Psychonomic Bulletin & Review
2009, 16 (5), 824-831
doi:10.3758/PBR.16.5.824

A dissociation between similarity effects in episodic face recognition

Andrew Heathcote, Emily Freeman, Joshua Etherington, Julie Tonkin, and Beatrice Bora
University of Newcastle, Callaghan, New South Wales, Australia

Memory similarity, the similarity between a test lure and memory traces, reduces confidence and accuracy in all forms of recognition memory. In contrast, Tulving (1981) showed that, in recognition memory for scenic pictures, *choice similarity*, the similarity between forced choice test alternatives, increased accuracy but decreased confidence. In the present study, we replicated both memory and choice similarity effects and the dissociation between accuracy and confidence with pictures of faces. State-trace analysis confirmed the dissociation and identified two dimensions underlying these effects, one associated with choice similarity and another associated with memory similarity. Further analysis showed that the effect of study–test lag was associated with the memory-similarity dimension.

Unstudied test items that are judged to be perceptually or conceptually similar to study items produce increased false recognition for a wide range of stimuli (e.g., words, numbers, shapes, and pictures). The resulting negative correlation between similarity and accuracy, which we call the *memory similarity effect*,[1] is a general regularity in recognition memory tested by either yes/no or two-alternative forced choice (2AFC) procedures (e.g., Wickelgren, 1977). A second regularity, a positive correlation between recognition accuracy and retrospective confidence (i.e., confidence judgments given at test) is attributed to memory trace strength's being the common basis for both decisions (Hart, 1967); stronger traces cause greater confidence and are associated with more accurate recognition decisions.

Hence, it was doubly surprising when Tulving (1981) reported a reversal of both regularities in 2AFC recognition decisions for scenic pictures followed by a confidence rating. The reversals were caused by *choice similarity*, the similarity between studied (target) and unstudied (lure) test alternatives. In higher choice similarity pairs (AA′), both items were halves of the same scenic picture, in which one half (A) was studied and the other (A′) was not. In lower choice similarity pairs (BC′), the lure (C′) was the unstudied half of a studied picture (C) that was not similar to the target (B). Since the lure in both choice similarity conditions was similar to a studied item, memory similarity was controlled. Tulving found that, for picture halves rated as higher in similarity, accuracy was increased but confidence was reduced for AA′ relative to BC′ test pairs. These results are all the more compelling because, in the same experiments, the usual memory similarity effect was obtained when choice similarity was controlled. Accuracy and confidence were greater for lower than for higher memory similarity pairs. However, for picture halves rated lower in similarity, only the choice similarity effect on confidence was reliable.

Despite the surprising nature of Tulving's (1981) results, to our knowledge only one replication has been performed: Dobbins, Kroll, and Liu (1998) used the same type of stimuli that Tulving used, but their choice similarity effect was weaker for both accuracy and confidence, likely because memory similarity was lower for their stimuli. We know of no attempt to determine whether the choice similarity effect and the dissociation between confidence and accuracy apply to other types of stimuli. In the three experiments reported here, we investigated whether these phenomena also occur in recognition memory for faces. The experiments differed only in their test procedures. Experiment 1 followed Tulving, with a 2AFC response followed by a confidence rating on a 3-point scale. In Experiment 2, participants simultaneously indicated their choice and confidence on a 6-point scale. Experiment 3 followed Dobbins et al. in requiring a "remember/know" response after a simultaneous choice and confidence rating.

Two models of the Tulving (1981) effect have been proposed. Although both were based on data collected with scenic images, the mechanisms that they propose might also apply to face stimuli. Clark's (1997) *single-process* model assumes that 2AFC is based on a single *evidence* variable equal to the difference in memory strength (i.e., the match between memory and test cues) between test items. Memory strengths for higher choice similarity pairs are positively correlated, not only facilitating average accuracy but also making more extreme differences less likely, which reduces average confidence. Dobbins et al.

A. Heathcote, andrew.heathcote@newcastle.edu.au

(1998) proposed a *dual-process* model to explain their finding that "remember" responses, and by implication decisions based on recollection, were less common but more accurate when choice similarity was higher. For "know" responses, accuracy was unaffected by choice similarity, whereas confidence was decreased when choice similarity was higher (see also Voss, Baym, & Paller, 2008, for an alternative dual-process account in terms of implicit and explicit memory processes).

