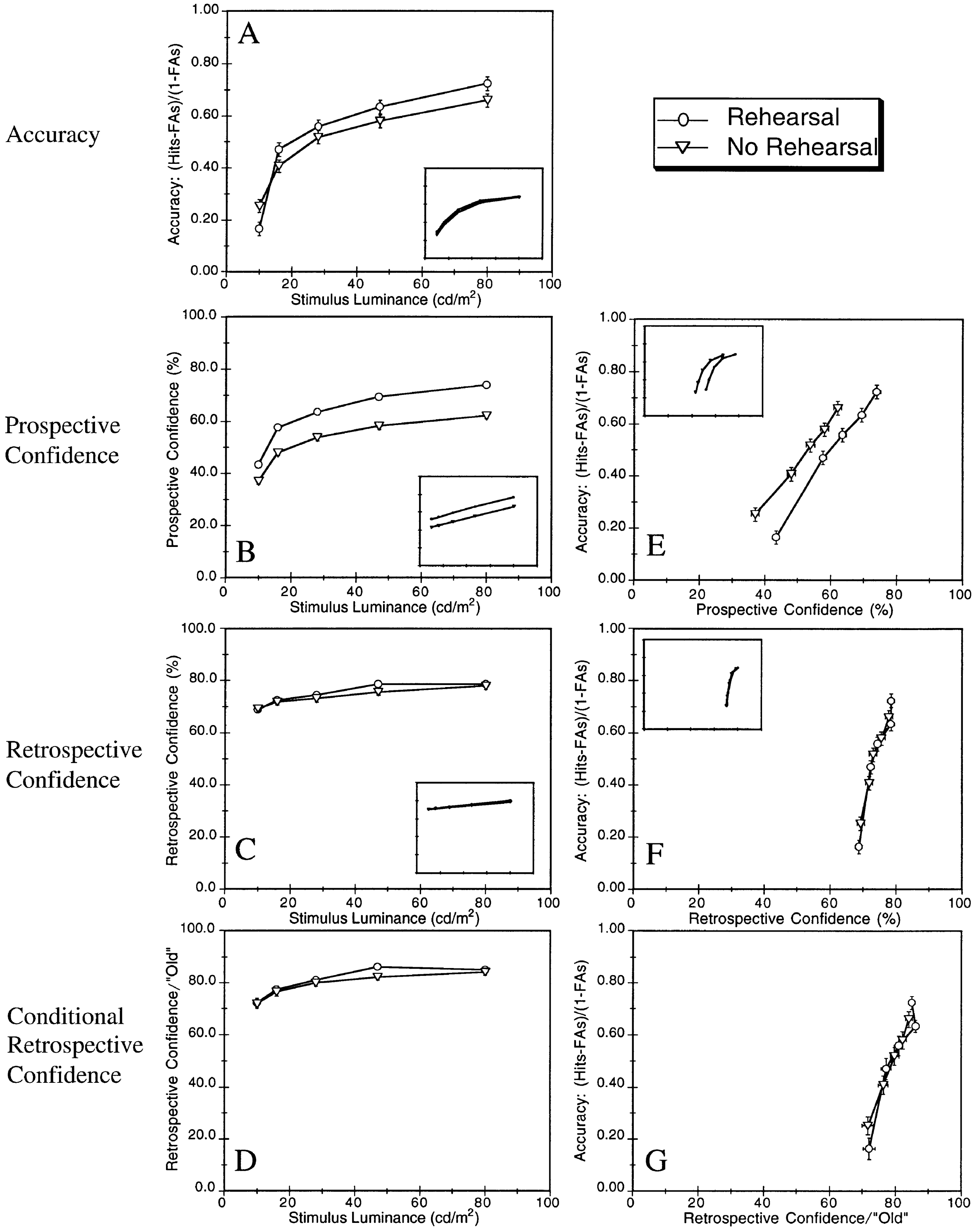

Figure 4. Experiment 2 data. (A–D) Accuracy, prospective confidence, retrospective confidence, and retrospective confidence conditioned on saying "old" as a function of stimulus luminance. (E–G) State-trace plots of prospective confidence, retrospective confidence, and conditional retrospective confidence against accuracy.

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 prospective confidence judgments are being made. This demonstrates that although a covert retrieval attempt might contribute to prospective confidence ratings (see, 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. As with any attempted dissociation between two dependent variables, it may be that we have not found an experimental situation that provides a dissociation between retrospective confidence and accuracy, and below we demonstrate that such a dissociation can be found. Thus the results of Experiments 1 and 2 illustrate those situations in which retrospective confidence and accuracy can be expected to be based on the same sources of information.

