Raw Conditional Probabilities Are a Flawed Index of Associative Strength: Evidence From a Single Trait Expectancy Paradigm

John J. Skowronski
Department of Psychology
The Ohio State University at Newark

Katherine Gannon
Department of Psychology
Texas Tech University

This article reports the results of an experiment testing Skowronski and Welbourne's (1997) proposition that an elevated conditional probability for the consistent-inconsistent recall sequence in the memory incongruity paradigm can occur due to chance. Participants in the experiment were given a trait expectancy about a target, then were asked to read sentences describing the target's behaviors. Later, participants attempted to recall those behaviors. As expected, results indicated an incongruency effect in recall: Participants recalled greater numbers of expectancy-inconsistent behaviors than expectancy-consistent behaviors. Also as expected, a heightened conditional probability was observed for the consistent-inconsistent recall sequence. However, this advantage disappeared when two other sequencing measures, the Conditional Probability Difference score and the Adjusted Ratio for Individual Sequences index, were used. Both of these indexes correct for chance orderings, suggesting that the elevated conditional probability observed for the consistent-inconsistent recall sequence in this paradigm is an artifact. This outcome also suggests that models of the incongruity effect that were based solely on associative links among items in a person representation may be, at best, incomplete.

Researchers have spent much effort attempting to understand how a priori impressions and expectations affect subsequent memory for a target person's characteristics. Early theorizing equated trait impressions or expectations with schemas, and predicted person memory results similar to those obtained by schema researchers (see Alba & Hasher, 1983). Initial results were straightforward. Perceivers with prior impressions or expectations about a target (a) preferentially remembered or recognized target characteristics that were consistent with the impression or expectation (compared to expectation-irrelevant characteristics; Rothbart, Evans, & Fulero, 1979) and (b) evinced impression-directed memory distortion, reconstruction, and guessing (Clark & Woll, 1981; C. Cohen, 1981).

The focus of this research was altered by Hastie and Kumar (1979). In apparent conflict with schema theory, Hastie and Kumar reported that behaviors that were inconsistent with a trait expectation were sometimes more likely to be recalled than expectancy-consistent behaviors. These findings generated substantial interest (see Rojahn & Pettigrew, 1992; Stangor & McMillan, 1992, for reviews and meta-analyses).

To explain Hastie and Kumar's data, Hastie (1980) suggested that incongruity effects in recall were due to enhanced processing of expectancy-inconsistent items. Hastie speculated that this processing was either a consequence of explaining the inconsistency (e.g., by engaging in attributional thinking), or of reconciling expectancy-inconsistent items with expectancy-consistent items (e.g., How could a nice person who donates money to charity sell cocaine to children?). Subsequent theory and research (e.g., Srull, 1981) has focused on the latter explanation.

More specifically, according to this latter approach, the incongruity effect in recall occurs because of associative linkages among items in the mental representation of the target. Reconciliatory activity supposedly produces numerous links between expectancy-inconsistent items and other items (see Figure 1). Attempts at recalling information use these...
links. Retrieval is thought to begin at the person node and to move down one of the pathways to an item in the network. Once the item is retrieved, it can cue recall of another item to which it is linked. Thus, as the number of links leading to a target item increases, the probability of recalling the target item also increases. Because expectancy-inconsistent items have a greater number of associative links than other items, they are more likely to be recalled.

This associative linkage explanation gained popularity because it leads to novel predictions about recall orderings, and these predictions appeared to be supported by the data. For example, because of the associative linkages formed as a result of reconciliation, it was thought that recall of an expectancy-consistent item was especially likely to be followed by recall of an expectancy-inconsistent item. This prediction is intriguing, for it violates the common sense expectation that an item ought to be followed in the recall list by an item with the same trait implications as the first. The results of several experiments that used a conditional probability measure as a dependent variable (e.g., Srull, 1981; Srull, Lichtenstein, & Rothbart, 1985) showed that consistent-inconsistent recall sequences did, indeed, have higher conditional probabilities than consistent-consistent recall sequences, apparently confirming the incongruity prediction.

However, in a recent article, Skowronski and Welbourne (1997) criticized the conditional probability measure, arguing that the measure imperfectly assesses the associative strength between items. Using several hypothetical examples, Skowronski and Welbourne demonstrated that the conditional probability measure can be dramatically affected by factors irrelevant to the associative strength between items, such as the overall number of items that are recalled and the exact content of the recall list.

Skowronski and Welbourne (1997) noted that these artifacts might cause incorrect interpretations of raw conditional probability data, and they cited the literature on the incongruity effect as one example of how this might occur. In several hypothetical examples, Skowronski and Welbourne demonstrated that the elevated conditional probability that accompanied the consistent-inconsistent recall sequence in prior studies could have occurred simply as a consequence of the fact that participants recalled a greater number of expectancy-inconsistent than expectancy-consistent items. Furthermore, comparison of calculated expected conditional probabilities to the actual conditional probabilities obtained in several studies often revealed a good fit between the actual probabilities and the expected probabilities. Hence, Skowronski and Welbourne argued that these raw conditional probabilities may not reflect a bias in the formation of linkages to expectancy-inconsistent items; instead, these probabilities are what would be expected by chance if the items were simply recalled in a random order.