Both models explain the confidence–accuracy dissociation in terms of differential effects on two underlying dimensions or latent variables that control either two simple processes or a single more complex process. In the dual-process model, each dimension is equivalent to a process (familiarity and recollection), each of which is controlled by a single latent variable (mean familiarity and recollection probability, respectively). In Clark's (1997) model, the single-process is controlled by two latent variables (evidence mean and variance). Rather than comparing the models directly,[2] we tested their shared assumption that two dimensions are required to explain the confidence–accuracy dissociation.

Following Busey, Tunnicliff, Loftus, and Loftus (2000), we used state-trace analysis (Bamber, 1979) to test the dimensionality of the relationship between confidence and accuracy. State-trace analysis uses a plot to test whether one psychological dimension is sufficient to explain the relationship between a pair of dependent variables without making assumptions that have been identified as problematic in other approaches to this question (Dunn & Kirsner, 1988). Results for one dependent variable (e.g., accuracy) are plotted against results for the other (e.g., confidence). Points on the plot represent results from experimental conditions. In our case, the conditions were the result of a factorial combination of high and low choice similarity and memory similarity manipulations. If all of the points on the state-trace plot could be joined by a single monotonic (i.e., always increasing or always decreasing) curve, all findings could be explained by variations on a single underlying psychological dimension. If a monotonic curve did not suffice, more than one dimension would be required.

Like us, Busey et al. (2000) studied episodic recognition memory for faces. They found that one dimension explained the relationship between confidence and accuracy (i.e., state-trace plots were monotonic) in two experiments that manipulated only study-related factors, either allowing or not allowing rehearsal combined with manipulations of stimulus duration (Experiment 1) or luminance (Experiment 2). In contrast, more than one dimension was required (i.e., the state-trace plot was nonmonotonic) in a third experiment that manipulated luminance and the match between study and test stimulus luminance. Confidence and accuracy dissociated because participants' confidence judgments were less affected than was their accuracy by the match between study and test stimuli: They were more confident for bright than for dim test stimuli, even when those bright test stimuli were less accurate because they were studied dim. Tulving's (1981) dissociation is analogous: Participants were more confident for low than for high choice similarity, even though accuracy was greater for high than for low choice similarity pairs when memory similarity was high.

In our Experiments 2 and 3, we also examined the effect of the time between study and test (i.e., *study–test lag*). Arguably, the single-process model makes a very constrained prediction: The effect of lag will dissociate with the effect of choice similarity, but *not* with the effect of memory similarity. That is, a monotonic curve will join points in a state-trace plot from different memory similarity and lag conditions, but a different curve will be required for high and low choice similarity conditions. If this were the case, it would demonstrate that a one-dimensional model is not a straw man in our paradigm, in the same way that Busey et al.'s (2000) findings about rehearsal and study duration showed that the one-dimensional model was not a straw man in their paradigm.

Our predictions about lag also test the generality of the mechanisms underlying the single-process model, because they are based on using those mechanisms to extrapolate from an empirical finding about lag effects in another paradigm. Ratcliff, McKoon, and Tindall (1994) found an effect of study position on the mean but not on the variance of memory evidence. In Clark's (1997) model, memory similarity affects mean evidence and choice similarity affects evidence variance. Hence, only one dimension (mean evidence) is required to predict the joint effects of lag and memory similarity.

EXPERIMENTS

Face stimuli were classified by gender and race (Black, Asian, and White). The experiments used the same 2 × 2 factorial design, crossing higher and lower choice similarity with higher and lower memory similarity, as in Tulving (1981, Experiment 2). Memory similarity was manipulated using pairs of faces from sets that were rated higher and lower on perceptual similarity (see Figure 1). Higher choice similarity resulted when test alternatives were members of the same pair. Lower choice similarity resulted when test alternatives were from different racial categories. Formation of lower choice similarity pairs from different races was a convenience based on the available stimulus set.[3]

Method

Participants. Introductory psychology students at the University of Newcastle, Australia, participated in exchange for course credit (38 participated in Experiment 1, 35 participated in Experiment 2, and 45 participated in Experiment 3).