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, which assume that studied and nonstudied 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 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, Ferguson, and R. C. L. Lindsay (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 subjects' 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 poststudy 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 he/she 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 luminance 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 eight 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 found a dissociation between confidence and accuracy, such that conditions that produce decreases in accuracy also produce increases in confidence.

Method

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

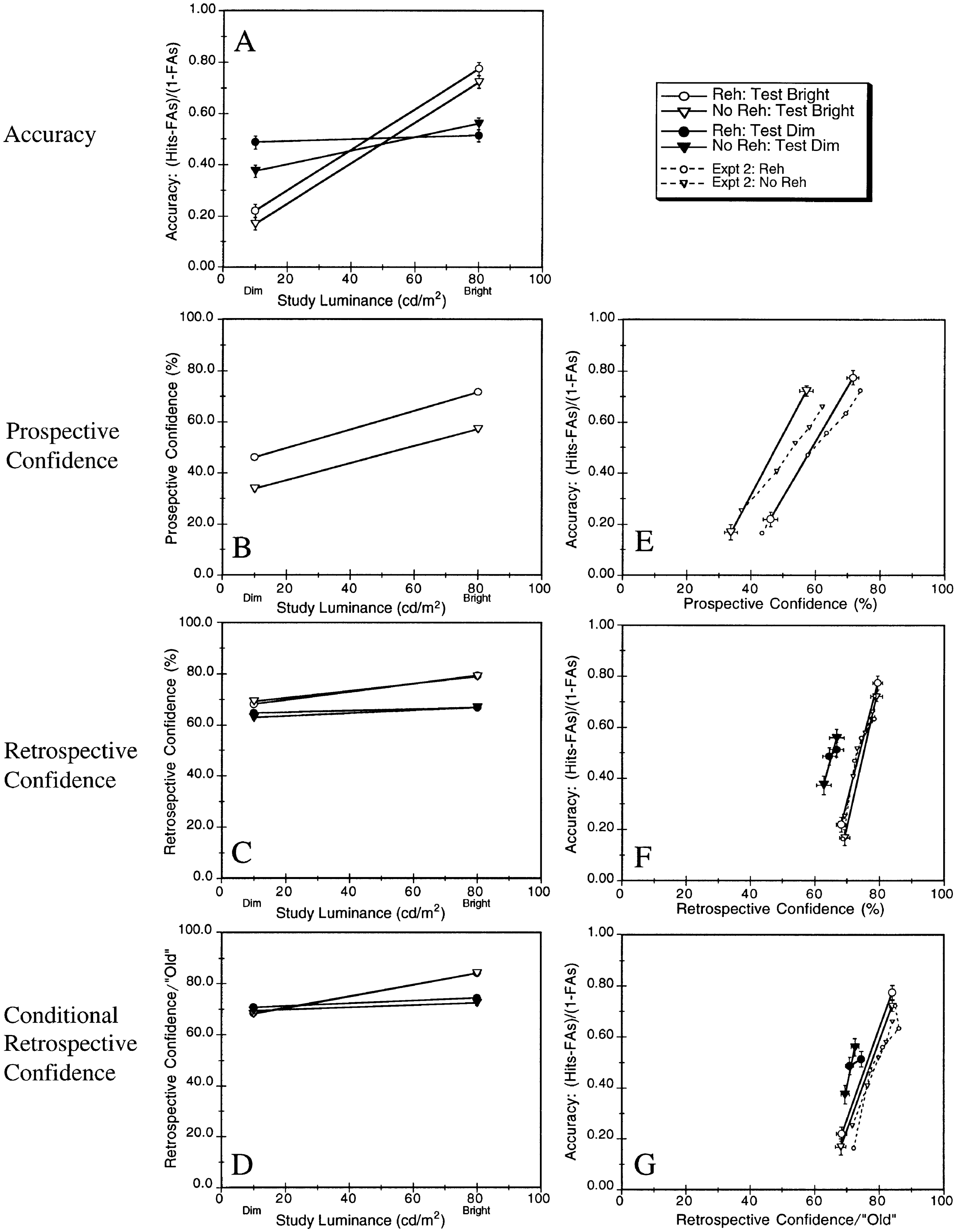

Figure 5. Experiment 3 data. (A–D) Accuracy, prospective confidence, retrospective confidence, and retrospective confidence conditioned on saying "old" as functions of study luminance and test luminance. (E–G) State-trace plots of prospective confidence, retrospective confidence, and conditional retrospective confidence against accuracy.

