European Journal of Social Psychology Eur. J. Soc. Psycho/. ***28,*** *437–456 (1998)*

Raw conditional probabilities are a flawed index of associative strength: evidence from a multi-trait paradigm

JOHN J. SKOWRONSKI[1 *], ANDREW L. BETZ[2], CONSTANTINE SEDIKIDES[3] and MATTHEW T. CRAWFORD[4]

1 *The Ohio State University at Newark, U.S.A.*
[2]GTE Laboratories, U.S.A.
[3]University of North Carolina at Chapel Hill, U.S.A.
[4]Indiana University at Bloomington, U.S.A.

Abstract

Skowronski and Welbourne (1997) argue that raw conditional probabilities may be a flawed index of associative strength in recall, and may need to be corrected for chance before they can be safely interpreted. Three experiments examined this idea in the context of an experimental paradigm used by Hamilton, Driscoll and Worth (1989). Participants in this paradigm were asked to read items describing a social target. The items each pertained to one of several different trait concepts, or were irrelevant to those concepts. Participants later recalled the items. The data supported Skowronski and Welbourne's conjecture. The raw conditional probabilities differed substantially from the chance-adjusted probabilities. The data from a second dependent measure, inter-item generation times, matched the pattern of adjusted conditional probabilities. In addition to their methodological implications, these results contradict the Complete Association Model of person representation proposed by Hamilton et al. *Finally, these data raise the possibility that traditional associative models of person memory, which were based on raw conditional probabilities (e.g. Srull & Wyer, 1989), are flawed.*

* Address for correspondence: John J. Skowronski, The Ohio State University at Newark, 1179 University Drive, Newark, OH 43055-1797, U.S.A. Tel: (614)366-9348. e-mail: skowronski.l@osu.edu.
Contract grant sponsor: NIMH; Contract grant number: ROI-MHSO730-OlAl.

CCC 0046–2772/98/030437–20$17.50
© *1998* John Wiley & Sons, Ltd.

Received 29 February 1996
Accepted 6 June 1997

INTRODUCTION

Person memory research has attempted to understand how trait expectancies that one has about others affect memory for information relevant to those expectancies (e.g. Hastie, 1980; Hastie & Kumar, 1979; Srull, 1981, 1983; see Rojahn & Pettigrew (1992) and Stangor & McMillan (1992) for meta-analytic reviews). Some of this research has relied on the conditional probability measure to index the strength of the association between separate items of information describing the target. Thus, a high probability of recalling an expectancy-inconsistent item after recalling an expectancy-consistent item (e.g. Srull, Lichtenstein, & Rothbart, 1985) was thought to reflect a strong association between the two items.

However, Skowronski and Welbourne (1997) argue that raw conditional probabilities are a flawed index of associative strength. They mathematically demonstrate that raw conditional probabilities can be affected by numerous factors, such as the number of items from different categories present in the original item pool. For example, if all presented items are recalled in a random order, the raw conditional probability of having a category B item follow a category A item will be quite different if the item pool contains seven items from category A and seven items from category B than if the item pool contains two items from category A and 12 items from category **B.** Obviously, differences between conditional probabilities in these two cases do not reflect differences in associative strength between members of category A and category B. Instead, such differences merely reflect conditional probabilities expected by chance.

Skowronski and Welbourne argue that failure to control for chance-based factors that influence raw conditional probabilities may lead researchers astray. For example, consider an experiment by Hamilton, Driscoll and Worth (1989; also see Driscoll, 1992). Hamilton *et al.* presented participants with a multiple trait expectancy about a target (i.e. friendly, intelligent, and adventurous; or unfriendly, unintelligent, and unadventurous). They then presented to participants a list of 30 behaviours describing the target. Either five items (all in one category) or 10 items (five from each of two categories) in this list were inconsistent with the trait expectancy, either 5 or 10 items were irrelevant to the expectancy, and 15 items (five from each of the three categories) were consistent with the expectancy. Examination of the mean conditional probabilities obtained from participants' recall protocols indicated that conditional probabilities were higher when the two items in a sequence were from the same trait dimension (e.g. adventurous–adventurous or adventurous–unadventurous) than when they were from different trait dimensions (e.g. adventurous–friendly or adventurous–unfriendly). Furthermore, Hamilton *et al.* noted that the conditional probabilities for the within-trait dimension recall sequences were not too different from each other, and the conditional probabilities for all between-trait dimension recall sequences were about equal to each other.

Hamilton *et al.* (1989) used these data (as well as recall data and inter-item generation times) to argue that participants used the trait expectancies as organizing themes in constructing a person representation of the target. Furthermore, their *Complete Association Model* (CAM) posits that participants thought about items describing the target relative to all other items that came from the same trait dimension. However, the CAM takes the conditional probabilities at face value, and assumes that these probabilities are unbiased indices of associative strength. This

assumption may be incorrect. To illustrate the problem, consider the following example.

Assume that a participant was told to expect that an actor was friendly, then read a list of five expectancy-consistent items, five expectancy-irrelevant items, and five expectancy-inconsistent items describing the target. Next, the participant attempts to recall the items. For simplicity's sake, assume that all items that were presented were recalled. If the order of recall is random, the expected conditional probability of an expectancy-consistent item following another expectancy-consistent item is different (4/14, or *0.286)* from the expected conditional probability of an expectancy-inconsistent item following an expectancy-consistent item *(5/14,* or *0.357).* The reason for this expected difference is simple: given recall of an expectancy-consistent item, there are only four chances that the next item recalled will be an expectancy-consistent item, but five chances that the next item recalled will be expectancy-inconsistent.

Skowronski and Welbourne (1997) argue that conditional probabilities should be interpreted only after they are adjusted for these expected conditional probabilities. One such adjustment is to subtract the expected conditional probability from the actual conditional probability. This adjustment is termed the CPDIFF index.