Apparatus and Procedure. Face images (105 × 120 pixel grayscale bitmaps) from the FERET database (Phillips, Wechsler, Huang, & Rauss, 1998) were grouped into 377 generally similar pairs. Pair similarity was rated by 10 first-year psychology students using a 5-point scale (1 = *very low* to 5 = *very high*). Pairs were rank ordered using average similarity ratings within gender and race categories; higher and lower similarity sets were created by selecting lowest and highest ranked pairs. Table 1 lists the rating results for each category for 240 critical pairs. Four faces appeared before and after the critical faces in the study list as untested primacy and recency buffers. Buffer faces and faces used for an initial practice study–test cycle were drawn randomly from the remaining 137 face pairs.

Figure 1. Study list construction. For each study list, 24 critical study items were randomly selected members from 12 high memory similarity pairs and 12 low memory similarity pairs. (A) Example black and white female study list showing half of the actual number of critical items. Examples of higher memory similarity items are designated by uppercase letters (A, B, C), and examples of studied lower memory similarity items are designated by lowercase letters (a, b, c). (B) Pair mates for studied faces, which were not studied. Examples of unstudied higher memory similarity items are designated by uppercase letters with a prime (A′, B′, C′), and examples from lower memory similarity items are designated by lowercase letters (a′, b′, c′). Note that items with the same letter in panels A and B (e.g., A and A′) are pair mates.

Responses were recorded via a six-button array in Experiment 1. The first button on the left, labeled "GO," was used to initiate study–test cycles. The remaining five buttons were labeled "L, R, 1, 2, 3," from left to right. In all experiments, Buttons 1–3 were used to make typicality ratings of faces in the study phase (1 = *very typical* to 3 = *very unusual*). Typicality ratings were elicited to ensure attention to the faces and were not analyzed further. In the test phase, Buttons 1–3 were used to make confidence ratings (1 = *guess* to 3 = *sure*). In Experiment 1, the "L" and "R" buttons were used to indicate whether the target face was on the left or right. An eight-button array, consisting of left and right hand clusters of three keys and a central pair of keys, was used to record confidence and accuracy in Experiments 2 and 3. In Experiment 2, the left and right clusters were labeled "3, 2, 1, 1, 2, 3," from left to right. Participants made their target face choice by pressing a button in either the left or the right cluster. The button pressed within the cluster indicated confidence.

In Experiment 3, we attempted to replicate the response procedure used by Dobbins et al. (1998). The left- and right-hand clusters were labeled "1, 2, 3" and "3, 2, 1," from left to right, and both buttons in the central pair were labeled "4," because participants were required to rate their confidence using a 4-point scale. Additionally, the central pair of keys was used to make remember–know judgments. Participants pressed the left button, labeled "remember," if they remembered seeing the face or particular elements of the face; they pressed the right button, labeled "familiar," if the face was familiar but they did not remember the face or any particular elements of the face.

Testing used a PC with a 1,168 × 856 pixel resolution monitor. The experimental session, which lasted from 40 to 55 min, began with participants' reading instructions on the screen at their own pace. During the study phase, the faces were displayed for 2 sec each, one at a time, in the middle of the screen. After each face had appeared, participants were prompted to make a typicality rating. The test phase began immediately after study. In the test phase, face pairs appeared one pair at a time. If no response was made after 6 sec, the next pair was displayed.

In each of the 11 study–test cycles (the first being practice), participants studied 32 faces presented in a random order, except that the

Table 1
Characteristics of the 240 Critical Experimental Face Pairs

Gender	Race	Similarity	Mean Rating (%)	Number
Female	Black	Lower	38	18
		Higher	67	18
	Asian	Lower	31	6
		Higher	64	6
	White	Lower	33	24
		Higher	67	24
Male	Black	Lower	33	18
		Higher	63	18
	Asian	Lower	37	18
		Higher	64	18
	White	Lower	34	36
		Higher	71	36

Note—Female pairs were used to create one Asian–White and three Black–White study lists. Male pairs were used to create three Asian–White and three Black–White study lists.