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 per 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 such subjects 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 .331, and the mean confidence rating for distractors was 59.60%. For faces tested bright, the mean false alarm rate across subjects was .246, and the mean confidence rating for distractors was 70.30%. The dim distractor false alarm rate was used to correct conditions that were tested dim, and likewise the bright distractor 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 left four panels show accuracy, prospective confidence, retrospective confidence, and conditional retrospective confidence as functions of study and test luminance. The bright-tested conditions are duplicates of Experiment 2 conditions, and their data are represented by open curve symbols, mimicking the curve symbols used in Figures 3 and 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 had 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 5E 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 5E–5G. 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 had a positive effect on accuracy and on both kinds of confidence. As foreshadowed earlier, there was a small effect of rehearsal on accuracy that in Experiment 3 occurred at both study luminance levels.

As in Experiment 2, rehearsal effect had a substantial effect on prospective confidence, but very little effect on retrospective confidence, as is shown in panels C, D, F, and G. And, as in Experiments 1 and 2, the rehearsal and no-rehearsal curves fall atop each other in the accuracy–retrospective confidence scatterplots shown in panels F and G.

Dim test pictures. As already noted, test luminance cannot have had 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 as opposed 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 pattern emerges: As indicated in panels F and G, retrospective confidence for dim-tested pictures decreased relative to retrospective confidence for bright-tested pictures. The accuracy–retrospective confidence scatterplots shown in panels F and G confirm this: For a given level of accuracy, subjects are less confident for dim- than for bright-tested pictures.

Dissociations of Confidence and Accuracy

The state-trace plots shown in Figure 5 reveal a dissociation between retrospective confidence and accuracy. The two sets of state-trace curves in Figures 5F and 5G map out the state spaces for items tested dim and tested bright. In both graphs, 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 relative to a dim test item. To see this, consider a face that was studied *dim*. As is evident in panels A and F, increasing the test luminance of a face studied dim *decreases* recognition accuracy (by .283 ± .047,[4] averaged over rehearsal condition).

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, for both the marginal retrospective confidence data and the confidence data were conditioned on an "old" response. 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.

Relation to Cue Utilization Theories

The metacognitive strategy just described falls into a larger class of cue utilization theories that have been proposed to account for choice and confidence judgments. For example, Gigerenzer et al. (1991) proposed that subjects learn the utility of a particular cue (or derive this utility based on assumptions about how memory works) and make their decision and confidence judgments on the basis of the perceived utility. In Experiment 3, subjects apparently assumed that bright test faces would always produce better performance, when, in fact, in our dim study–bright test condition, this was not true. The cue utilization theories constitute a single-dimensional model: Choice and confidence are both based on the perceived utility of the cue. However, we are investigating the *accuracy*–confidence relation, since this is usually of interest to the applied fields of face recognition and eyewitness identification. What Experiment 3 demonstrates is that (1) subjects do make these assumptions about the cues in recognition memory paradigms, which therefore extends the cue utilization approach to face recognition; and (2) subjects are insensitive to the poor quality of the underlying information when making confidence judgments in the dim study–test bright condition. If subjects had been sensitive to the poor quality of the match between the test item and the contents of memory, they could have overcome the misleading information suggested by the perceived utility of the cue.

The cue utilization approach, and in particular the probabilistic mental model theory proposed by Gigerenzer et al. (1991), suggests that subjects learn the utility of particular cues and use this information in conjunction with information retrieved from memory when making prospective confidence judgments. If this is indeed the case, then subjects have a much more optimistic view of the benefits of rehearsal than is warranted by its effects on accuracy. This can be seen in the data in all three experiments, where prospective confidence and accuracy show a dissociation on the basis of rehearsal. This in part may reflect the difference between verbal materials and our image-based face. Subjects may have more experience in the educational system rehearsing verbal materials, and their success in this domain may be translated to the face recognition domain in the form of overconfidence in the benefits of rehearsal relative to other independent variables that affect the quality of the initial information.

