BRIEF REPORTS

The conjunction fallacy: A test of averaging hypotheses

EDMUND FANTINO, JAMES KULIK, STEPHANIE STOLARZ-FANTINO, and WILLIAM WRIGHT
University of California, San Diego, La Jolla, California

The conjunction fallacy, in which individuals report that the conjunction of two events is more rather than less likely to occur than one of the events alone, is a robust phenomenon. We assessed the possibility that an analysis in terms of functional measurement methodology might be consistent with occurrence of the fallacy. A 3×3 design in which we varied the judged likelihood of the two components constituting the conjunction permitted us to assess the possibility that subjects judge the likelihood of conjunctions by averaging the likelihood of their component parts. The results were consistent with this possibility, and this interpretation was supported by analysis of the results in terms of functional measurement methodology.

Continued interest in apparent anomalies of human judgment and decision making has not always resulted in definitive accounts of why they occur. The present report addresses one particularly intriguing such anomaly, the conjunction fallacy in probability judgment discussed at length by Tversky and Kahneman (1982, 1983) and in more recent treatments of decision making (e.g., Abelson, Leddo, & Gross, 1987; Birnbaum, Anderson, & Hynan, 1990; Fantino & Stolarz-Fantino, 1991; Gavanski & Roskos-Ewoldson, 1991; Gigerenzer, in press; Massaro, 1994; Stolarz-Fantino & Fantino, 1990; Stolarz-Fantino, Fantino, & Kulik, 1996; Wells, 1985; Yates & Carlson, 1986). Subjects evincing the conjunction fallacy report that the conjunction of two events is more rather than less likely to occur than one of the events alone. This phenomenon is illustrated by an example used in the present research (see Appendix); the example is patterned after those in Tversky and Kahneman (1982).

Logically, Ralph is at least as likely to play in a heavy metal band for a hobby as he is to play in a heavy metal band for a hobby *and* happen to be a building inspector. Yet “Ralph plays in a heavy metal band” is judged less likely by most subjects than the conjunction. The robustness of this effect is well documented in Tversky and Kahneman (1983) and by our own attempts to eliminate it. For example, Stolarz-Fantino and Fantino (1990) reasoned that if instructions provided discriminative stimuli

favoring logical responses, the conjunction effect would be eliminated. However, a greater proportion of University of California, San Diego (UCSD) students showed the conjunction fallacy under the instruction “Your judgments should be made in terms of their probability and not simply in terms of whatever intuitive appeal is generated by the description above” than under standard instructions. Tversky and Kahneman have suggested that subjects frequently use a representativeness heuristic in arriving at their judgments rather than reasoning in terms of probabilities; that is, subjects purportedly consider the degree to which the specific stimulus components appear similar to or are consistent with the characteristics indicated in the background description. Thus, “Ralph is a building inspector who plays in a heavy metal band for a hobby” is more representative of the description of Ralph’s personality than is the simple statement “Ralph plays in a heavy metal band for a hobby.”

The present study considers the possibility that subjects average the likelihood of the two components in arriving at a judgment of the likelihood of the conjunction. Averaging implies that the conjoint probability is equal to or lies between the component probabilities, whereas the normative multiplicative model implies that the conjoint probability is equal to or lower than the component probabilities. Tversky and Kahneman (1983) note that “an averaging process . . . may be responsible for some conjunction errors, particularly when the constituent probabilities are given in a numerical form” (p. 306). Such a possibility has at least two implications, which we assessed: (1) If the likelihood of a conjunction is treated as the average of the likelihood of its components, then we should be able to predict the rated likelihood of the former from the rated likelihood of the latter; (2) if the conjunction consists of components that are judged about equally likely,

This research was conducted in accordance with human subjects guidelines at this institution. Informed consent was obtained from the subjects. The research was supported by NIMH Grant MH-20752 and NSF Grant BNS 91-08719 to the University of California, San Diego. Correspondence concerning this article should be addressed to E. Fantino, Department of Psychology 0109, University of California, San Diego, La Jolla, CA 92093 (e-mail: efantino@ucsd.edu).

then incidence of the conjunction fallacy should be less than when the conjunction consists of components of divergent likelihoods (since, in the latter case, the average likelihood falls more clearly above the less likely of the two components). Indeed, scrutiny of the standard conjunction-fallacy cases used in previous work suggests that most have involved components of highly divergent likelihoods, for example, a highly likely job combined with an unlikely hobby, or vice versa. The case in which both components constituting the conjunction are unlikely is one of two cases that are particularly interesting. According to the averaging view, frequency of the fallacy should be comparable in the low-low and high-high cells (but lower than for cases in which the events are of divergent likelihoods).