1st and last 4 were buffer items. Each study list used faces that were all of the same gender, half from one race and half from another (see Figure 1). High choice similarity pairs were created by pairing a studied face with its unstudied pair mate. Low choice similarity pairs were created by pairing a studied face from one race with an unstudied face from another race (see Figure 2). The order of the 16 test pairs and the side on which the target was presented were randomized.

RESULTS

We excluded 8, 4, and 7 participants from analysis in Experiments 1–3, respectively, either because their accuracy was less than 60% (accuracy for other participants was above 70%) or because they did not follow instructions to use all confidence ratings (i.e., rarely or never using the lowest confidence rating). Integer confidence ratings (r = 1–3) were converted to percentages using the formula $100 \times (r - 1)/2$. Following Tulving (1981), we calculated confidence on the basis of all responses, both correct and incorrect.

ANOVAs examined accuracy and confidence as a function of experiment, choice similarity, and memory similarity factors. Effects with $p < .05$ were considered reliable. *Experiment* did not reliably interact with any other variables, so Figure 3 shows results averaged over experiments.[4] There was a reliable main effect of choice similarity [$F(1,96) = 17.1$, $p < .001$], with high more accurate than low by 2.3%. The main effect of memory similarity was also reliable [$F(1,96) = 51.9$, $p < .001$], with low more accurate than high by 2.3%. Consistent with the almost identical main effects of each factor, their interaction was not reliable ($F < 1$).

For confidence, there were reliable main effects of choice similarity [$F(1,96) = 26.0$, $p < .001$], with the low more confident than high by 3.5%, and of memory similarity [$F(1,96) = 47.9$, $p < .001$], with low more confident than high by 2%. The interaction between choice and memory similarity was reliable [$F(1,96) = 6.7$, $p = .011$], because the difference in confidence due to memory similarity was larger for high (2.8%) than for low (1.2%) choice similarity pairs. Note, however, that the effect of choice similarity was reliable in both low [$t(98) = 3.23$, $p < .002$] and high [$t(98) = 4.1$, $p < .001$] memory similarity conditions.

Figure 4 displays state-trace results. For Experiment 1 (panel A), a two-factor explanation of the confidence–accuracy dissociation is clearly supported. Within each level of choice similarity (i.e., points joined by lines), both confidence and accuracy increase as memory similarity decreases. In contrast, within each level of memory similarity, accuracy increases, but confidence decreases as choice similarity increases. Panels B and C of Figure 4 show state-trace results for Experiments 2 and 3, broken down by short versus long lag. The lag factor was created using a median split; the average short and long lags were 16.5 and 29.5 study and test events, respectively. For low choice similarity test pairs, lag was calculated using the study position of the unstudied test item's studied pair mate. Conditions with the same choice similarity and lag are joined by lines. All four lag and memory similarity conditions within each level of choice similarity can be joined by a single monotonic curve (within experimental error). These results indicate that the effects of memory similarity and study–test lag can be explained by a single dimension. In contrast, the curve for higher choice similarity conditions is displaced upward and to the left (i.e., more accurate and less confident) of the function for lower choice similarity, indicating that a second dimension is required to explain the effect of choice similarity.

DISCUSSION

For faces, we observed the same confidence–accuracy inversion that Tulving (1981) and Dobbins et al. (1998) found with scenic pictures. Increased choice similarity improved accuracy but decreased confidence when the effect of memory similarity was controlled. Hence, the choice similarity effect appears to occur along a general visual similarity dimension that applies to both scenes and faces. The potentially greater range of variation between halves of scenic pictures than between pairs of faces may explain why our results differed from Tulving's in two respects: (1) Our choice and memory similarity effects were smaller, although highly reliable, and (2) Tulving obtained a choice similarity effect only for his higher memory similarity pairs, whereas we found reliable effects for our lower and higher memory similarity pairs. Both differences were likely caused by face pairs sharing a great deal of structural similarity, whereas scenic picture halves can be quite dissimilar. Hence, the difference between stimuli in lower and higher choice and memory similarity pairs is likely to be much greater for scenes than for faces, causing larger effects for scenes. In the same vein, our lower memory similarity pairs were likely still sufficiently similar to support a reliable choice similarity effect, whereas Tulving's lower memory similarity pairs were likely not sufficiently similar.