Signal Detection Accounts of Experiment 3

The discontinuous state-trace plots for retrospective confidence shown in Figure 5 suggest that confidence and accuracy are based on different memory representations (or are nonmonotonic expressions of a single rep-

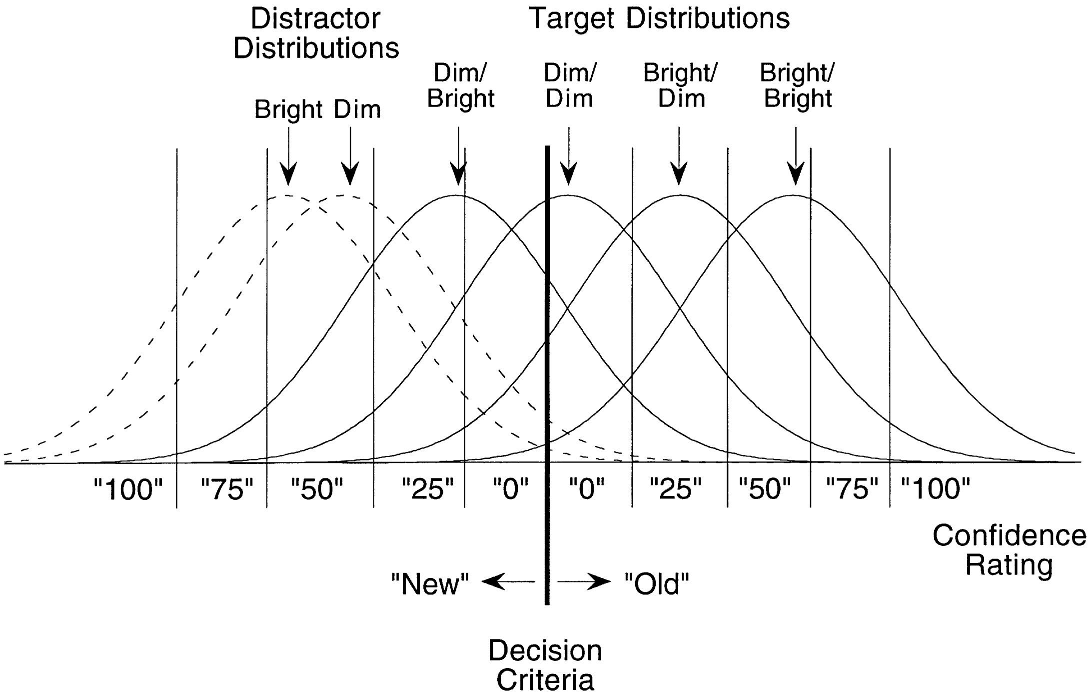

Figure 6. Representation of Experiment 3 conditions according to a signal detection theory account. Faces shown in different conditions are arranged along a single dimension such as familiarity, with distractor faces shown as dashed lines (for simplicity we combine across the rehearsal manipulation). The scale of this hypothetical representation has been expanded for clarity between the different conditions.

resentation) and require a multidimensional model by our definition. SDT provides one multidimensional model in which the discontinuous state-trace plots result from the existence of two distractor distributions (the dim and bright distractors).[5] This may produce a situation with a single memory representation, but because subjects do not have full knowledge of the locations of the two distractor distributions, they cannot adjust their confidence criteria optimally and discontinuous state-trace plots result. In this section we explore a signal detection model as one possible multidimensional model.[6] Note that tests of a signal detection account typically require a distractor distribution, and because we do not get prospective confidence ratings for distractors, we cannot use SDT to account for our prospective confidence data, nor can we use it to compare prospective and retrospective confidence judgments.

Within a signal detection framework, target and distractor distributions are represented along a single dimension (usually termed *familiarity* or *memory match*), and confidence ratings result from the application of criteria along this axis. Figure 6 shows a hypothetical representation of the Experiment 3 conditions, with the scale enlarged to enhance the differences between the conditions. For simplicity, we combined across rehearsal conditions, which did not have a large effect on the retrospective confidence or accuracy data. The four target conditions and two distractor conditions are represented by distributions along a unidimensional familiarity scale. The subject imposes a decision criterion such that if a test face engenders a feeling of familiarity that is greater than the decision criterion, he/she responds "old"; he/she responds "new" otherwise. After making this decision, subjects use the distance between the obtained familiarity and the decision criterion to make a confidence rating. This is done by placing four confidence criteria on either side of the decision criterion. The locations of the confidence criteria are determined by the subject; SDT does not specify their locations unless additional assumptions are made. For example, an optimal decision rule in a situation with only one target and one distractor distribution would be to place the decision criterion at the point at which the two distributions cross.