The theoretical implications of this argument are important, potentially striking at the core of linkage-based explanations of the incongruity effect in recall. These theories have postulated that a bias in the formation of inter-item associative linkages during the construction of a person representation is responsible for the incongruity effect in recall, and the conditional probability data have been crucial in establishing that a bias exists in these inter-item linkages. If these conditional probability data are flawed and do not reflect a bias toward the formation of linkages to expectancy-inconsistent items, then linkage-based explanations for the incongruency effect in recall are seriously challenged.

To better understand the gist of the Skowronski and Welbourne (1997) argument, consider the following simpli-
fied example. Assume that a participant in an experiment is given a trait expectancy about a hypothetical target, and later reads sentences describing the target. On a subsequent recall task, assume that the participant recalls four expectancy–consistent sentences, three expectancy–consistent sentences, and two expectancy–irrelevant sentences. If the first item recalled is an expectancy–consistent item, which item type is most likely to be recalled as the second item in the list? If items are simply recalled in a random order, an expectancy–inconsistent item is most likely to be recalled next: After all, there are four of those available. In comparison, there are only two expectancy–irrelevant and two expectancy–consistent items available (remember that one expectancy–consistent item has already been recalled). This example illustrates that, if an incongruity effect occurs in recall, then it is likely that the conditional probability of the consistent–inconsistent recall sequence will be quite high, simply by chance.

Skowronski and Welbourne (1997) suggested that raw conditional probabilities must be adjusted for these chance expectations before the probabilities can legitimately be interpreted. They proposed two adjustment methods. The first of these adjustment methods involves the calculation of a Conditional Probability Difference (CPDIFF) score, which is simply the difference between the actual conditional probability and the expected conditional probability (the expected probability is calculated as if the recall ordering was random). A CPDIFF score of zero indicates that the conditional probability obtained is exactly what it should be if recall orderings were chance; values greater than zero reflect conditional probabilities that are greater than chance, and values less than zero reflect conditional probabilities that are below chance. The second adjustment method involves the calculation of an Adjusted Ratio for Individual Sequences (ARIS) score, based on the well-known Adjusted Ratio of Clustering score (see Roenker, Thompson, & Brown, 1971). The ARIS score is a ratio formed by dividing (a) the difference between the number of repetitions of a sequence and the number of repetitions expected by (b) the difference between the maximum possible number of repetitions of a sequence and the number expected. An ARIS score of zero suggests that the number of repetitions of a sequence is exactly as would be expected by chance. Values greater than zero represent the occurrence of sequences more often than expected, but expressed as a proportion of the maximum number of above-chance sequences possible. For example, if four occurrences of a sequence were expected, eight were possible, and a participant had five, the ARIS score would be (5 − 4) / (8 − 4) or .25. Values below zero represent the occurrence of a sequence at below-chance rates.

The experiment reported in this article attempted to examine whether the raw conditional probability results obtained in earlier research using the incongruity memory paradigm always reflect associations between items in the person representation, or whether such effects can merely be a consequence of chance, driven by the incongruity effect in recall. The experiment used the CPDIFF score and ARIS index to correct for chance, so that the results obtained from these two indexes could be compared to the results provided by calculation of the raw conditional probabilities. In addition, the results provided by the two chance-corrected indexes could also be compared.

The outcomes on these two chance-corrected conditional probability indexes are theoretically important. If explanations of the incongruity effect in recall that are based on a bias in linkages to expectancy–inconsistent items are to be believed, then a heightened conditional probability of consistent–inconsistent recall sequences must be present in the data, even when the probability data are adjusted for chance. The presence of an incongruity effect in the recall data without evidence of such a linkage bias would suggest that incongruity effects in recall are not solely caused by linkage biases.

Research by Skowronski, Betz, Sedikides, and Crawford (1998) using the CPDIFF score has already demonstrated that apparently meaningful recall sequence raw conditional probabilities may not be meaningful at all, but instead can reflect chance recall orderings. However, these results were obtained in the context of a multi-trait expectancy study, and there is relatively little evidence of incongruity effects in recall or incongruity resolution in such studies (e.g., see Driscoll, 1992; Hamilton, Driscoll, & Worth, 1989). Hence, the results of Skowronski et al. (1998) may not appear in circumstances where incongruity effects in recall and evidence for incongruency resolution routinely occur. Furthermore, Skowronski et al. did not use the ARIS index, and it is not known whether that index will yield results identical to that provided by the CPDIFF score. The analyses of Skowronski and Welbourne (1997) suggested that there could sometimes be differences.