Consider the effects that calculating the CPDIFF index might have on the Hamilton *et al.* (1989) data. The expected conditional probabilities in the Hamilton *et al.* study are lower for sequences in which the items come from the same trait category (e.g. friendly–friendly) than for sequences in which items come from denotatively opposite categories (e.g. friendly–unfriendly). Subtracting the expected conditional probabilities from the raw conditional probabilities should yield higher difference scores for sequences in which items come from the same trait category than for sequences in which items come from different trait categories. Thus, these adjusted probabilities would suggest that, given recall of an item, it was easier to next recall an item that comes from the same trait category as the original item than an item from the opposite trait category. This outcome is inconsistent with the CAM's assumption that inter-item associations in a person representation are just as strong for behaviours with implications for denotatively opposite trait categories as for behaviours with implications for the same trait category.

Overview

We conducted three experiments. These experiments explored the idea that the raw conditional probabilities obtained in an experiment need to be adjusted for chance before they can be interpreted correctly. The experiments also explored the implications of these adjusted conditional probability data, as well as accompanying recall and inter-item generation time data, for Hamilton *et al.'s* (1989) CAM.

These experiments used variants of the multi-trait expectancy method that was employed by Hamilton *et al.* (1989, Experiment 1). For example, in our initial experiment participants engaged in an impression formation experiment. They first read a paragraph indicating that a target (Bob) was either adventurous, intelligent, and friendly (positive expectancy) or unadventurous, unintelligent, and unfriendly (negative expectancy). Participants then read a series of behaviours that Bob performed. The behaviours were consistent, inconsistent, or irrelevant to the traits in the expectancies. The exact numbers of behaviours that were consistent with the

expectancy and that were inconsistent with the expectancy were varied across groups. Participants in the one-exposure condition read the behaviours only once; participants in the two-exposure condition read the behaviours twice. After completing a filler task, participants attempted to recall the behaviours by reciting each behaviour into a tape recorder.

Four dependent measures were calculated from the recall lists provided by each participant. The first of these was the consistency of the recalled items with the expectancy (consistent, inconsistent, or irrelevant). The second dependent measure was the conditional probability of a recall sequence (described in more detail in the Methods section below). The third dependent measure was the CPDIFF score, which is simply the difference between the obtained conditional probability and the expected conditional probability. The final dependent measure (inter-item generation time) was the amount of time that passes between completion of the recital of an item and the beginning of the recital of the next item.

As noted in the Methods section below, the three experiments that we conducted each used different manipulations of participants' processing goals in an attempt to explore whether such manipulations affected the content and organization of the recall list (see Seta & Hayes, 1994). These manipulations had minimal impact on the data. Furthermore, although the data for the adjusted conditional probabilities were clear and consistent across each of the three experiments, results for the recall and time measures were equivocal. The minimal impact of the processing goals manipulations and the need to enhance power for two of the dependent measures induced us to combine the data from the three experiments into a single analysis. Each of the three experiments, and the analytic technique used to combine the results of the three experiments, are described in the next section.

METHODS

Experiment 1

Participants

Participants were 96 undergraduates from the University of Wisconsin Madison, and were tested one at a time. Participation in the experiment was one of the options that could be chosen to fulfil research requirements in an introductory psychology course.

Stimulus Behaviours

Five behaviours exemplified each of three positive traits (adventurous, intelligent, and friendly) and five behaviours exemplified each of three negative traits (unadventurous, unintelligent, and unfriendly). Other behaviours were irrelevant to these traits.

Stimulus Booklets

Participants each received a 30-page stimulus booklet. Each page described one of Bob's behaviours. There were 12 different versions of the booklet. Six booklets

contained predominantly negative behaviours. In three of these booklets, there were 15 negative, 10 neutral, and five positive behaviours. In the other three of these booklets, there were 15 negative, five neutral, and 10 positive behaviours. Six additional booklets contained predominantly positive behaviours. In three of these booklets, there were 15 positive, 10 neutral, and five negative behaviours. The other three booklets contained 15 positive, five neutral, and 10 negative behaviours. Table 1 presents the different combinations of behaviour sets used in the 12 booklets. The items were presented in a random order in all 12 booklets, with the constraint that two items from the same trait dimension never occurred in succession.

Procedure

On entering the laboratory, participants were told that they would receive information about a person named Bob, and that they were to form an impression of him *(impression set* instructions). The participants were then provided with either a positive or a negative person expectancy. In the *positive expectancy* condition, participants were told:

> In the packet of materials in front of you (turned face down) you will find a number of behaviours performed by an individual named Bob. Bob tends to be much more friendly and sociable than the average person. He enjoys making new friends, meeting with old ones, and generally tends to value social activities. Bob is also very intelligent. His sharp, quick mind has always helped him to excel at almost anything he does. Bob tends to be very adventuresome, as well. He enjoys new and exciting activities and actively seeks out adventures and experiences with uncertain outcomes. When I say begin, please turn the packet over and read each behaviour carefully as you form your impression of Bob.

In the *negative expectancy* condition, participants were told:

> In the packet of materials in front of you (turned face down) you will find a number of behaviours performed by an individual named Bob. Bob tends to be a bit less friendly and sociable than the average person. He enjoys spending time alone, and generally doesn't value social activities. Bob is also of below average intelligence. His mind works slow and has always prevented him from excelling at anything he does. Bob tends to be very timid, as well. He is scared of doing new things and prefers sticking to a daily routine and activities he knows he can do. When I say begin, please turn the packet over and read each behaviour carefully as you form your impression of Bob.

Paced by an audiotape providing .6 seconds of reading time for each, participants then read the behaviours. After all behaviours had been read, participants in the *one-exposure condition* completed a 3-minute filler task. Participants in the *two-exposure condition* were told:

> In order for you to double check or revise your impression of Bob, we would like you to read the information one more time.

Participants in this condition were again paced through the booklet by the audio-tape. After completing this second reading, participants completed the 3-minute filler

task. After the filler task, all participants engaged in a surprise recall task. They were told:

> Earlier in this experiment, you were asked to read a series of sentences performed by a person named Bob. We would like you to recall as many of these behaviours as possible. We realize that it would be impossible for you to remember the behaviours word for word, but please try to recall as many as you can as precisely as possible. We will be tape recording your recall of Bob's behaviours, so please speak loudly into the tape recorder.

Participants were given 4 minutes to recall the behaviours. After completion of the recall task, participants were debriefed and dismissed.