Experiment 3 contains two types of distractors (bright and dim test faces), and this manipulation may create two distractor distributions. If subjects adopt a single set of decision and confidence criteria for all conditions, then a shift in the distractor distributions could affect mean confidence and accuracy differently, resulting in a discontinuous state-trace plot. For example, shifting both a target and a distractor distribution leftward by an equal amount (e.g., all bright test faces, including bright distractors and targets tested bright) would not change accuracy for faces tested bright, but would decrease the conditional confidence for the target conditions. This would produce a discontinuity in the state-trace plot. However, if subjects could adjust their confidence criterion optimally, then confidence would continue to track accuracy, as in Experiments 1 and 2.

This example illustrates the possibility that the two distractor distributions, when combined with a single set of criterion cutoffs, could produce the discontinuous state-trace plots seen in Figure 5. Finding that such a model accounts for the data would not invalidate the state-trace conclusions, because the functions that map the single memory dimension (in this case, familiarity) to accuracy and confidence are not monotonically related. Accuracy is determined by the distance between the target and relevant distractor distribution, while confidence is a function of the distance from the rated familiarity to the decision criterion. In addition, two sources of information are now required to determine confidence and accuracy: the location along the familiarity axis and the location of the distractor distribution. These two sources of information are consistent with the claim that this version of SDT is a multidimensional model.

Finding a good fit of a signal detection model extends the state-trace analysis by identifying the nature of the decision rules that map the memory representation onto confidence and recognition judgments. However, as we will show below, there are aspects of the data that this version of SDT cannot account for, although these are entirely consistent with our metacognitive explanations for our discontinuous state-trace plots.

Receiver-operating characteristic (ROC) functions were constructed for Experiment 3 in the following fashion. For simplicity we combined across the rehearsal/no-rehearsal manipulation, because, as noted, this manipulation did not produce large effects in either retrospective confidence or accuracy judgments. On each trial we combined the old–new decision with the five confidence levels to produce a cumulative distribution with 10 values. The response probabilities in each confidence bin were use to construct ROC functions by calculating the cumulative probabilities starting with the 100% old category and working to the 100% new category. These distributions were converted to probabilities and plotted against the relevant distractor distribution to produce the functions shown in Figure 7. The probabilities have been converted to z scores, which produce linear functions if the underlying distributions are Gaussian. Theoretical predictions were generated by assuming that the underlying distributions were Gaussian and that a single set of confidence criteria is applied to all targets and distractors. The bright distractor distribution was fixed at zero and the means of the other five distributions in Figure 6 were allowed to vary as free parameters. We first consider an equal-variance case, although this model is somewhat simplistic in light of much of the literature that suggests that the target distribution has greater variability than the distractor distribution. However, the failures of this model serve to illustrate particular aspects of the data that we wish to highlight, and in fact the model will fail in a way that is not typically seen in recognition memory.

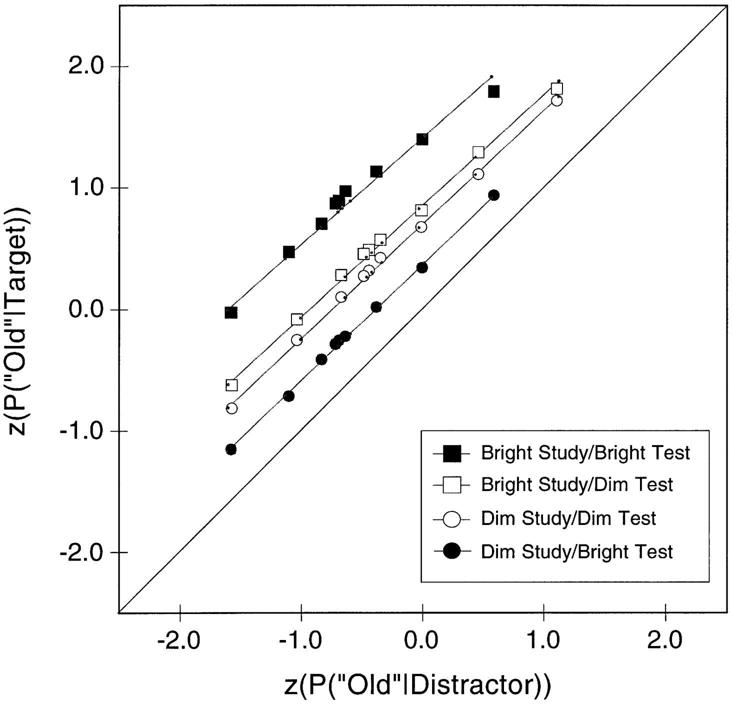

Figure 7. Receiver-operating characteristic functions for the four conditions of Experiment 3, along with the fit of an unequal-variance signal detection model. Small dots along each line correspond to predicted probabilities according to the model for each confidence bin.