METHODS

Participants

In return for partial course credit, 107 participants from introductory psychology classes at Texas Tech University completed the experiment. Two participants failed to follow instructions, and their data were dropped from all analyses. Of the remaining 105 participants, 53 were men and 52 were women.

Stimulus Material and Design

Following the procedure used by Srull (1981), a short description of the target was provided to participants prior to the presentation of behavioral information about the target.
Three different trait dimensions were used in this study (sociability, intellectual orientation, meticulousness), with two specific types of expectancies (positive and negative) for each dimension (e.g., intellectual expectancy or unintellectual expectancy).

The expectancy paragraphs all followed a similar format (following Srull, 1981). Examples of this format are provided in the following intellectual and unintellectual expectancy paragraphs.

**Intellectual expectancy.** Some people say that Lily is basically an intellectual person who likes to engage in pursuits that are thought-provoking and that help her gain knowledge. She generally is not fond of intellectual pursuits that she considers to be intellectually inferior.

**Unintellectual expectancy.** Some people say that Lily is basically an unintellectual person who likes to do things that are fun and that don’t ask her to think too hard. She generally is not fond of intellectual pursuits that ask her to think and that she considers to be boring.

Behaviors appeared in a booklet, one per page. Each booklet contained five behaviors consistent with the expectancy, five behaviors inconsistent with the expectancy, and five behaviors irrelevant to the expectancy. Examples of behaviors used for the target described by either an intellectual expectancy or an unintellectual expectancy are as follows:

- Has several posters of rock stars hanging on her living room wall. (unintellectual)
- Spends most of her spare time shopping. (unintellectual)
- Played chess with her brother on the weekend. (intellectual)
- Went to the opera at least once a month. (intellectual)
- Took food out of the cabinet and fed her dog. (neutral)
- Ate breakfast before going to work. (neutral)

The behaviors in each booklet were presented in a random order unique to each participant.

**Procedure**

Two to eight individuals participated in each experimental session. Participants were seated at individual tables that were relatively isolated, so that interaction between participants was minimized.

The participants were told that the experiment generally concerned the way people typically process social information. They were informed that they would be reading behaviors performed by a woman named Lily. They were asked to form simple judgments about her, which they would be asked about at a later time. Participants were instructed that they would have 6 s to read each of the behaviors. The experimenter then read one of the six expectancy statements. Participants were instructed to begin reading the behavior on the first page and to turn to the next behavior when they heard a beep. As in Srull (1981), the beep tape provided 6 s of reading time per behavior.

When participants had finished reading the last behavior in their booklets, they were asked to try and remember the names of all 50 states in the United States and their capitals. They were instructed to record these on a blank sheet of paper. After 10 min, the participants were instructed to remember as many of Lily’s behaviors as they could. Participants were given 7 min to recall the behaviors. Finally, participants were thanked and debriefed.

**RESULTS**

The items that were recalled by each participant were independently examined by an undergraduate coder who was naïve to the hypotheses of the experiment, and by one of the authors of this article. Items were scored as correctly recalled if the participant’s report captured the trait-relevant gist of the sentence. Participant reports that did not capture this gist, or that reflected other information (e.g., a trait judgment about the target), were scored as intrusions. The two coders agreed on the classification of 96% of the items (Cohen’s $k = .94$); inconsistencies were resolved through discussion.

**Recall Content**

The number of items recalled by each participant was entered into a 3 (trait dimension: sociability, intelligence, meticulousness) $\times$ 3 (item type: expectancy-consistent, expectancy-inconsistent, expectancy-irrelevant) mixed analysis of variance (ANOVA) with repeated measures on the latter factor. The results of prior research suggest that an item type main effect should emerge. A typical pattern of recall in incongruity experiments is for expectancy-inconsistent items to be better recalled than both expectancy-consistent and expectancy-irrelevant items. A second outcome sometimes observed is for expectancy-consistent items to be better recalled than expectancy-irrelevant items.

The results of the analysis were consistent with these prior findings. Expectancy-inconsistent items were recalled most frequently ($M = 3.08$), expectancy-consistent items were re-

---

1 To calculate the kappa statistic (J. Cohen, 1960), each coder independently placed each item into one of four categories: expectancy-consistent, expectancy-inconsistent, expectancy-irrelevant, and intrusion (e.g., listed item was not presented). Inspection of these classification data revealed that the most frequent disagreement between coders occurred when one coder placed an item into the intrusion category and the other coder placed the item into one of the other three categories. Other disagreements (e.g., one coder placing an item in the expectancy-consistent category and the other in the expectancy-inconsistent category) were relatively rare.
called next most frequently \((M = 2.44)\), and expectancy–irrelevant items were recalled least frequently \((M = 2.20)\), \(F(2, 204) = 19.53, p < .0001\), mean square error \((MSE) = 1.10\), partial \(\eta^2 = .16\.