Support for this view comes from a number of studies. Both Gavanski and Roskos-Ewoldson (1991) and Wells (1985), who tested high- and low-probability combinations, did in fact find that the incidence and the magnitude of the fallacy were greater when high and low components were involved relative to high-high and low-low components. Abelson et al. (1987) compared subjects' ratings of the conjoint probability of two explanations of scenario outcomes with the probabilities of the component explanations. They were able to describe their data with a geometric-mean model that, although procedurally a multiplicative model, behaves like an averaging model in that the resulting predictions nearly always fall between the individual likelihoods of the factors being integrated (Abelson et al., 1987). Furthermore, the tendency of the geometric-mean model to yield a prediction closer to the smaller of the two component ratings mimics the well-known property of differential weighting observed in averaging tasks, wherein the more unlikely or unusual and thus more salient components play a more important role in the overall judgment (Oden & Anderson, 1971).

Research by Anderson (1965, 1981) and his colleagues has consistently shown that person perception judgments conform to an information-averaging model. The primary purpose of the present studies was to more formally assess this view with Anderson's (1981) functional measurement methodology.

Our emphasis on Anderson's view stems from its wide applicability. It should be noted, however, that other quantitative models have been applied, with some success, to the conjunction fallacy. These include Massaro's fuzzy-logical model of perception (Massaro, 1994) and algebraic models of judgment, including those of Birnbaum et al. (1990), Einhorn (1985), Wyer (1976), and Yates and Carlson (1986). According to Massaro (1994), subjects interpret the conjunction task as one of "pattern recognition," in which feature overlap determines probability judgment. Wyer's (1976) model for probability judgments assumes that subjects' estimates of conjoint probabilities are intermediate to those expected on a purely normative (multiplicative) and an averaging approach. Birnbaum et al. (1990) uses the multiplicative rule to express conjoint

probabilities, but weights the components in the geometric average to determine the intuitive probability of the conjunction. These free parameters enable it to predict conjunction-rule violations. The strengths and limitations of these and other models were recently reviewed by Hertwig (1995). Our purpose is not to choose between the various models but to assess the adequacy of predictions based on functional measurement methodology. This methodology is particularly promising because of its general applicability in a wide range of judgment tasks. Is the methodology adequate to provide an account consistent with violations of the conjunction rule?

The present study represents an extension of this line of work on the conjunction fallacy to a 3×3 within-subjects design (i.e., high, medium, low likelihoods of each of the two events). The primary advantage of this design is that, by using the functional measurement methodology, it allows a more formal assessment of the averaging view than heretofore. While we are primarily interested in the applicability of a weighted averaging model (described in the Results section), we also evaluate, for comparative purposes, the usefulness of a simple averaging of the component probabilities and of Abelson et al.'s (1987) geometric-mean model.

METHOD

Subjects

Three hundred thirty-six UCSD students served in the experiment.

Stimulus Materials

In the present experiment, we developed the set of eight statements about Ralph (shown in the Appendix) after pretesting with a group of UCSD students. The statements included three related to Ralph's hobby, which were judged high, moderate, or low in representativeness or likelihood (gliders, scuba, heavy metal, respectively), three analogously related to Ralph's job (building inspector, ranger, kindergarten teacher), and two fillers. The nine possible conjunctions formed from the six critical items were arranged, one to a page, in booklet form. Although all of our subjects rated the likelihood of all nine conjunctions, they were asked not to go back to a page once it was completed, and there was no evidence that they ever did so. All subjects first rated the likelihood (0–100) of the eight simple statements described above. These likelihood ratings confirmed those of our pretesting, with analyses of variance (ANOVAs) indicating highly significant differences ($ps < .0001$) among the intended high-, medium-, and low-likelihood hobbies ($M = 64.46, 29.20$, and 11.53 , respectively) and occupations ($M = 64.08, 40.37$, and 21.51 , respectively). The subjects then rated the likelihood of the nine conjunctions. Each appeared equally often as the first to be rated in the subjects' booklets.

Procedure

All of the data were collected in UCSD classrooms. The subjects were asked by their instructors to participate voluntarily in the experiment. No extra credit or other rewards were available.