In this version the standard deviations of all Gaussian distributions were fixed at 1.0. The decision criteria and eight confidence criteria were allowed to vary as free parameters, which produces a model with 14 free parameters that will fit 54 data points. The free parameters were adjusted according to a maximum likelihood measure (G^2) using the Solver function in Excel.

The fits of an equal-variance version of SDT were reasonable, with some notable exceptions that will prove problematic for this version of SDT. In particular, the model has difficulty accounting for confidence for the bright test faces. The model must assume that bright distractors engender less familiarity than dim distractors. As long as the confidence criterion remains fixed for the two distributions, a shift in a distribution (either left or right) results in mean confidence going up on one side and down on the other. As long as the variance of the distribution is held constant, a left or right shift cannot simultaneously increase mean confidence for both "old" and "new" responses.

The Experiment 3 data violate the above relation, as shown in Table 1. Bright distractors have significantly greater mean confidence for *both* "old" and "new" responses, which is inconsistent with an equal-variance signal detection model. This model predicts higher confidence for dim distractors when subjects say "old." The signal detection model could potentially account for the distractor confidence values by assuming that bright and dim distractor distributions have different variances. For example, if the bright distractor distribution increased its variance as it shifted leftward, it would place more responses in the higher confidence regions and therefore account for the increase in mean confidence for both "old" and "new" responses. To test this, we fit a version

Table 1
Mean Confidence (%) for Bright and Dim Distractors, Conditioned on Whether Subjects Say "old" or "new"

Distractor Condition	Data "old"	Data "new"	Equal-Variance Theory "old"	Equal-Variance Theory "new"	Unequal-Variance Theory "old"	Unequal-Variance Theory "new"
Bright	64.3 (62.3–66.3)	71.7 (70.7–72.8)	59.3	70.0	61.8	71.6
Dim	58.7 (56.9–60.4)	59.5 (58.5–60.4)	63.4	64.0	58.9	60.5

Note—Values in parentheses are 95% confidence intervals.

Table 2A
Model Parameters for the Signal Detection Fits of Experiment 3, Fixed Standard Deviation

	Distractor		Target Condition			
	Bright	Dim	Dim/Bright	Dim/Dim	Bright/Dim	Bright/Bright
Mean of distribution	0.000	0.331	0.407	0.991	1.154	1.618
Standard deviation	1.000	1.000	1.000	1.000	1.000	1.000

Confidence Criterion	Value
100-New	—
75-New	−0.630
50-New	−0.040
25-New	0.364
0-New	0.632
0-Old	0.706
25-Old	0.741
50-Old	0.897
75-Old	1.214
100-Old	1.723

Table 2B
Model Parameters for the Signal Detection Fits of Experiment 3, Variable Standard Deviation

	Distractor		Target Condition			
	Bright	Dim	Dim/Bright	Dim/Dim	Bright/Dim	Bright/Bright
Mean of distribution	0.000	0.328	0.385	0.922	1.071	1.606
Standard deviation	1.000	0.799	1.050	0.851	0.873	1.138

Confidence Criterion	Value
100-New	—
75-New	−0.569
50-New	−0.019
25-New	0.351
0-New	0.596
0-Old	0.665
25-Old	0.697
50-Old	0.840
75-Old	1.133
100-Old	1.611

of the signal detection model in which the variances of the dim distractors and the four target conditions were free parameters (10 free parameters in all). A single set of confidence criteria and a single decision criteria were used for all conditions. The resulting fit is shown in Figure 7.

The fitted parameter values are shown in Tables 2A and 2B for all SDT fits. As anticipated above, the best fit of the dim distractor variance is one that is less than the bright distractors. The variance of the dim distractors is 0.80, while that of the bright distractors is fixed at 1.0. This change in variance (as well as a leftward shift of bright distribution) allows the model to qualitatively account for the distractor confidence values shown in Table 1. However, even this model cannot account for the very high mean confidence value found when subjects say “old” to bright distractors.

Note that the increase in variance occurs only slightly as a result of an increase in *d′* (or the location of the distribution); instead, bright test items produce higher variance estimates than dim test times regardless of the position of the distribution along the axis. This is entirely consistent with the metacognitive strategy that we have proposed to account for our discontinuous state-trace plot. Subjects assume they will do better with bright test items, and this increase in confidence results in an increase in the variance of the distributions.