The results of planned pairwise comparisons indicated that only the mean for the expectancy–inconsistent condition was significantly different from the means for the other two conditions; smaller of the two significant \(F\)'s \((1, 204) = 6.53, p = .01\); nonsignificant \(F(1, 204) = .92, p = .34\). No other main effects or interactions emerged from the analysis.

Recall Sequence Analyses: Conditional Probabilities, CPDIFF, and ARIS

The order in which participants listed the items they recalled was used to calculate the conditional probability measure, the CPDIFF score, and the ARIS index for all recall sequences. The CPDIFF score and the ARIS index were calculated as described in Skowronski and Welbourne (1997); see p. 4 for a description of how to calculate the expected probabilities used in the CPDIFF measure; see pp. 9–10 for a description of how to calculate the ARIS index.

Because of the fact that it well tolerates missing data, hierarchical pooled within-participant regression, rather than ANOVA, was used for our inferential analyses of these three measures (see Cohen & Cohen, 1983, chap. 11, for more information on this regression technique). Hierarchical pooled within-participant regression can be thought of as the regression analog of a repeated measures ANOVA. As in within-participant ANOVAs, these pooled within-participant regression analyses partial the total within-participant variance into the following: (a) variance due to participants, (b) variance due to the effects of interest, and (c) error variance. They do so by assigning a unique dummy code to each participant, allowing the creation of a participants factor in the regression models. This participants factor consumes 102 degrees of freedom (number of participants [105] minus number of conditions [3]) in our models. The net effect of these procedures is to adjust each observation for a participant’s overall response tendency. Alternatively, another way to think about this procedure is that in evaluating the relation between the predictors of interest and the dependent measures, the regression partials out the tendencies that participants might exhibit in their responses.

The hypotheses generated by associative linkage explanations for incongruity effects in recall focus on comparisons of the conditional probabilities among sequences with a common initial item in the recall sequence (e.g., see Wyer & Srull, 1989, pp. 185–186). Hence, we conducted three sets of regression analyses. The first set focused on recall sequences in which the first item recalled was an expectancy–consistent item, the second set focused on recall sequences in which the first item recalled was an expectancy–inconsistent item, and the third set focused on recall sequences in which the first item recalled was an expectancy–irrelevant item. To explore whether the effects obtained in the first set of analyses were consistent across the three trait dimensions used in the study, we conducted a second set of regression analyses focusing on possible Recall Sequence × Trait Dimension interactions.

**Recall sequences beginning with an expectancy-consistent behavior.** The typical finding in the incongruity literature, as well as the theoretically predicted outcome (see Wyer & Srull, 1989, p. 185), is that the conditional probability for the consistent–inconsistent recall sequence is higher than the conditional probabilities for the consistent–consistent and consistent–irrelevant recall sequences. The conditional probability means for the recall sequences presented in Table 1 are consistent with both theory and this previous research, and the regression analyses of the data summarized by these means yielded a significant sequence effect, \(F(2, 179) = 8.66, p = .0003, \text{MSE} = .13\), partial \(\eta^2 = .09\). Planned pairwise comparisons conducted at the .05 level indicated that the mean conditional probability for the congruent–incongruent recall sequence was higher than either of the other two conditional probabilities, smaller of the \(F\)'s \((1, 179) = 12.39, p = .0007\), and that the conditional probabilities for the other two recall sequences did not differ, \(F(1, 179) = .01, p = .95\). The results of the analyses also indicated that the recall sequence effect did not significantly interact with trait dimension, \(F(4, 175) = 1.50, p = .21, \text{MSE} = .13\), partial \(\eta^2 = .03\), suggesting that the pattern of conditional probabilities was similar across the three trait dimensions used in the experiment.

The main research question of this article was whether the pattern described in the previous paragraph would hold, even when the data were adjusted for chance sequencing. The means for one of the chance-adjusted measures that we used, the CPDIFF measure, are presented in Table 1. Two aspects of those means are immediately apparent. First, the values of all of the means are close to 0, which is the value that represents the occurrence of a sequence at the frequency expected by chance. Hence, these means suggest that the observed frequencies of all of the recall sequences were close to those expected if the items were simply recalled in a random order. Furthermore, as confirmed by the results of the regression analyses, \(F(2, 179) = 2.18, p = .12, \text{MSE} = .12\), partial \(\eta^2 = .02\), there are only small, and nonsignificant, differences in the CPDIFF scores observed for the different sequences. Planned pairwise comparisons conducted among the means at the .05 level confirmed that there were no differences that

---

2 As noted in Tabachnick and Fidell (1996, p. 53) the strength of association index partial \(\eta^2\) is calculated by dividing \(SS_{\text{effect}}\) by \(SS_{\text{effect}} + SS_{\text{error}}\). Because the denominator of this index contains only variance attributable to the effect of interest and error, they argued that this index may be preferable to \(\eta^2\), which is calculated by using \(SS_{\text{total}}\) in the denominator. However, practically speaking, the use of \(\eta^2\) or partial \(\eta^2\) made relatively little difference in this study. The association indexes calculated from the sums of squares were similar, regardless of the index used.
were statistically significant, largest $F(1, 179) = 3.19, p = .08$. Hence, the heightened frequency for the consistent–inconsistent recall sequence that was reflected in the conditional probability analyses disappears when the probabilities are corrected for chance.