Expectancies and Stimulus Behaviour Combinations

The booklets were coded to yield a four-level *stimulus behaviour combination* factor that describes the number of behaviours that were consistent, irrelevant, and inconsistent with the initial expectancy. An example of this re-coding is as follows. When the expectancy was positive, the 15 positive/10 irrelevant/five negative booklets depicted in Table 1 would contain 15 expectancy-consistent, 10 expectancy-irrelevant, and five expectancy-inconsistent behaviours. By comparison, when the expectancy was negative the same booklet would contain five expectancy-consistent, 10 expectancy-irrelevant, and 15 expectancy-inconsistent behaviours.

Dependent measures

Four dependent measures were derived from each participant's tape-recorded recall list. The order in which behaviours were recalled was used to calculate our first dependent measure, *conditional probability.* Up to eight conditional probabilities were calculated for each participant, and reflect both the first and second item in a sequentially recalled pair. The eight recall sequences can be classified on two dimensions. The first is *recall sequence* (consistent–consistent, inconsistent–consistent, consistent–inconsistent, and inconsistent–inconsistent). The second is whether both items in the recall sequence come from the same trait dimension *(within-trait dimension),* or from different trait dimensions *(between-trait dimension).* Recall sequences involving irrelevant items were ignored.

The second of our dependent measures was the *adjusted conditional probability.* We used the CPDIFF measure described by Skowronski and Welbourne (1997) to adjust the raw probabilities. This adjustment simply involves using the recalled items to calculate the expected conditional probability for a sequence, then subtracting the expected probability for that sequence from the obtained probability. Values near 0 on this variable suggest that the obtained probabilities were close to those expected by chance; values above 0 suggest that obtained probabilities were above those expected by chance, and values below 0 suggest that the conditional probabilities were below that expected by chance.

The third of our dependent variables, *inter-item generation time,* was determined by timing with a stopwatch the interval between the last word of a recalled behaviour and

Table 1. Composition of the stimulus booklets used in Experiments 1—3 (entries reflect number of behaviours of each type)

	Adventurous	Intelligent	Friendly	Unadventurous	Unintelligent	Unfriendly	Irrelevant
15 positive/10 irrelevant/5 negative Booklet 1	5	5	5	5			10
Booklet 2	5	5	5		5		10
Booklet 3	5	5	5			5	10
15 positive/5 irrelevant/10 negative Booklet 4	5	5	5	5	5		5
Booklet 5	5	5	5		5	5	5
Booklet 6	5	5	5	5		5	5
10 positive/5 irrelevant/15 negative							
Booklet 7	5	5		5	5	5	5
Booklet 8	5		5	5	5	5	5
Booklet 9		5	5	5	5	5	5
5 positive/10 irrelevant/15 inconsistent Booklet 10	5			5	5	5	10
Booklet 11		5		5	5	5	10
Booklet 12			5	5	5	5	10

the first word of the subsequently recalled behaviour on the audiotapes. In our analyses we used only the times for the eight recall sequences described previously.

The fourth dependent measure was *recall.* A gist criterion was used to determine items that were correctly recalled. Each correctly recalled behaviour was coded in terms of its consistency with the participant's expectancy, and constitutes the *recalled behaviour consistency* factor in analyses of these data. For each participant, the proportion of events recalled for the *expectancy-consistent, expectancy-inconsistent* and *expectancy-irrelevant* categories was determined by dividing the number of recalled events in each category by the total number of events presented in each category (5, 10 or 15, depending on the stimulus behaviour combination).

Experiment 2

Participants

Participants were 96 undergraduates from The Ohio State University at Newark, and were tested one at a time. Participation in the experiment was one of the options that could be chosen to fulfil research requirements in an introductory psychology course.

Procedure

Most elements of the procedure used in Experiment 2 almost exactly duplicated Experiment 1. An exception involved a between-participants processing goal manipulation. At the start of Experiment 2, participants in *impression set* conditions were told:

> In this study, you will be presented with information about a person named Bob. As you read through the information, we would like you to think about Bob's personality traits.

Later in the experiment, these participants were provided with the same expectancy instructions used in Experiment 1. A second group of participants received *evaluative set* instructions:

> As a part of this study, you will be presented with information about a person named Bob. As you read through the information, try to decide if Bob is a person you would like or dislike.

The expectancy instructions given to participants in this evaluative set condition also were altered. Participants in the *positive evaluative expectancy* condition were told:

> In the packet of information in front of you, you will find a number of behaviours performed by an individual named Bob. Bob is a person that other people seem to like. Whenever anyone has anything to say about Bob, it usually seems to be good. When I say begin, please turn the packet over and read each behaviour carefully as you try to decide for yourself whether Bob is a person you would like or would dislike.

Participants in the *negative evaluative expectancy* condition were told:

> In the packet of information in front of you, you will find a number of behaviours performed by an individual named Bob. Bob is a person that other people seem to dislike. Whenever anyone has anything to say about Bob, it usually seems to be bad. When I say begin, please turn the packet over and read each behaviour carefully as you try to decide for yourself whether Bob is a person you would like or would dislike.

Experiment 3

Participants

Participants were 96 undergraduates from the University of Wisconsin—Madison, and were tested one at a time. Participation in the experiment was one of the options that could be chosen to fulfil research requirements in an introductory psychology course.

Procedure

Only minor changes were made to the procedure used in Experiment 2. Participants in Experiment 3 were not asked to form an impression. Instead, they were given *memory set* instructions (for a similar manipulation, see Hamilton, Katz, & Leiner, 1980):

> The study you will be participating in today involves memorizing information about a person named Bob. Please concentrate on memorizing this information, as it is important for the experiment.

In addition, participants in Experiment 3 were not provided with trait expectancies *(no expectancy* condition). In Experiments 1 and 2, the stimulus presentation combinations were defined in terms of the consistency of the behaviours with the expectancy. Because no expectancies were provided in Experiment 3, the stimulus behaviour combinations are defined solely in terms of the behaviours that make up the combinations. Consequently, we used the four types of booklets to define two variables. The first is the *dominant behaviour* in the booklet (positive or negative); the second is the *stimulus behaviour combination* (number of behaviours consistent/ irrelevant/inconsistent with the dominant behaviour: 15/10/5 and 15/5/10). The data from Experiment 3 closely resembled the data from the 15/10/5 and 15/5/10 conditions used in Experiments 1 and 2. Hence, for purposes of the combined analyses, the data from each participant in Experiment 3 were assigned to one of those two conditions.