Our findings with 2AFC tests of memory for facial stimuli may have implications in the applied domain of eyewitness identification. Juries tend to assume a positive correlation between confidence and accuracy (Penrod &

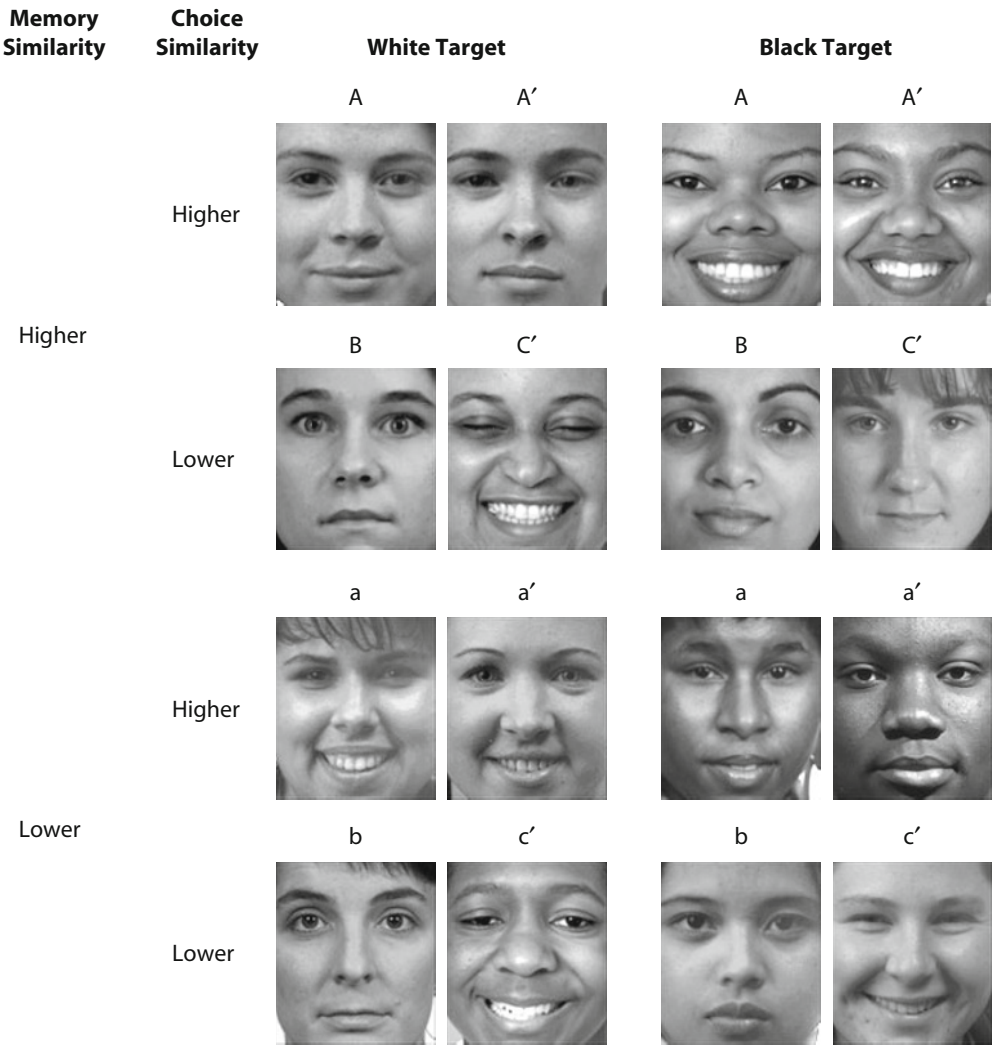

Figure 2. Test list construction. Each test list presented participants with two face pairs from each of the eight race × choice similarity × memory similarity conditions. The figure shows a half-length example test list constructed from the studied and unstudied faces shown in Figure 1. For example, the low choice similarity and low memory similarity test pairs (bc′) are constructed from studied (b) and unstudied (c′) items in Figure 1. Note that members of higher choice similarity test pairs (AA′) look similar to each. For higher memory similarity test pairs (BC′), the unstudied pair member (C′) looks similar to their studied but not tested pair mate (C) shown in Figure 1.