This pattern was consistent across the three trait dimensions used in the experiment: The Recall Sequence × Trait Dimension interaction was also not significant, but just barely, $F(4, 175) = 2.34, p = .06, MSE = .11$, partial $\eta^2 = .05$. Examination of the means for this interaction trend indicated that it was theoretically unimportant. It was due primarily to a very low CPDIFF value for the consistent–irrelevant recall sequence for the intelligence dimension ($M = -.17$). The means for all the other conditions were near 0 (highest $M = .06$; lowest $M = -.07$).

The data from the CPDIFF measure suggest that the heightened conditional probability of the consistent–inconsistent recall sequence can be a theoretically meaningless artifact. This conclusion is supported by the data from the ARIS scores. The ARIS score means for the consistent–consistent, consistent–inconsistent, and consistent–irrelevant recall sequences are presented in Table 1. Two aspects of those means are immediately apparent. First, the value of all of the means is close to 0, the value that represents the occurrence of a sequence at chance frequencies. More importantly, as confirmed by the results of the regression analyses, $F(2, 179) = 1.22, p = .30, MSE = .53$, partial $\eta^2 = .01$, there are only small, and nonsignificant, ARIS differences among the sequences. Planned pairwise comparisons conducted among the means at the .05 level confirmed that there were no differences that were statistically significant, largest $F(1, 179) = 1.52, p = .23$. The Recall Sequence × Trait Dimension interaction was also not significant, $F(4, 175) = 1.68, p = .16, MSE = .52$, partial $\eta^2 = .04$.

Recall sequences beginning with an expectancy–inconsistent behavior. The pattern of conditional probabilities described in the literature for inconsistent–consistent, inconsistent–inconsistent, and inconsistent–irrelevant recall sequences is mixed. Some studies find the highest conditional probabilities for the inconsistent–consistent recall sequence, some find the highest conditional probabilities for the inconsistent–inconsistent recall sequence, and still others find no difference among the sequences (e.g., see Srull, 1981; Srull et al., 1985). The associative linkage incongruity model suggests that the conditional probabilities for the inconsistent–inconsistent and the consistent–consistent recall sequences ought to be about equal (see Wyer & Srull, 1989, p. 185).

The conditional probability means for these recall sequences, presented in Table 1, are consistent with the predictions of the associative linkage incongruity model. The regression analyses did not yield a significant recall sequence effect, $F(2, 191) = 1.64, p = .20, MSE = .09$, partial $\eta^2 = .02$. A planned pairwise comparison conducted at the .05 level confirmed that the mean conditional probabilities for the theoretically important inconsistent–inconsistent and inconsistent–consistent recall sequences did not differ, $F(1, 191) = 3.14, p = .08$. The nonsignificant Recall Sequence × Trait Dimension interaction, $F(4, 187) = 1.67, p = .16, MSE = .09$, partial $\eta^2 = .03$, indicates that this pattern was consistent across the three trait dimensions used in the experiment.

Adjusting the conditional probabilities by calculating the CPDIFF measure also yielded nonsignificant results. The mean CPDIFF scores for the recall sequences beginning with expectancy–inconsistent items are presented in Table 1. Two aspects of those means are immediately apparent. First, as with the CPDIFF scores for sequences beginning with an expectancy–consistent item, the values of all of the means for recall sequences beginning with expectancy–inconsistent items were close to 0. This suggests that these sequences were all occurring with a frequency near to that expected by chance. Second, the recall sequence effect testing differences among the recall sequence CPDIFF means was not significant, $F(2, 191) = 2.75, p = .07, MSE = .08$, partial $\eta^2 = .03$. A planned pairwise comparison conducted at the .05 level confirms that the mean CPDIFF scores for the theoretically im-

<table>
<thead>
<tr>
<th>Recall Sequence</th>
<th>Conditional Probability</th>
<th>CPDIFF</th>
<th>ARIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>c-c</td>
<td>.24</td>
<td>-.01</td>
<td>.05</td>
</tr>
<tr>
<td>c-i</td>
<td>.42</td>
<td>.03</td>
<td>.08</td>
</tr>
<tr>
<td>c-x</td>
<td>.23</td>
<td>-.06</td>
<td>.05</td>
</tr>
<tr>
<td>i-c</td>
<td>.28</td>
<td>-.05</td>
<td>.01</td>
</tr>
<tr>
<td>i-i</td>
<td>.23</td>
<td>-.07</td>
<td>.11</td>
</tr>
<tr>
<td>i-x</td>
<td>.32</td>
<td>.02</td>
<td>.12</td>
</tr>
<tr>
<td>x-c</td>
<td>.32</td>
<td>-.02</td>
<td>.04</td>
</tr>
<tr>
<td>x-i</td>
<td>.36</td>
<td>-.04</td>
<td>.00</td>
</tr>
<tr>
<td>x-x</td>
<td>.22</td>
<td>.01</td>
<td>.08</td>
</tr>
</tbody>
</table>