Using Data from the Three Experiments in a Single Analysis

We analysed all four measures using pooled within-participant hierarchical regression analyses (see Cohen & Cohen, 1983). Although cumbersome, within-participant

regression analysis provides considerable analytic power, and does not have ANOVA's disadvantages when confronted with missing data (a chronic problem in these studies) or missing cells (a result of combining data from the three experiments). In our analyses, each participant was assigned a dummy code, and these dummy codes were used in the regressions. The within-participant analyses were then conducted in a layered fashion, with an initial regression model including only main effects, a second regression model including both main effects and two-way inter-actions (interpreting only the interactions), a third model including main effects, two-way interactions and three-way interactions (interpreting only the three-way interactions), and so forth.

The predictors in the conditional probability and inter-item generation time analyses were *recall sequence* (consistent–consistent, consistent–inconsistent, inconsistent–consistent, inconsistent–inconsistent), *dimensionality* (within-trait dimension or between-trait dimension), *stimulus behaviour combination* (the numbers of consistent/irrelevant/inconsistent behaviours: 15/5/10, 15/10/5, 10/5/15 and 5/10/15), *number of exposures* (1, *2), initial expectancy* (negative, positive, or none), and *processing set* (impression, evaluation, memory). The first two predictors are within-participant predictors, and the remainder are between-participant predictors. In the analyses of the recall data, the recall sequence predictor was deleted from the analyses and *a recalled behaviour consistency* (consistent, irrelevant, inconsistent) predictor was added.

RESULTS

Raw and Adjusted Conditional Probabilities

We expected that adjusting the raw conditional probabilities for chance by calculating the CPDIFF measure would have three important effects on the data. First, we anticipated that adjusting the data for chance would generally heighten the difference in conditional probabilities between the within-trait dimension and the between-trait dimension recall sequences, with the within-dimension sequences having higher CPDIFF scores than the between-dimension sequences. This should occur because, on average, there were more between-trait items than within-trait items presented to participants, and thus, participants should be likely to recall more between-trait dimension items. Adjusting for this disparity should favour the within-trait dimension recall sequences.

Second, we expected that the CPDIFF measure would reveal differences in conditional probability among the various recall sequences, especially the within-trait dimension sequences. The reason for this expectation is straightforward: as we noted in the Introduction section of this article, the expected conditional probabilities of within-trait dimension consistent-consistent and inconsistent-inconsistent recall sequences should be lower than the expected conditional probabilities for the other two sequences. Thus, subtracting the expected conditional probabilities from the raw conditional probabilities should yield higher difference scores for sequences in which items come from the same trait category than for sequences in which items come from different trait categories. These adjusted probabilities would suggest that, given recall

of an item, it was easier to next recall an item that comes from the same trait category as the original item than an item from the opposite trait category. Such an outcome would be inconsistent with the CAM's assumption that inter-item associations in a person representation are just as strong for behaviours with implications for denotatively opposite trait categories as for behaviours with implications for the same trait category.

Finally, the behaviour combination manipulation presents different numbers of expectancy-consistent and expectancy-inconsistent items to different groups of participants. The net impact of this manipulation is to hold the number of within-dimension items relatively constant across the sequences, but to widely vary the number of between-dimension items presented. We expected that the raw conditional probabilities for the between-trait dimension sequences would be related to these variations. For example, when there were a large number of expectancy-congruent items presented, and few expectancy-incongruent items, the raw conditional probabilities for recall sequences in which the second item in the sequence was an expectancy-congruent item should be high, and the raw conditional probabilities for recall sequences in which the second item in the sequence was an expectancy-congruent item should be low. Adjusting the data for chance by calculating the CPDIFF measure should reduce or eliminate these effects.

The data were consistent with these three expectations. Table *2,* which presents a comparison of the raw conditional probabilities and the CPDIFF's for the theoretically important recall sequence x dimensionality interaction, provides one important example of the changes that occur when the conditional probability data are adjusted for chance. The dimensionality effect predicted by the CAM, in which within-trait dimension sequences are more likely than between-trait dimension sequences, appears in our data, but only for the CPDIFF measure, $F(1, 1713) = 138.44$, $p < 0.0001$. In fact, in the raw conditional probability data, the mean between-trait dimension sequence conditional probability is actually slightly higher ($M = 0.194$) than the mean within-trait dimension probability ($M = 0.171$), $F(1, 1713) = 6.24$, $p < 0.02$.

The raw conditional probability data in Table *2* also reveal a significant recall sequence x dimensionality interaction, $F(3, 1682) = 9.06$, $p < 0.0001$. The means for this interaction unexpectedly suggest that there were differences in mean conditional

Table 2. The conditional probabilities and adjusted conditional probabilities (CPDIFF) by levels of dimensionality and recall sequence

	Dimensionality	
Recall sequence	Between-trait dimension	Within-trait dimension
Conditional probabilities		
Consistent–consistent	0.182	0.218
Consistent–inconsistent	0.187	0.126
Inconsistent–consistent	0.226	0.145
Inconsistent–inconsistent	0.170	0.192
Adjusted conditional probabilities (CPDIFF)		
Consistent–consistent	-0.078	0.108
Consistent–inconsistent	-0.044	0.010
Inconsistent–consistent	-0.040	0.002
Inconsistent–inconsistent	-0.068	0.083

probability among the recall sequences. *Post-hoc* analyses showed that, although there were slight differences among the means of the four between-trait dimension sequences, $F(3, 676) = 2.78$, $p < 0.05$, the interaction was largely a function of differences among the within-trait dimension sequences $F(3, 805) = 11.79$, $p < 0.0001$. Pairwise *post-hoc* Tukey tests suggest that the means form two coherent groups. Specifically, the two sequences in which the items came from the same category are not significantly different from each other, the two sequences in which the items came from different trait categories are not significantly different from each other, but the two sequences in which items came from the same trait category were both significantly different from both of the sequences in which items came from different trait categories. Thus, within-dimension recall sequences in which both traits came from the same trait category (e.g. friendly–friendly) had higher conditional probabilities than sequences in which the items came from denotatively opposite categories (e.g. friendly–unfriendly), regardless of the relations of these behaviours to the expectancies.