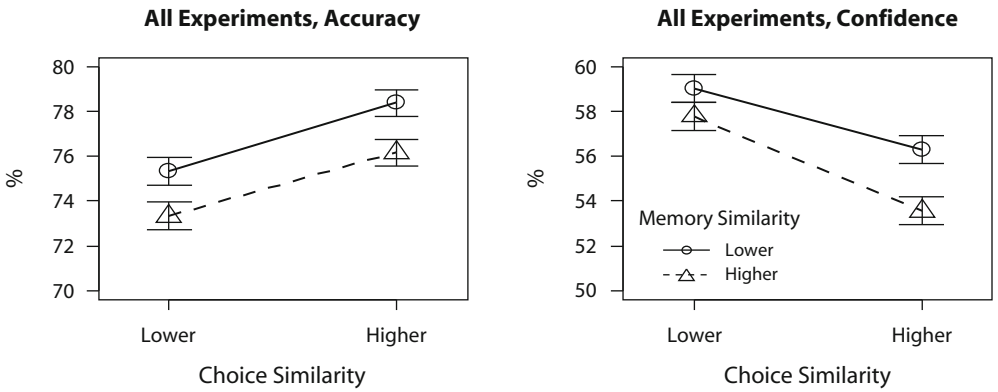

Figure 3. Confidence and accuracy as a function of choice and memory similarity from Experiments 1–3, averaged over experiments. Standard error bars were calculated using Loftus and Masson's (1994) method for a within-subjects design.