Note. CPDIFF = Conditional Probability Difference score; ARIS = Adjusted Ratio for Individual Sequences; c = expectancy–consistent; i = expectancy–inconsistent; x = expectancy–irrelevant.
portant inconsistent–inconsistent and inconsistent–consistent recall sequences did not significantly differ, \( F(1, 191) = 0.23, p = .65 \). Hence, correcting for chance, the sequences were occurring equally often. The Recall Sequence \( \times \) Trait Dimension interaction was also not significant, \( F(4, 187) = 0.95, p = .44, MSE = .08 \), partial \( \eta^2 = .02 \), indicating that the pattern of CPDIFF scores was consistent across the three trait dimensions used in the experiment.

Similar results were obtained for the ARIS index (see Table 1 for means). The results of the regression analyses indicate that the recall sequence effect was not significant, \( F(2, 191) = 2.61, p = .08, MSE = .47 \), partial \( \eta^2 = .03 \). A planned pairwise comparison conducted at the .05 level confirmed that the mean ARIS scores for the theoretically important inconsistent–inconsistent and inconsistent–consistent recall sequences did not significantly differ, \( F(1, 191) = 1.03, p = .31 \). The Recall Sequence \( \times \) Trait Dimension interaction was also not significant, \( F(4, 187) = 0.86, p = .49, MSE = .47 \), partial \( \eta^2 = .02 \), indicating that the pattern of ARIS scores was consistent across the three trait dimensions used in the experiment.

Recall sequences beginning with an expectancy–irrelevant behavior. The associative linkage model of the incongruity effect (see Wyer & Srull, 1989, p. 185) suggests that the conditional probability for the irrelevant–consistent recall sequence ought to be higher than the conditional probability for the irrelevant–inconsistent and irrelevant–irrelevant recall sequences, an outcome obtained by Srull (1981). The conditional probability means for these recall sequences, presented in Table 1, are not consistent with the predictions of the associative linkage model: The conditional probability for the irrelevant–consistent recall sequence is only the second highest of the three. Furthermore, the regression analyses did not yield a significant recall sequence effect, \( F(2, 173) = 1.80, p = .17, MSE = .14 \), partial \( \eta^2 = .02 \). In addition, planned pairwise comparisons conducted at the .05 level indicated that the conditional probability for the irrelevant–congruent recall sequence was not significantly higher than the conditional probability of either of the other two recall sequences, larger of the two \( F(1, 173) = 2.93, p = .08 \). Finally, the non-significant Recall Sequence \( \times \) Trait Dimension interaction, \( F(4, 169) = 1.10, p = .36, MSE = .14 \), partial \( \eta^2 = .03 \) indicates that there were no substantial differences in the patterns of conditional probabilities across the three trait dimensions used in the experiment.

Adjusting the conditional probabilities by calculating the CPDIFF measure also yielded nonsignificant results. The mean CPDIFF scores for the recall sequences beginning with expectancy–inconsistent items are presented in Table 1. As with the CPDIFF scores reported earlier, the value of all of the means for recall sequences beginning with expectancy–irrelevant items were close to 0. This suggests that these sequences were all occurring with a frequency near to that expected by chance. In addition, the recall sequence effect was not significant, \( F(2, 173) = .73, p = .49, MSE = .13, \) partial \( \eta^2 = .01 \). Planned pairwise comparisons conducted at the .05 level indicated that the CPDIFF score for the irrelevant–congruent recall sequence was not significantly higher than the CPDIFF scores for either of the other two recall sequences, larger of the two \( F(1, 173) = .38, p = .56 \). Finally, the Recall Sequence \( \times \) Trait Dimension interaction was also not significant, \( F(4, 169) = .61, p = .66, MSE = .13 \), partial \( \eta^2 = .01 \), indicating that the pattern of CPDIFF scores was consistent across the three trait dimensions used in the experiment.

Similar results were obtained for the ARIS index (see Table 1 for means). The results of the regression analyses indicate that the differences among the recall sequence ARIS means were not significant, \( F(2, 173) = .56, p = .58, MSE = .49 \), partial \( \eta^2 = .01 \). Planned pairwise comparisons conducted at the .05 level indicated that the ARIS score for the irrelevant–congruent recall sequence was not significantly higher than the ARIS scores of either of the other two recall sequences, larger of the two \( F(1, 173) = .16, p = .67 \). The Recall Sequence \( \times \) Trait Dimension interaction was also not significant, \( F(4, 169) = .79, p = .54, MSE = .49 \), partial \( \eta^2 = .02 \), indicating that the pattern of ARIS scores was consistent across the three trait dimensions used in the experiment.