To summarize the results of the signal detection modeling, we find that an equal-variance version of the SDT model with a single set of decision and confidence criteria cannot account for several aspects of the data. Most notably, the model could not account for the fact that mean confidence for bright distractors increased for both “old” and “new” responses over dim distractors, which is typically not the way the equal-variance model fails to predict recognition data. A version of the model that allows unequal variances could qualitatively account for this effect, although it has difficulty with the very high confidence values that result when subjects say “old” to bright distractors. One interpretation of the unequal variance model is that the increases in variance for bright test distributions is a result of a metacognitive strategy in which subjects assume that brighter test faces always produce better performance, regardless of the response.

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 generally assume (erroneously) that a brighter test face will 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 nonanalytic mnemonic processes may provide the basis for much of the confidence ratings and produce strong correlations between confidence and accuracy.

GENERAL DISCUSSION

The principal 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 (or, equivalently, are based on nonmonotonic functions of a common source 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 JOLs, FOKs, 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 metacognitive judgments are being made (Koriat, 1997). Cues such as ease of processing are thought to be intimately tied to the stimulus, and are therefore described as intrinsic cues. Cues relating to the study conditions are thought of as extrinsic cues. Both of these are analytic in nature, in that they involve heuristics that subjects overtly use to make their confidence ratings (i.e., "I had longer to study that item, therefore I must have a better memory for it"). There is also a nonanalytic, 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 suggests a number of conclusions about the mechanisms underlying metamemorial 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 sec 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, whereas 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. Thus these aspects of our data fail to coincide with the predictions of Koriat's accessibility hypothesis.

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 (1995) accessibility hypothesis, in which subjects move from the use of analytic heuristics applied to intrinsic and extrinsic cues at study to a nonanalytic 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 has described 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.[7]

Although we have proposed a two-state model to account for the dissociations of confidence and accuracy seen in Experiment 3, Clark (1997) 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 forced-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 distractor to an item in memory. This assumption is implemented by subtracting the distractor 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 distractors. Predictions on each trial can be captured by subtracting the distractor strength from the target strength. As the variability of this target-strength-minus-distractor-strength distribution increases (as a result of the intervening pictures), accuracy goes down (more distractor 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 (1997) 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 implications 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 in the making of 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.

Other work has suggested that multiple dimensions may be at work in recognition memory and confidence judgments. Dobbins, Kroll, and Liu (1998) and Yonelinas, Kroll, Dobbins, Lazzara, and Knight (1998) have suggested that confidence and accuracy can be dissociated on the basis of "remember–familiar" judgments. They have suggested that confidence may track the level of familiarity fairly directly, but that confidence and recollection may be more tied to processes under strategic control. These processes might include the metacognitive heuristics described by cue utilization theories, including the strategy apparently employed by our subjects in Experiment 3 who assumed that bright test items are always better than dim test items.

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, on the basis of Experiment 3, we would have to recommend that eyewitnesses who perceive a crime at night should view a lineup in the dark! Clearly this is a solution that only a defense attorney could love. In addition, we should point out that we used the same pictures at study and at test, which is rarely the case in the legal setting unless an eyewitness views a photo lineup twice.

This difficulty suggests a current research line. Aficionados of encoding specificity (e.g., 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 laboratories.

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 & 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.

Bothwell, R., Deffenbacher, K., & Brigham, J. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology*, **72**, 691-695.

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

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, **22**, 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., & Penrod, S., & Stuve, T. (1988). Juror decision making in eyewitness identification cases. *Law & Human Behavior*, **12**, 41-55.

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

Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence–accuracy inversions in scene recognition: A remember–know analysis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1306-1315.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, **98**, 506-528.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, **13**, 8-20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Theory and data. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 5-16.

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**, 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 Experimental Psychology: Learning, Memory, & Cognition*, **23**, 1394-1409.

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

King, J. F., Zechmeister, E. G., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology*, **95**, 329-343.

Koriat, A. (1993). How do we know what we know? The accessibility model of 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.

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, **9**, 215-218.

Lindsay, R. C. L., Wells, G., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology*, **66**, 79-89.

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

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., & Ruthruff, E. R. (1994). A theory of visual information acquisition and visual memory with special application to intensity-duration tradeoffs. *Journal of Experimental Psychology: Human Perception & Performance*, **20**, 33-50.

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

Pelli, D. G., & 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, D. J. Herrmann, J. D. Read, D. Bruce, D. G. Payne, & M. P. Toglia (Eds.), *Eyewitness memory: Theoretical and applied perspectives* (pp. 107-130). Mahwah, NJ: Erlbaum.