That this pattern was statistically reliable in the raw data was a surprise to us. The conclusions offered in the Hamilton *et at.* (1989) article led us to believe that there would be no differences among the within-trait dimension recall sequences. However, when we attempted to reconcile our data with those obtained by Hamilton *et at.*, we discovered that reconciliation was unnecessary. A closer examination of their data suggests that they obtained almost the same pattern of conditional probabilities in their within-trait dimension condition as we obtained in our experiments. Collapsing across their two list composition conditions (see Hamilton *et at.*, 1989, Table 4), the means for the four recall within-trait dimension recall sequences were as follows (all means are unweighted): consistent–consistent $\boldsymbol{M} = 0.285$, consistent–inconsistent $\boldsymbol{M} = 0.230$, inconsistent–consistent $M = 0.230$, inconsistent–inconsistent $\boldsymbol{M} = 0.250$. Although the differences among the sequences are slightly smaller in their experiment than in our three experiments, the pattern is the same: given recall of an item, it is more likely that the next item recalled will come from the same trait category as the original item than from the denotatively opposite category.

Interestingly, Hamilton *et at.* (1989, pp. 933–934) did not report *any* inferential statistical tests of their conditional probabilities. Thus, it is unclear whether they based their conclusion of no among-sequence differences on an unreported analysis, or on a simple visual examination of the data. Certainly, given the results of our own statistical analyses, their conclusion that there were no differences among the recall sequences in their study must be regarded with some scepticism. When analysed individually, this pattern was significant in all three of our experiments. The appropriate conclusion seems to be that there is evidence for differences in conditional probabilities among the within-trait dimension recall sequences. That evidence exists in both our own experiments, and in the pattern of means obtained by Hamilton *et at.* Obviously, these data contradict the assertion of the CAM that no such differences should exist.

As indicated by the means in Table 2, adjusting the raw probabilities for chance by calculating the CPDIFF measure merely serves to sharpen the recall sequence x dimensionality interaction, $F(3, 1682) = 19.27$, $p < 0.0001$. As with the raw probabilities, *post-hoc* exploration of this interaction revealed that it was largely due to the effects occurring in the within-trait dimension condition. Indeed, separate examination of the mean adjusted conditional probabilities in the between-trait dimension

condition did not yield a significant main effect for recall sequence. By comparison, the recall sequence main effect in the within-trait dimension condition was significant, $F(3, 805) = 23.81$, $p < 0.0001$. As with the raw conditional probability data, *post-hoc* Tukey tests suggest that the means form two coherent groups. The two sequences in which the items came from the same category are not significantly different from each other, the two sequences in which the items came from different trait categories are not significantly different from each other, but the two sequences in which items came from the same trait category were both significantly different from both of the sequences in which items came from different trait categories.

The prediction that the CPDIFF measure helps to eliminate the effects of the different presentation lists is tested by the behaviour combination x recall sequence x dimensionality interaction. As expected, this interaction is significant in the raw conditional probability data, $F(7, 1616) = 2.47$, $p < 0.02$, but is not significant in the adjusted data, $F(7, 1616) = 0.71$, $p > 0.67$. Examination of the data confirms our expectations about the effects of the CPDIFF measure. For example, consider the conditional probabilities for the between-trait dimension consistent—inconstent recall sequence. When there are relatively few expectancy-inconsistent behaviours presented to participants (15/5/10 and 15/10/5 conditions), the raw conditional probabilities are relatively low ($Ms = 0.117$ and 0.170). However, when there are many expectancy-inconsistent behaviours (10/5/15 and 5/10/15 conditions), the raw probabilities are quite high ($Ms = 0.248$ and 0.303). As one would expect, by accounting for differences in the number of the different types of items recalled (which, of course, reflects the number of items presented), the CPDIFF measure drastically reduces these differences ($Ms = -0.019, -0.061, -0.083, -0.028$, respectively). A similar outcome occurs for the between-trait dimension incongruent—congruent recall sequences, which show substantial differences in the raw conditionals across conditions ($Ms = 0.291, 0.268, 0.163, 0.072$). The first two means are for conditions in which there were 15 expectancy-congruent behaviours, while the latter two conditions had only 10 and five. As before, these large behaviour combination differences are virtually eliminated by the CPDIFF adjustment for chance ($Ms = -0.014, -0.051, -0.052, -0.056$, respectively).[1]

Few other significant results of the analyses were of theoretical significance. One of these was a significant instruction x dimensionality interaction on the CPDIFF measure, $F(1, 1682) = 4.31$, $p < 0.04$. The means for this weak effect suggest that the evaluative set instructions may have reduced dimensionality effects in CPDIFF scores (between-trait dimension $M = -0.043$, within-trait dimension $M = 0.027$) relative to the impression set (between-trait dimension $M = -0.059$, within-trait dimension $M = 0.059$) and memory set (between-trait dimension $M = -0.054$, within-trait dimension $M = 0.051$) conditions. This could indicate that participants did pay more attention to the evaluative implications of the items in the evaluative set condition. However, the means for the significant number of repetitions x instruction x dimensionality interaction, $F(1, 1616) = 6.00$, $p < 0.02$, qualify this effect, suggesting that it occurred only when participants had the opportunity to view the behaviour list twice.

[1]It should be noted that the patterns of adjusted conditional probabilities cannot be attributed to the idiosyncratic effects of the CPDIFF measure. We also explored a somewhat cruder adjustment measure than the CPDIFF measure; this measure simply involved dividing the obtained conditional probability for a sequence by the number of behaviours of each type that were presented. The overall pattern of data for this alternative adjusted measure resembles the pattern for the CPDIFF measure.