First-Order and Residualized Correlations Among the Probability Measures

The arguments presented by Skowronski and Welbourne (1997) suggested that the ARIS and CPDIFF corrections to the conditional probability data should reduce or eliminate the differences in conditional probabilities associated with some of the recall orderings if those differences are due to chance. This is exactly the pattern that was observed in the analyses presented previously. However, the Skowronski and Welbourne arguments also imply that these corrections should not substantially alter the orderings in those data across participants. That is, participants who have a high raw conditional probability score for a given recall sequence relative to other participants should also have high ARIS and CPDIFF scores for that sequence. Hence, there should be substantial correlations among a participant’s scores on these three measures.

We conducted two sets of analyses to examine these correlations. One set of analyses assessed all possible zero-order correlations among the three measures. This was done separately for each of the nine recall sequences that were described in the aforementioned analyses. The data clearly reflect substantial correlations among the three measures. Of the 27 correlations (among three measures for each of the nine recall sequences), 4 were above .95, 20 were between .85 and .949, and 3 were between .75 and .849 (largest \( r(75) = .99 \); smallest \( r(103) = .77 \), all \( p < .0001 \)). A second set of analyses examined these correlations after the effects of the
and .949, and 3 were between .75 and .849 (largest of the residuals for each index that were output from our regression analysis). The recall sequence had been accounted for (i.e., we correlated variables entered into the regression model (participants and sequence). The high correlations in the zero-order data were maintained in the residualized data. Of the 27 correlations among the residuals, 4 were above .95, 20 were between .85 and .949, and 3 were between .75 and .849 (largest \( r_{75} = .98 \); smallest \( r_{103} = .81 \), all \( p < .0001 \)).

The results of these correlation analyses are clear. Although the CPDIFF and ARIS indexes affect the extent to which the various recall sequences differ from each other, they do not have much of an effect on the extent to which a participant differs from other participants on a given recall sequence.

**DISCUSSION**

The data from this experiment lead to several straightforward conclusions. The first of these conclusions is that the conditional probability measure is a flawed measure of associative strength. Skowronski and Welbourne (1997) argued that the conditional probability measure can be distorted by several factors unrelated to associative strength, such as the length of the recall list or the content of the recall list. The data from this experiment provide an example of this distortion: The heightened conditional probability associated with the consistent–inconsistent recall sequence was eliminated when the data were corrected for chance. This outcome suggests that the heightened conditional probability observed for this recall sequence in our experiment was an artifact: a simple consequence of the incongruity effect in recall.

A second conclusion that can be derived from the results of this experiment is that the CPDIFF and ARIS indexes both can be effective in eliminating chance from assessments of recall sequencing. However, this does not imply that the results obtained by using these indexes will always be the same. For example, examination of the data from this experiment suggests that statistical tests involving the CPDIFF index might be slightly more powerful than tests involving the ARIS index, largely because the variance associated with the CPDIFF index tended to be lower than the variance associated with the ARIS index. More research is needed to see if this effect is generally obtained. Other differences between the indexes may emerge. For example, Skowronski and Welbourne (1997) argued that the CPDIFF score should be sensitive to the length of the recall list, whereas the ARIS score ought not to evince such sensitivity. Practically speaking, this means that the CPDIFF score can be biased by the results of a few participants with long recall lists. However, despite such possible differences, both the CPDIFF and ARIS led to the same conclusions in this experiment: (a) Adjusting for chance expectations, there were no significant differences in the frequencies with which participants exhibited various recall sequences; (b) the recall sequences produced by our participants were occurring at about the level that would be expected if those participants were recalling the items in random orders.

It should be noted that although this article focuses on the analysis of recall sequences, the flaws in the conditional probability measure that are described by Skowronski and Welbourne (1997) and that are documented in this article have implications that go well beyond the study of person memory. Virtually any study designed to analyze associative relations among items in a sequence are potentially affected, particularly when there are a number of different kinds of items that may repetitively occur in the sequence. For example, investigators who study how an individual behaves in a job setting might be interested in the extent to which certain on-the-job behaviors might prompt other behaviors. Researchers who study real-world social interaction sequences might be interested in the extent to which certain behaviors in one partner prompt the production of behaviors in the other partner (and vice-versa). Conditional probabilities would seem to be ideal for the investigation of such problems. However, the data presented in this article suggest that such raw conditional probabilities might be misleading, providing incorrect information about the extent to which one behavior prompts another. The CPDIFF and ARIS measures can help to avoid this problem.