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

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

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

Stretch, V., & Wixted, J. T. (1998a). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1397-1410.

Stretch, V., & Wixted, J. T. (1998b). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1379-1396.

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

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

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

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

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. L. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology*, **66**, 688-696.

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

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 681-690.

Yonelinas, A. P., Kroll, N. E., Dobbins, I., Lazzara, M., & Knight, R. T. (1998) Recollection and familiarity deficits in amnesia: Convergence of remember–know, process dissociation, and receiver operating characteristic data. *Neuropsychology*, **12**, 323-339.

NOTES

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 $S+T$.

2. Of course, as with the acceptance of any null hypothesis, there may have been a dissociation between retrospective confidence and accuracy that we did not observe. However, this null finding is replicated in Experiment 2, and we dissociate retrospective confidence and accuracy in Experiment 3.

3. 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 5–10 cd/m^2 range might be only four to five 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 that provides 4,096 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.

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

5. We thank John Wixted for making this point and motivating the signal detection analysis.

6. SDT is considered a multidimensional model by our definition because accuracy is a function of the distance between the target and distractor distributions (i.e., d'), while confidence is a function of the distance to the decision criterion. With a single distractor distribution, these two values become monotonic because the distance to the decision criterion is the same for all conditions, and therefore this version of SDT is single dimensional by our definition.

7. Alternatively, subjects may lack the ability to anticipate the decay in effectiveness of rehearsal over time. However, this does not imply that this is a single-dimensional model, since prospective confidence and accuracy are based on different sources of information (subjects assume that rehearsal will help a lot when they are making prospective confidence ratings, while in fact it helps relatively little in terms of accuracy).

(*Continued on next page*)

APPENDIX
What Constitutes a Single-Dimensional Model?

The definitions of single- and multidimensional models as defined by state-trace analysis, though exact, may not correspond to what have traditionally been viewed as single-dimensional models in the literature. To illustrate the nature of multidimensional models, below we provide examples taken from the present work to demonstrate how each situation requires a single- or multidimensional model.

Model 1: Prospective confidence is based on the value of a strength variable (S_1) at time T_1, and accuracy is based on strength (S_2) at time T_2. Let $S_1 \neq S_2$.

Interpretation: This model is a single-dimensional model even though $S_1 \neq S_2$, and a continuous state-trace plot will be produced.

Model 2: Prospective confidence at time T_1 is based on S_1, which is some function of strength due to luminance and rehearsal. Accuracy at time T_2 is based on S_2, which is also a function of strength due to luminance and rehearsal. However, strength due to rehearsal fades quickly, while strength due to luminance fades slowly. Thus when accuracy is assessed at time T_2 the strength is not as high as it should be to produce a continuous state-trace plot.

Interpretation: In most interpretations this case requires a multidimensional model, because it has two dimensions: Dimension 1, based on luminance, is S_L, and Dimension 2, based on rehearsal, is S_R. They combine together to produce S. This model will produce a discontinuous state-trace plot. There are, however, situations in which the underlying model is multidimensional and yet the state-trace plot is continuous. This is the same situation that is faced in traditional hypothesis testing in which the null hypothesis is not rejected yet a true difference between the conditions exists.

Model 3: Prospective confidence is based on strength (S) at time T_1 and on metacognitive certainty (C) at time T_1. Accuracy is based only on strength (S) at time T_2.

Interpretation: This is a multidimensional model in which the two dimensions are strength (S) and certainty (C). Metacognitive certainty could also be multidimensional, which would provide a further rejection of the single-dimensional model.

Model 4: Retrospective confidence is based on the placement of the confidence criterion in a signal detection model, while accuracy is based on the location of the distractor and target distributions along the axis.

Interpretation: If the experiment involves two identifiably different distractor distributions (e.g., bright and dim faces, low- and high-frequency words), then this situation could produce a discontinuous state-trace plot and therefore imply a multidimensional representation. This results from the fact that with two distractor distributions, d' (the distance between each target and relevant distractor distributions) is independent of the placement of the confidence criterion. However, if only a single distractor distribution is present, then a single-dimensional model is more likely. There are, however, conditions in which the processes that produce the value along the signal detection axis (usually termed *sense of prior occurrence* or *match to memory*) are themselves multidimensional. This would produce a discontinuous state-trace plot even with a single set of distractor stimuli. Such a condition could arise if both familiarity and recollective processes determined recognition performance, and confidence is a function of only one process (e.g., recollection).

(Manuscript received July 15, 1998;
revision accepted for publication June 16, 1999.)