The impact of the number of exposures on the organization of the recall list was also apparent in two more interactions involving the CPDIFF measure. The first of these, a number of exposures x dimensionality interaction, $F(1, 1682) = 5.58$, $p < 0.02$, suggests that repeated exposure to the list strengthened the overall tendency for the recall list to be organized by trait dimension (one exposure: between-trait dimension $M = -0.045$, within-trait dimension $M = 0.037$; two exposures: between-trait dimension $M = -0.063$, within-trait dimension $M = 0.064$). The means for the recall sequence x number of exposures x dimensionality interaction, $F(3, 1616) = 3.62$, $p < 0.02$, simply suggest that this pattern occurred more strongly for the consistent—consistent and inconsistent—inconsistent sequences than for the other two recall sequences.

In theory, the amount of time that passes between the end of one item in the recall list and the start of the next item should yield results that parallel the CPDIFF measure. That is, both of these measures ought to assess the ease of item retrieval, so that when the CPDIFF measure is high, generation time ought to be low, and vice versa. We were particularly interested in two effects in the time data: a dimensionality effect indicating that generating times were faster for within-trait dimension recall sequences than for between-trait dimension sequences (replicating Hamilton *et al.*, 1989), and a recall sequence x dimensionality interaction suggesting that it was easier to recall an item from the same trait category as the previously recalled item than an item from the denotatively opposite trait category. Analyses of the individual studies indicated that the dimensionality effect emerged quite readily. However, although the means were consistent with the hypothesized recall sequence x dimensionality inter-action, this interaction was not statistically reliable in any of the individual analyses. We hoped that increasing statistical power by combining the results of the three experiments would yield a significant interaction.

However, power was not the only obstacle in these data. Within each experiment, the distribution of inter-item generation times was positively skewed, and some of the times were quite long (one was over 3 minutes long). Obviously, such long generation times and severe non-normality in the distribution of times cause analytic problems. To minimize the impact of the long times and to normalize the distributions, we analysed the data using both a logarithmic transformation and an inverse transformation (1/(1 + time), see Fazio, 1990). Both of these transformations yielded similar results, but for simplicity we focus on the logarithmic transformation data.

After doing a logarithmic transformation on each inter-item generation time, the transformed times were subjected to pooled within-participant hierarchical regression analyses. The factors in these analyses were the same as the ones used in the analyses of the conditional probabilities.

The means (back-transformed to time for easier interpretation) depicted in Table 3 suggest that the generation time data do indeed support the results of the CPDIFF measure. As expected, times were faster when an item came from the same trait dimension as a previous item ($M = 3.13$ seconds) than when it came from a different trait dimension ($M = 6.11$ seconds), $F(1, 1729) = 94.45$, $p < 0.0001$. Furthermore,

Table 3. The inter-item generation times (back-transformed from the logarithmic transformation, in seconds) by levels of dimensionality and recall sequence

	Dimensionality	
Recall sequence	Between-trait dimension	Within-trait dimension
Consistent–consistent	6.42	2.68
Consistent–inconsistent	5.52	3.97
Inconsistent–consistent	6.23	4.05
Inconsistent—inconsistent	6.17	2.89

the means for the significant recall sequence x dimensionality interaction, $F(3, 1698) = 2.92$, $p < 0.04$, suggest that inter-item generation times are faster when both items come from the same trait category than when they come from denotatively opposite categories. However, it should be noted that, despite the fact that we combined the data from multiple experiments and used appropriate data trans-formations, this interaction is relatively weak. Thus, no pairwise *post-hoc* comparisons among the means are statistically reliable. Nonetheless, the data suggest that the two effects of greatest theoretical importance that emerged in the CPDIFF data also emerge in the time data.

Interestingly, Hamilton *et al.* (1989, Experiment 1, Table *5)* obtained similar results. The means that they present also suggest that recall of the second item in a recall sequence was significantly faster if the two items were from the same trait dimension (unweighted $M = 4.00$ seconds) than from different trait dimensions (unweighted $M =$ 10.60 seconds), an outcome that we replicated in our studies. However, because the recall sequence x dimensionality interaction was nonsignificant in their analyses, one might suggest that our data diverge from theirs. Examination of the means suggests that this conclusion is incorrect. The time means from Hamilton *et al.* Experiment 1, are again strikingly similar to ours: times were faster when both items came from the same trait category (unweighted $M = 4.47$ seconds), and slower when the items came from denotatively opposite categories (unweighted $M = 6.65$ seconds), regardless of the initial expectancy. In our view, the nonsignificance of the recall sequence x dimensionality interaction in their analyses is understandable, given that: (a) the time data are quite messy (this interaction was also not statistically reliable in any of our individual analyses), (b) compared to our regression analyses, Hamilton *et al.* used a relatively low-power analysis (a `pseudosubject' ANOVA), and (c) they had many fewer participants than we had after we combined our experiments. Thus, in our view, there is essential convergence between our data and the data of Hamilton *et al.* The inter-item generation times both support the CPDIFF data reported earlier, and contradict the predictions of the CAM.

Only two other outcomes of the time analyses are of potential interest. The first of these is a main effect for number of exposures, $F(1, 243) = 4.52$, $p < 0.04$, indicating that inter-item generation times were faster in the two-exposure condition *(M = 3.92)* than in the one-exposure condition *(M = 4.93).* This effect suggests that retrieval of items was easier after two list exposures, possibly reflecting the heightened accessibility of the items.

A second effect of possible interest was a main effect for instructional set, $F(1, 243) = 9.31$, $p < 0.01$. The means for this effect indicate that inter-item generation times were fastest in the memory set condition, *(M = 3.65),* middling in the impression

set condition, (*M* = 4.31), and slowest in the evaluative set condition (*M* = 6.15). The interpretation of this unexpected effect is unclear. However, one interpretation may be that the slowing of generation times in the impression and evaluation conditions may reflect the formation of multiple associations between items. Research into the fan effect (Anderson, 1974) suggests that multiple associates to an item may slow retrieval of any one of those associates. Because forming impressions is thought to cause people to form a coherent person representation, whereas simply memorizing items does not (Hamilton *et al.*, 1980), the time data might suggest that items in the evaluation and impression conditions might have more links to other items than those in the memory condition. These multiple links may be responsible for the slowing of inter-item generation times observed in these conditions. However, this notion is highly speculative: the remainder of the data from our experiments do not suggest that such inter-item links were formed. Hence, interpretation of this effect remains unclear.