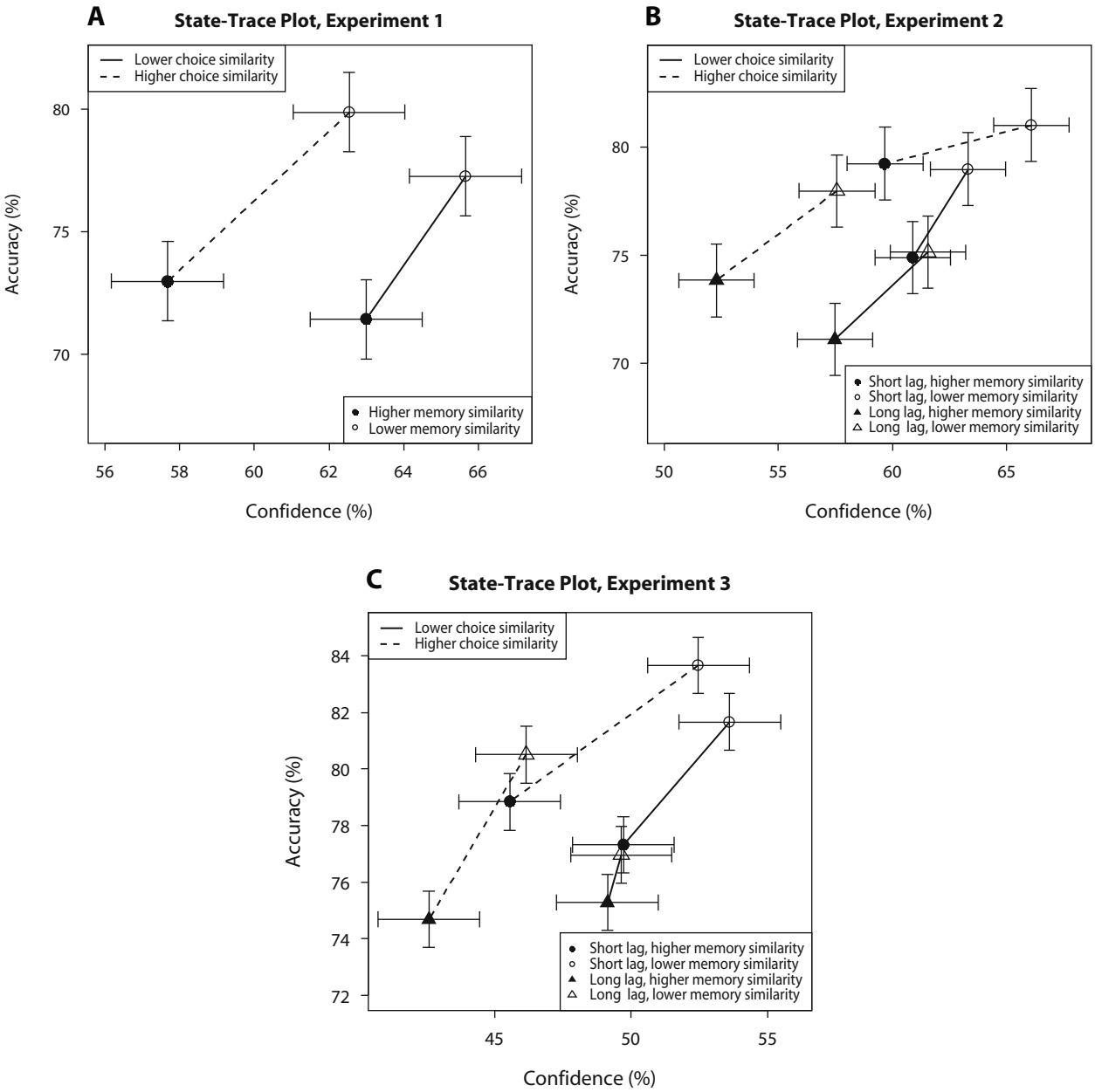

Figure 4. State-trace plots of accuracy as a function of confidence as a function of choice and memory similarity conditions. For Experiments 2 and 3, results were further divided on the lag between study and test. Standard error bars were calculated using Loftus and Masson's (1994) method for a within-subjects design.

Cutler, 1995), whereas our results imply that they can be negatively correlated. Even though we found a smaller confidence–accuracy dissociation than did Tulving (1981) with natural scenes, the dissociation is likely to be quite pervasive, because of the high level of similarity among all faces. Clearly, however, more investigation is needed to explore these implications in more ecologically valid paradigms. For example, although 2AFC tests resemble criminal identification lineups, there are many differences, such as more than two choices and the ability to make no choice in lineups (see Clark, 2003).

We obtained the same choice and memory similarity effects with the response methods used in previous investigations—choice followed by a confidence rating (Tulving, 1981) and a simultaneous choice and confidence response followed by a remember–know classification (Dobbins et al., 1998)—as well as with a simultaneous choice and confidence rating alone (our Experiment 2). One of the reasons we ran the latter condition was to check a potential speed–accuracy trade-off (Reed, 1973) explanation of the choice similarity effect suggested by Tulving's statement that "highly similar test items may induce par-

ticipants to . . . examine the evidence more thoroughly" (p. 495). Test response times were strongly affected by confidence (higher confidence decisions were quicker) and to a lesser degree by accuracy (correct choices were quicker). However, when the effects of differences in confidence and accuracy between memory and choice similarity conditions were controlled, neither choice nor memory similarity had a reliable effect.[5] Hence, speed–accuracy trade-off was unlikely to be the cause of the choice similarity effect.

For all response methods, state-trace analysis consistently indicated that at least two psychological dimensions are required to explain the dissociation between choice and memory similarity effects on confidence and accuracy. Study–test lag effects in Experiments 2 and 3 could be explained by the same dimension as that used for memory similarity effects, but both dissociated from choice similarity.[6] Qualitatively, the dissociation that we found and the one found by Busey et al. (2000) are similar in that they both relate to differences between conditions that might be evident to participants at test (i.e., brightness and similarity between test alternatives). A potential explanation for both dissociations suggested by Busey et al. is that confidence judgments are affected by erroneous beliefs about the effects of test differences on accuracy. However, differences evident at test may not always be necessary to cause a dissociation. Voss et al. (2008) reported that dividing attention during study of abstract visual stimuli resulted in decreased confidence but increased accuracy in a 2AFC test using high choice similarity pairs.

The state-trace results also have strong implications for quantitative memory models. In Clark's (1997) single-process model, the state-trace results are consistent with changes in mean memory evidence's underpinning lag and memory similarity effects and with changes in evidence variance's underpinning choice similarity effects. In Dobbins et al.'s (1998) dual-process model, these results are that consistent lag and memory similarity have the same pattern of effect on familiarity and recollection and that choice similarity has a different pattern of effect. Given Yonelinas and Levy's (2002) suggestion that study–test lag affects familiarity but not recollection, these results could be used as a basis for extending Dobbins et al.'s model to address lag and memory similarity effects using a familiarity-based mechanism. In general, our results demonstrate the power of state-trace analysis to provide guidance for the development of different process models without requiring a commitment to the detailed assumptions made by any one model.