RESULTS AND DISCUSSION

The mean proportion of subjects committing the conjunction fallacy for each of the nine conjunctions (with likelihood of hobby listed first) was: high-high, .78; high-medium, .80; high-low, .81; medium-high, .81; medium-

medium, .75; medium-low, .75; low-high, .85; low-medium, .81; low-low, .75. A within-subjects ANOVA found a significant interaction [$F(4,1188) = 4.35, p < .002$] and no main effects. Inspection of the means indicates, first, that the incidence of the conjunction error is comparable to that reported in prior work in absolute terms; more interesting, the effect is somewhat more likely when the conjunctions' two components are most divergent in likelihood, that is, high-low (.85 and .81; $M = .83$) than when the conjunctions involve components of equivalent likelihoods, high-high, medium-medium, and low-low (.78, .75, and .75 along the diagonal, $M = .76$). These results provide indirect support for an averaging view.

Testing the Averaging Model

With what success did the average judged likelihood of the two components predict the judged likelihood of the conjunction itself? To address this issue, we first calculated, for each subject, the correlation across the nine conjunctions between the subjects' averaged judgments of the two components with their respective conjunction judgments. We then computed the average correlation (across subjects) and found that there was a highly significant relationship [$r(334) = .702, p < .001$]. Thus, on average, approximately 49% of the variance in subjects'

conjunctions could be accounted for by a model that simply averaged the separate component likelihoods that constituted a particular conjunction.

To more formally assess the averaging model as well as alternative models, we next applied functional measurement methodology to the present data (Anderson, 1981). On the basis of the assumption that the factors being conjoined were independent, it would seem objectively desirable for subjects' likelihood ratings of a conjunction to approximate the product of the component ratings. This would be required if subjects' subjective likelihood estimates conformed to the objective criteria specified by the mathematical laws of probability. In that case, the factor plots of a conjunctive combination of job and hobby information would appear as in Figure 1a, where the dotted line represents the expected pattern of response to the single-component hobby-only stimuli (H1, H2, and H3 represent low-, moderate-, and high-likelihood hobbies; J1, J2, and J3 are low-, moderate-, and high-likelihood jobs). Under that model, the conjunction effect would not be observed. On the other hand, if subjects' conjunction ratings approximated an additive function, the pattern of results would be expected to resemble those in Figure 1b, with observed parallelism between the conjunction and single-component factor plots. In that case, the exact height of the single-component plot would

Figure 1. (a-c) Job $\times$ hobby conjunction models (ideal data). H1, H2, H3 refer to increasingly likely jobs; the dashed curve is for hobby only, the solid curves for different combinations of hobby and job. (d) The obtained crossover interaction supports an averaging model. Subjects judge the likelihood of a possible job (curve parameter) and hobby (horizontal axis) combination under the influence of a person description. Crossover of the dashed curve (hobby only) is critical evidence in favor of an averaging model.

vary with the relative degree of likelihood of the levels of job, but the two would not be expected to interact. A third possibility is that the subjects' conjunction ratings approximate an averaging model, in which case the single-component plot would be expected to exhibit significantly more slope than the conjunctions (Figure 1c). This would result in an interaction between the conjunction and single-factor judgments, and would be statistically detectable as a significant increase in response to H1–J2 over H1-only and as a significant decrease in response to H3–J2 over H3-only.

To test the usefulness of the models in predicting conjunction effects in this experiment, we compared the models' predictions of response patterns with the obtained response patterns for the two-way conjunction and single-component likelihood ratings of the 281 subjects who committed the conjunction fallacy. The pattern of results is displayed in Figure 1d. From the perspective of an information-integration theoretic approach, these results are best explained in terms of a differential weighted averaging model (Oden & Anderson, 1971; Shanteau, 1975). According to that model, the observed rating response, R , would be interpreted as a weighted average of the individual component probabilities conforming to the following algebraic description:

$$R = C + \frac{\sum w_i s_i}{\sum w_i},$$

where s_i are the subjective likelihoods of the individual components of a conjunction judgment, w_i are the corresponding weights, and C is an additive constant characteristic of baseline response. From the perspective of information averaging, the weights are psychologically interpreted as indicators of importance. The observed nonparallelism of the conjunctive judgments is explained by the less likely components' being weighted more heavily.

Visual inspection of the graphic representation in Figure 1d indicates that, in combination with the high-likelihood hobby information H3 (builds radio-controlled gliders), the additional information that Ralph might have the medium-likelihood job J2 (park ranger) causes an overall reduction in the subjective likelihood response, exactly as predicted by the averaging model. At the opposite end of the hobby-likelihood spectrum, the pattern of observed results also supports the averaging hypothesis. In combination with the low-likelihood hobby information H1 (plays in a heavy metal band), the additional information that Ralph might have a medium-likelihood job leads to an increase in the overall likelihood estimate. Thus, the impact of the medium-likelihood job information has opposite effects, depending on the level of hobby likelihood with which it is conjoined. From a statistical perspective, the critical difference predicted between the two-way and single-component ratings can be assessed by the specific comparison of the appropriate means, as discussed above. Planned comparison of the conjunctive con-

dition H1–J2 (low-likelihood hobby by medium-likelihood job) with the condition H1-only (low-likelihood hobby) reveals a significant difference in the expected direction [$F(1,280) = 15.14, p < .001$]. At the opposite extreme of hobby likelihood, the comparison of the H3–J2 condition (high-likelihood hobby by medium-likelihood job) with H3-only (high-likelihood hobby) also reveals the predicted significant effect [$F(1,280) = 109.04, p < .001$].