Recall

The proportion of expectancy-consistent, expectancy-irrelevant, and expectancy-inconsistent behaviours recalled by each participant were calculated and analysed using the pooled within-participant regression analyses described in the Methods section of this article. The most important potential effect in this analysis concerns recall of the different types of behaviours (consistent, irrelevant, inconsistent). Person memory research has frequently found evidence of increases in recall for expectancy-inconsistent items (e.g. Hastie, 1980; Srull, 1981). This effect has been interpreted as evidence for reconciliatory processes: that is, participants attempt to understand how targets could perform behaviours that are inconsistent with both the expectancy and with other behaviours in the set. However, this inconsistency effect has not been present in previous multiple-expectancy studies (e.g. Driscoll, 1992; Hamilton *et al.*, 1989).

The results from our three experiments confirm the absence of an inconsistency effect in recall with multiple expectancies. Although there was a main effect for the recalled behaviour consistency factor in our analyses, $F(2, 766) = 5.27$, $p < 0.01$, the means indicate that the expectancy-consistent behaviours were better recalled (*M* = 0.427) than the expectancy-inconsistent behaviours (*M* = 0.395). The expectancy-irrelevant behaviours were most poorly recalled (*M* = 0.386). From these data, it seems reasonable to conclude that reconciliation of inconsistency was probably not occurring at a high level.

The absence of interactions in these analyses indicates that this recall pattern did not vary substantially across our three experiments. Only two interactions involving the recalled behaviour consistency factor were significant. The first of these involved behaviour combination, $F(6, 752) = 7.71$, $p < 0.0001$. The means for this interaction indicate that set size effects were occurring. For both expectancy-consistent and expectancy-inconsistent items, the proportion of items recalled was highest when the number of behaviours of that type presented was smallest. The second interaction involving the recalled behaviour consistency factor was with the number of exposures, $F(2, 752) = 3.44$, $p < 0.04$. The means for this interaction indicate that the recall advantage that occurred with two exposures (number of exposures $F(1, 348) = 37.78$,

$p < 0.0001$) was greatest for expectancy-irrelevant items ($M_{diff} = 0.151$), middling for expectancy-consistent items ($M_{diff} = 0.114$), and smallest for expectancy-inconsistent items ($M_{diff} = 0.083$). This pattern of results suggests that the expectancy manipulation may have had a more substantial effect on the types of items recalled in the one-exposure condition than in the two-exposure condition. No other effects or interactions of theoretical interest emerged from the analyses.

GENERAL DISCUSSION

The data from our three experiments have both methodological and theoretical implications. The methodological implications are straightforward: raw conditional probabilities are flawed as an index of associative strength, and must be corrected for chance before they can be interpreted correctly. The conditional probability results that we present in this article illustrate this point in three ways. First, the greater ease of recalling sequential within-trait dimension items rather than between-trait dimension items would have been missed in our data, if not for the CPDIFF correction. Second, the recall sequence x dimensionality interaction demonstrating differences in the conditional probabilities associated with the within-trait dimension recall sequences was sharpened and clarified by application of the CPDIFF correction. Third, the CPDIFF correction eliminated the effects of varying the number of between-trait dimension items presented on the between-trait dimension conditional probabilities.

Our results have two major theoretical implications. The first of these applies to Hamilton *et al.'s* (1989) Complete Associative Model of person representation. That model predicts that there ought to be no differences in the conditional probabilities among the four within-trait dimension recall sequences used in our experiments, and no differences in inter-item generation times as well. Our three experiments demonstrated that such differences emerge, even in the uncorrected raw conditional probabilities. The time results are weaker than the conditional probability results, but the message from both measures is the same: after retrieval of an item, it is easier to retrieve an item that comes from the same trait category as the initial item than an item that comes from a different trait category. Moreover, this pattern is not only evident in our data; it was evident in Hamilton *et al.'s* original data. We were able to detect these effects because we used more participants than Hamilton *et al.* and also used more powerful statistical analyses.

Given the inability of the CAM to describe our results, it seems reasonable to consider other theoretical alternatives. Two alternatives, one based on the spread of excitation in associative networks and one based on retrieval strategies, are reasonable possibilities.

The network-style model would suggest that a relatively sparse person representation of an actor was created in our experiments, and that the items in this person representation remain linked to the naturally occurring semantic network. In the sparse person representation, the items are connected to the person node (Bob), but not to each other. Thus, on recall of an item, people may follow a link from the item to the trait that is linked to the item in the naturally-occurring semantic network (e.g. helped an elderly lady cross the street_____kind), and then follow a pathway back

down from the trait category to another item (e.g. kind—helped his brother mow the' lawn). This should be relatively easy (high-probability) and quick. In comparison, to retrieve an item from the opposite trait category, one has to follow the pathway from the initially recalled item to the trait to which the item is linked (e.g. helped an elderly lady cross the street—kind), then follow a link between the trait concept and it opposite (kind–unkind), then another link down to an item linked to that new tram (e.g. unkind _____ stole candy from a baby). This should be harder (moderate proba bility), and should take longer. The same process can be used to describe retrieval o two items from different trait categories. However, because the link between a trail and its opposite is likely to be stronger than the link between a trait and a semantically unrelated trait, retrieval of items from denotatively unrelated trait categories should be least probable, and should take the most time. Hence, this process can account for all the data obtained in our experiments: (a) no incongruity effect in recall (few inter-I item linkages); (b) same-trait recall sequences that are both fast and relatively high in probability; (c) opposite-trait recall sequences that are only moderately fast and moderate in probability; and (d) denotatively unrelated recall sequences that are ~ relatively slow and low in probability.