The third conclusion that can be derived from the data that is reported in this article is relevant to current theory in social cognition. More specifically, our data suggest that the associative linkage explanation that has long been used to explain incongruity effects in recall may be incomplete at best. In the early studies (e.g., Srull, 1981), substantial support for that model came from conditional probabilities. The data in this article, along with the analysis provided by Skowronski and Welbourne (1997) and the data provided by Skowronski et al. (1998), suggest that the raw conditional probability data may have been misleading in those early conditional probability studies, leading to untoward confidence in the associative linkage explanation for incongruity effects in recall.

The possibility that the associative linkage explanation for incongruity effects in recall is incomplete is provocative, and for several reasons, we urge caution. One reason for caution is that there are several experimental tests of the associative linkage model that do not rely on conditional probabilities, but instead use response times or production times; these experiments have provided results that are consistent with the associative linkage model (e.g., Sherman & Hamilton, 1994; Srull et al., 1985). Furthermore, although correcting for chance eliminated the conditional probability advantage enjoyed by consistent–inconsistent recall sequences in our own data, it is possible that such corrections might not eliminate that advantage in the data obtained in other studies. For example, the analyses of Skowronski and Welbourne (1997) sug-
gest that some of the data in the Srull (1981) experiments would likely show evidence of the heightened occurrence of consistent–inconsistent recall sequences, even when chance factors are taken into account. A similar conclusion is obtained when the conditional probability data reported by Srull et al. (1985, Experiment 4) are examined.

Given these data, we cannot claim that associative linkages between expectancy–consistent and expectancy–inconsistent items do not form. However, we note that regardless of the outcomes obtained by these other studies, the data in this article suggest that these associative linkages may not provide the only mechanism, and may not even provide the primary mechanism, for incongruity effects in recall. That is, the data presented in this article showed a robust incongruity effect, despite an absence of evidence in the chance-corrected indexes for enhanced inter-item linkages. When their conditional probabilities are corrected for chance, similar data have been obtained by Asuncion and Lam (1995) and Driscoll (1992, Experiment 2).

The fact that incongruity effects can occur in recall without evidence of enhanced consistent–inconsistent associative linkages points to another mechanism for the incongruity effects. At least three specific candidates for this mechanism exist in the literature. One candidate, suggested by Hastie (1980), is explanatory or attributional processing of expectancy–inconsistent items. Such explanatory or attributional processing would cause items to be coded and stored at a deeper level, or perhaps with a more distinctive (and hence, more retrievable) memory trace. A second, somewhat similar explanation comes from research by von Hippel, Jonides, Hilton, and Narayan (1993). These researchers suggested that people attend less closely to expectancy–consistent than to expectancy–inconsistent information. The reason for this difference is that the schema used to encode expectancy–consistent information confers an advantage, allowing perceivers to attend only to certain elements of the expectancy–consistent behavior descriptions. In comparison, no such advantage is conferred on expectancy–inconsistent information. Hence, the mental representation of expectancy–inconsistent events should be stronger or more distinctive in memory than the mental representation of expectancy–consistent events. Finally, Babey, Queller, and Klein (1998) suggested a “summary-plus-exception” model in which trait-inconsistent memories are not linked to other specific expectancy–consistent behaviors but to a summary representation of the trait-consistent knowledge. Retrieval of the summary representation also causes retrieval of the expectancy–inconsistent behaviors, heightening their later probability of recall. Note that all of these mechanisms predict the same outcome: an incongruity effect in recall. They differ only in the reason underlying the development of the strong or distinctive memory trace thought to be associated with expectancy–inconsistent information.

Finally, the data presented in this article call for renewed attention to the area of person memory. If (and we emphasize the word if) our data seriously challenge the associative linkage model of person memory, then new models of person representation must be developed and tested. Furthermore, there have always been problems in the area that have remained unsettled. For example, as noted in the review by Stangor and McMillan (1992), one of the findings that differentiates work in stereotyping from work on trait expectancies is that the stereotyping work often yields a congruity effect in recall instead of the incongruity effect obtained in trait expectancy research. Although several theories might be applicable to the problem (e.g., Brewer’s, 1988, dual process theory) this discrepancy remains curiously unexplored, not to mention unexplained (but see Driscoll & Gingrich, 1997). The apparent success of the associative linkage model of person memory may have caused researchers in person memory to move on to greener pastures, ignoring or downplaying such problems. The data in this article suggest that they ought to return: The area may be more unsettled than anyone may previously have imagined.

ACKNOWLEDGMENTS

This article was supported by National Institutes of Health Grant MH–50730.

We thank Bryce Dewispelare, Krista Collins, and James Brumagen for helping with the data collection; Cynthia Rose for her coding of the recall data; Marilyn Brewer and the Social Cognition Group at The Ohio State University; and Bernadette Park, Jim Sherman, and the many reviewers of previous drafts of this article for their helpful comments.

REFERENCES