This crossover interaction supports the weighted averaging model (see Anderson, 1981, p. 65, for a discussion). The averaging model thus provides a simple explanation for the observed phenomena. If the moderately likely possibility that the target person is a park ranger is averaged in with the highly likely possibility that the person spends his leisure time building radio-controlled gliders, the overall likelihood estimate will be lowered from what it would have been without the job information. Likewise, the addition of this moderately likely job information will raise the response to the very unlikely possibility that the person's hobby involves playing in a heavy metal band. The observed data thus agree quite well with an information-averaging hypothesis.

The Geometric-Mean Model

Although our experiment was not explicitly designed to test the explanatory capability of Abelson et al.'s (1987) geometric-mean model, a comparison with Anderson's (1981) weighted averaging model is informative. As noted earlier, the two models make very similar predictions. Nonetheless, we can test the relative suitability of these two models on the basis of the pattern of coefficients each requires to model our data. Both models provide coefficient terms which can vary with the levels of independent variable, allowing the models to adjust to observed factorial design data on a cell-by-cell basis. This powerful feature shared by both models must be constrained by substantive interpretation in order to prevent excessive modeling power from rendering the resulting explanations vacuous. Therefore, we will compare the two models on the basis of substantive interpretation of each model's coefficient matrix. In Anderson's model, the substantive interpretation is that the coefficients represent the importance of the individual levels of a factor for the overall judgment of likelihood of the conjunction. This interpretation, combined with the well-known effects of salience on social judgment (Taylor & Fiske, 1978), leads to the prediction that less likely components will be weighted more heavily. This results in an observed nonparallelism of the factor plots' fanning in toward the unlikely extreme. This prediction nicely fits our observed data. The averaging model explains the observed differential weighting pattern with the simple assumption that unusual or unlikely components play a more important role in the overall conjunction likelihood judgment.

The predictions from the geometric mean model of Abelson et al. (1987) result in a somewhat different situation. The data shown below are the result of solving for

k , Abelson et al.'s coefficient of exaggeration, on the basis of our own observed data, using the formula

$$k = C/(AB)^{\frac{1}{2}},$$

with C taken as the observed rating response in each conjunction pair AB , and with A and B taken from the marginal single-component ratings. In this model, the interpretation is that the coefficient k reflects the amount that the geometric mean is adjusted up or down on a case-by-case basis, depending on how coherently the components fit together. Thus, Abelson's k is a measure of what has elsewhere been referred to as configural effect. The resulting k matrix is the pattern of coefficients k required for Abelson et al.'s model to fit our data (again with hobby likelihood listed first): high-high, 1.014; high-medium, 0.958; high-low, 0.916; medium-high, 0.979; medium-medium, 0.950; medium-low, 0.884; low-high, 0.755; low-medium, 0.740; low-low, 0.568. It is obvious from this resulting pattern that the coefficients k systematically increase with levels of factor across both A (job) and B (hobby).

Abelson et al.'s (1987) substantive interpretation would indicate that as the levels of likelihood of the individual components decrease, those individual components become less compatible with all other components. This assumption is required in order for the model to fit our observed data and conform to its own proposed substantive interpretation. Specifically, the required assumption is that lower likelihood components systematically tend toward increasing configural conflict. Given that the components in our study were selected to be conceptually orthogonal, such an assumption is probably unwarranted. Anderson's (1981) weighted-averaging model requires an assumption that appears more reasonable to us, namely that unlikely information is relatively more important than likely information. Thus, we believe that our results are more consistent with a weighted averaging model. However, it should be stressed that the Abelson et al. model is a viable one for understanding the present data. Moreover, the other models noted earlier may also be adequate. In conclusion, the robustness of the conjunction fallacy is consistent with the possibility that subjects respond to the conjunctions to a significant degree by averaging the likelihood of their component parts, a possibility supported by our functional measurement analysis. In that people have been shown to follow (not necessarily consciously) what is essentially an averaging strategy for many other types of decisions (e.g., Anderson, 1981), we suggest that the conjunction fallacy, although anomalous from the standpoint of formal logic, may derive substantially from a process (i.e., averaging) that is anything but anomalous. An implication for future research is whether or not subjects' tendencies to average change with experience. Averaging may develop because something akin to averaging provides an adequate strategy for decision making in everyday behavior. Would experience with the task strengthen the tendency to average, or would it lead sub-

jects to adopt a more normative approach? More intensive study of individual subjects might provide an answer.