The effects described in the paragraph above are relatively passive, occurring because of the spread of excitation in an associative network. A second possibility is that our conditional probability results occurred because of the use of a conscious retrieval strategy. That is, people may explicitly use the trait implications of an item to recall a subsequent item (That was a kind item, what other kind items are there?'), and if that fails, they may switch to another trait. Because of the natural semantic and associative linkages among traits, the next trait examined will likely be the denotative opposite of the trait that was initially used (I'm out of kind items ... how about unkind ones?'). If this process fails to yield item retrieval, people may then proceed to another denotatively unrelated trait. The outcomes obtained in our experiments are consistent with this conscious retrieval strategy.

We emphasize that these probabilities seem to be the two that were most reasonable to us. Certainly, other mechanisms might be invoked (e.g. see Klein & Loftus, 1990). It remains to future research to examine in more detail the mechanisms that might explain the outcomes that we obtained in our experiments.

A second theoretical implication of our results is broader, and pertains to the idea that the conditional probabilities in single-trait expectancy studies reflect inconsistency resolution. Several of these single-expectancy studies have found an increase in the consistent–inconsistent conditional probability, and this increase has classically been interpreted as evidence of inconsistency resolution. However, Skowronski and Welbourne (1997) argue that this effect might be a statistical artifact of an inconsistency effect in recall. That is, if a greater number of expectancy-inconsistent than expectancy-consistent items are recalled, then the conditional probability of consistent–inconsistent recall sequences ought to be the highest of the four, *simply by chance.* One of the paradoxes of the inconsistency-resolution explanation for the conditional probabilities obtained in single-expectancy experiments is why heightened conditional probabilities have emerged in the consistent–inconsistent sequence, but not in the inconsistent–consistent sequence. Both sequences would seem to be reasonably related to reconciliation among disparate items, but it is only the consistent–inconsistent sequence that has generally seen an increase in conditional probability. Skowronski and Welbourne (1997) suggest that this disparity is a

natural consequence of an inconsistency effect in recall it would be expected by chance. Our results provide suggestive evidence favouring this artifact hypothesis. However, more direct tests are obviously desirable, and these direct tests are currently underway.

CONCLUSION

In the last few years, the area of person memory has been relatively quiescent. The results that we present in this paper, and the associated conjectures, suggest that this quiescence may soon come to an end. The person memory models that dominated the 1980s (as prototypically represented by Srull & Wyer, 1989; Wyer & Srull, 1989) may have been built on a faulty foundation: raw conditional probabilities. If our suspicions in this matter are correct, then the field may have to return to the drawing board in an attempt to understand how information about a person is represented in memory, and how these representations affect recall. It promises to be an interesting time.

ACKNOWLEDGEMENTS

We thank David Hamilton for sending us the stimulus materials that he and his colleagues used in their research. We thank the participants in the 1993 and 1997 Duck Conference on Social Cognition (also known as Duck U.) for their feedback on the studies described in this article. We thank Denise Driscoll, Donal Carlston, Eliot Smith, the Social Cognition group at The Ohio State University (particularly William von Hippel and Marilynn Brewer) and the anonymous reviewers of this article for their helpful commentary on earlier drafts of this paper. Special thanks to Jim Uleman for his suggestions about adjusting the raw probabilities.

This manuscript was supported by NIMH grant RO1-MHSO730-01A1.

REFERENCES

Anderson, J. R. (1974). Verbatim and propositional representation of sentences in immediate and long-term memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 149-162.

Cohen, J., & Cohen, C. (1983). *Applied multiple regression/correlation for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Driscoll, D. M. (1992). Multi-trait person impressions. Unpublished doctoral dissertation,University of California at Santa Barbara, Santa Barbara, CA.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick, & M. S. Clark (Eds), *Review of personality and social psychology: Research methods in personality and social psychology* (Vol. 11, pp. 74-97). Newbury Park, CA: Sage Publications.

Hamilton, D. L., Driscoll, D. M., & Worth, L. T. (1989). Cognitive organization of impressions: Effects of incongruency in complex representations. *Journal of Personality and Social Psychology, 57*, 925-939.

Hamilton, D. L., Katz, L. B., & Leirer, V. O. (1980). Organizational processes in impression formation. **In R.** Hastie, T. M. Ostrom, E. B. Ebbesen, R. S. Wyer, Jr., D. L. Hamilton, & D. E. Carlston (Eds), *Person Memory: The cognitive basis of person perception* (pp. 121–153). Hillsdale, NJ: Erlbaum.

Hastie, R. (1980). Memory for behavioral information that confirms or contradicts a personality impression. **In R.** Hastie, T. M. Ostrom, E. B. Ebbesen, R. S. Wyer, Jr., D. L. Hamilton, & D. E. Carlston (Eds), *Person memory: The cognitive basis of social perception* (pp. 155–177). Hillsdale, NJ: Erlbaum.

Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology, 37*, 25–38.

Klein, S. B., & Loftus, J. (1990). Rethinking the role of organization in person memory: an independent trace storage model. *Journal of Personality and Social Psychology,* ***59,*** 400–410.

Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. *British Journal of Social Psychology,* **31,** 81–109.

Seta, C. E., & Hayes, N. (1994). The influence of impression formation goals on the accuracy of social memory. *Personality and Social Psychology Bulletin,* **20,** 93–101.

Skowronski, J. J., & Welbourne, J. (1997). Conditional probability may be a flawed measure of associative strength. *Social Cognition,* ***15,*** 1–12.

Srull, T. K. (1981). Person memory: Some tests of associative storage and retrieval models. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 440–462.

Srull, T. K. (1983). Organizational and retrieval processes in person memory: An examination of processing objectives, presentation format, and the possible role of self-generated retrieval cues. *Journal of Personality and Social Psychology,* **44,** 1157–1170.

Srull, T. K., & Wyer, R. S. Jr. (1989). Person memory and judgment. *Psychological Review,* ***96,*** 58–83.

Srull, T. K., Lichtenstein, M., & Rothbart, M. (1985). Associative storage and retrieval processes in person memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **11,** 316–345.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-consistent and expectancy-inconsistent information: A review of the social and social development literatures. *Psychological Bulletin,* **111,** 42–61.

Wyer, R. S., Jr., & Srull, T. K. (1989). *Memory and cognition in its social context.* Hillsdale, NJ: Erlbaum.