AUTHOR NOTE

We thank Mark Steyvers for providing the set of faces, which were sorted into generally similar pairs by members of his laboratory. We thank Ken Norman, Jason M. Watson, and an anonymous reviewer for helpful comments. Address correspondence to A. Heathcote, School of Psychology, University of Newcastle, Psychology Building, University Ave., Callaghan 2308, Australia (e-mail: andrew.heathcote@newcastle.edu.au).

REFERENCES

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

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

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.

CLARK, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology*, **17**, 629-654.

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.

DUNN, J. C., & KIRSNER, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, **95**, 91-101.

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

HEATHCOTE, A., BORA, B., & FREEMAN. E. (2009). *Recollection and confidence in two-alternative forced choice episodic recognition*. Manuscript submitted for publication.

LOFTUS, G. R., & MASSON, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, **1**, 476-490.

MEISSNER, C. A., & BRIGHAM, J. C. (2001). Thirty years of investigating the own-race bias in memory for races: A meta-analytic review. *Psychology, Public Policy, & Law*, **7**, 3-35.

PENROD, S., & CUTLER, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, & Law*, **1**, 817-845.

PHILLIPS, P. J., WECHSLER, H., HUANG, J., & RAUSS, P. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image & Vision Computing Journal*, **16**, 295-306.

RATCLIFF, R., MCKOON, G., & TINDALL, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 763-785.

REED, A. V. (1973). Speed–accuracy trade-off in recognition memory. *Science*, **181**, 574-576.

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

VOSS, J. L., BAYM, C. L., & PALLER, K. A. (2008). Accurate forced-choice recognition without awareness of memory retrieval. *Learning & Memory*, **15**, 454-459.

WICKELGREN, W. A. (1977). *Learning and memory*. Englewood Cliffs, NJ: Prentice Hall.

YONELINAS, A. P., & LEVY, B. J. (2002). Dissociating familiarity from recollection in human recognition memory: Different rates of forgetting over short retention intervals. *Psychonomic Bulletin & Review*, **9**, 575-582.

NOTES

1. We avoided the more commonly used term *target–lure similarity*, because that could also apply to the similarity between choices in a two-alternative forced-choice test (i.e., what we call *choice similarity*). The term *memory similarity* emphasizes the relationship between a memory trace and a test lure, which is what our memory similarity manipulation affects while controlling for choice similarity.

2. We collected remember–know judgments in Experiment 3 to pilot an experiment used to compare the models. For the present purposes, Experiment 3 serves as a check on whether making remember–know responses changes choice and accuracy effects (it did not). Lack of space does not allow us to report remember–know results for Experiment 3 here. A model comparison based on remember–know responses is reported elsewhere (Heathcote, Bora, & Freeman, 2009).

3. The use of race does not confound the effects on which we focus, due to counterbalancing. Our study format, which mixes examples of different races within a study list, is also likely to weaken race effects, which are stronger when race is blocked (Meissner & Brigham, 2001). Consistent with race's not affecting our results, the experiment reported by Heathcote et al. (2009) showed the same pattern of results, using single-race lists with low choice similarity pairs created by pairing faces with different genders.

4. A reviewer noted that, when lag was included as a factor, there were slight deviations from the confidence–accuracy inversion at short lags in Experiment 2 and at long lags in Experiment 3. Because the interaction with lag was not consistent across experiments and because an ANOVA on the Experiment 2 and 3 data including a lag factor did not produce any reliable interactions, we attribute these deviations to measurement error.

5. The same was true with the other two response methods for the time to make the first response and for the sum of the times to make both responses. Where two responses were required, our participants appeared to make both decisions before making the first response, because the time for the second response was fast and was unaffected by choice and memory similarity.

6. Lag effects for Experiment 1 could not be analyzed, because we did not save information about each test item's study position. Given the similarity of other effects between Experiments 1–3, it seems unlikely that lag effects would differ much between Experiment 1 and the other experiments.

(Manuscript received August 29, 2008;
revision accepted for publication May 13, 2009.)