REFERENCES

- ABELSON, R. P., LEDDO, J., & GROSS, P. H. (1987). The strength of conjunctive explanations. *Personality & Social Psychology Bulletin*, **13**, 141-155.
- ANDERSON, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology*, **70**, 394-400.
- ANDERSON, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.
- BIRNBAUM, M. H., ANDERSON, C. J., & HYNAN, L. G. (1990). Theories of bias in probability judgment. In J. P. Caverni, J. M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 477-498). Amsterdam: Elsevier, North-Holland.
- BEHNHORN, H. J. (1985). *A model of the conjunction fallacy*. Chicago: University of Chicago, Center for Decision Research.
- FANTINO, E., & STOLARZ-FANTINO, S. (1991). Rational analysis and illogical inference. *Behavioral & Brain Sciences*, **14**, 494.
- GAVANSKI, I., & ROSKOS-EWOLDSON, D. R. (1991). Representativeness and conjoint probability. *Journal of Personality & Social Psychology*, **61**, 181-194.
- GIGERENZER, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, **103**, 592-596.
- HERTWIG, R. (1995). *Why Dr. Gould's homunculus doesn't think like Dr. Gould: The "conjunction fallacy" reconsidered*. Konstanz: Hartung-Gorre.
- MASSARO, D. W. (1994). A pattern recognition account of decision making. *Memory & Cognition*, **22**, 616-627.
- ODEN, G. C., & ANDERSON, N. H. (1971). Differential weighting in integration theory. *Journal of Experimental Psychology*, **89**, 152-161.
- SHANTEAU, J. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica*, **39**, 83-89.
- STOLARZ-FANTINO, S., & FANTINO, E. (1990). Cognition and behavior analysis: A review of Rachlin's *Judgment, decision and choice*. *Journal of the Experimental Analysis of Behavior*, **54**, 317-322.
- STOLARZ-FANTINO, S., FANTINO, E., & KULIK, J. (1996). The conjunction fallacy: Differential incidence as a function of descriptive frames and educational context. *Contemporary Educational Psychology*, **21**, 208-218.
- TAYLOR, S. E., & FISKE, S. T. (1978). Salience, attention, and attribution: Top of the head phenomena. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 11, pp. 249-288). New York: Academic Press.
- TVERSKY, A., & KAHNEMAN, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84-98). New York: Cambridge University Press.
- TVERSKY, A., & KAHNEMAN, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, **90**, 293-315.
- WELLS, G. L. (1985). The conjunction error and the representativeness heuristic. *Social Cognition*, **3**, 266-279.
- WYER, R. S., JR. (1976). An investigation of the relations among probability estimates. *Organizational Behavior & Human Performance*, **15**, 1-18.
- YATES, J. F., & CARLSON, B. W. (1986). Conjunction errors: Evidence for multiple judgment procedures, including "signed summation." *Organizational Behavior & Human Decision Processes*, **37**, 230-253.

APPENDIX

Example of the Statements Rated by Subjects

Ralph is 34 years old. He is intelligent though not especially creative, and his friends describe him as somewhat compulsive

and dull. In college, he did well in the physical sciences but was weak in the humanities and social sciences.

Please indicate the likelihood of each of the following statements about Ralph by entering a percentage on the line to the left of the statement—for example, “0” would be virtually impossible, and “100” virtually certain. You can think of the continuum of likelihood as looking like this:

0	100
virtually	virtually
impossible	certain

Since the statements are not mutually exclusive, the numbers (each from 1 to 100) need not sum to 100.

- _____ Ralph is a disc jockey.
- _____ Ralph builds radio-controlled gliders for a hobby.
- _____ Ralph is a building inspector.
- _____ Ralph collects stamps for a hobby.
- _____ Ralph plays in a heavy metal band for a hobby.
- _____ Ralph is a park ranger.
- _____ Ralph is a kindergarten teacher.
- _____ Ralph scuba dives for a hobby.
- _____ Ralph is a building inspector who plays in a heavy metal band for a hobby.

(Manuscript received January 5, 1996;
revision accepted for publication July 15, 1996).