Comments on papers by Hammond, Metzger, Wilson, and Pribram

These comments were made by WR in response to the following four papers on the theme of *Knowing*. The papers were read at the Second Banff Conference on Theoretical Psychology held in 1969.

Inductive knowing
Kenneth R. Hammond, University of Colorado,

The phenomenal-perceptual field as a central steering mechanism
Wolfgang Metzger, University of Münster,

Memory organization and question answering
Kellogg V. Wilson, University of Alberta,

Neurological notes on knowing
Karl H. Pribram, Stanford University

The papers and comments were subsequently published in *The Psychology of Knowing*, eds. Joseph R. Royce and Wm. W. Rozeboom, New York, Gordon & Breach, 1972.

The comments of WR now follow.

COMMENTS ON PROFESSOR HAMMOND'S PAPER

William W. Rozeboom

Ever since I first read Hursch, Hammond, and Hursch (1964) some years ago, the multiple-regression embodiment of Brunswik's lens model has seemed to me to be an outstanding example of the small but significant technical advances which in aggregate transform intuitive speculations into a hard science. And since I have no serious quarrel with anything Hammond has said here, I would like to take this opportunity to clarify some features of the model which its past literature has left unpleasantly obscure.

First, I had best briefly derive the basic lens-model equations. (I will assume some elementary knowledge of multiple regression and the covariance statistic as set forth, e.g., in Rozeboom, 1966, Chapter 4.) Let $X_1, ..., X_m, Y_1, Y_2$, (and $Y_3, ..., Y_n$ for the n-system case) be a set of variables jointly distributed over some population of events. Then variable Y_i ($i = 1, ..., n$) can be partitioned as a sum of three mutually orthogonal components

$$Y_i = \dot{Y}_i + \tilde{Y}_i + E_i,$$

where $\dot{Y}_i$ is the linear regression of Y_i upon $X_1, ..., X_m$, $\dot{Y}_i + \tilde{Y}_i$ is Y_i's curvilinear regression upon $X_1, ..., X_m$ (i.e., $\tilde{Y}_i$ is the curvilinear regression's residual after the linear regression is partialled out), and E_i is the residual of Y_i unaccounted for in any way by the X_k. It is then easily shown that the covariance between Y_1 and Y_2 (and similarly for any others of the Y_i) analyzes as

$$\text{Cov}(Y_1, Y_2) = \text{Cov}(\dot{Y}_1, \dot{Y}_2) + \text{Cov}(\tilde{Y}_1, \tilde{Y}_2) + \text{Cov}(E_1, E_2) \quad (1)$$

Now, for any two variables A and B, $\text{Cov}(A, B) = \sigma_A \sigma_B r_{AB}$; while if $\hat{A}$ is the linear, or curvilinear, regression of A upon a set of predictor variables, $\sigma_{\hat{A}}$ equals σ_A times the linear, or curvilinear, correlation of A with those predictors. Hence if the relation between focus variables Y_1 and Y_2 is mediated entirely by cue variables $X_1, ..., X_m$, while for simplicity and without loss of generality the variables are scaled to have unit variances, we have

$$\text{Cov}(E_1, E_2) = 0,$$

$$\text{Cov}(Y_1, Y_2) = r_{Y_1 Y_2},$$

$$\text{Cov}(\dot{Y}_1, \dot{Y}_2) = R_1 R_2 r_{\dot{Y}_1 \dot{Y}_2},$$

$$\text{Cov}(\tilde{Y}_1, \tilde{Y}_2) = \sigma_{\tilde{Y}_1} \sigma_{\tilde{Y}_2} r_{\tilde{Y}_1 \tilde{Y}_2} = r_{\tilde{Y}_1 \tilde{Y}_2} \sqrt{\eta_1^2 - R_1^2} \sqrt{\eta_2^2 - R_2^2},$$

where R_i and η_i are respectively the multiple linear and curvilinear correlations of Y_i with the X_k. Hence from (1),

$$r_{Y_1Y_2} = r_{\dot{Y}_1\dot{Y}_2} R_1R_2 + r_{\tilde{Y}_1\tilde{Y}_2} \sqrt{\eta_1^2 - R_1^2}\sqrt{\eta_2^2 - R_2^2}, \tag{2}$$

which is Hammond's second equation (p. 299) except for an improved analysis of the nonlinear residual.[1]

A second way to analyze the linear component of Cov (Y_1, Y_2) is to note that the variance of the difference between any two variables A and B is $\sigma^2_{A-B} = \sigma^2_A + \sigma^2_B - 2\,\mathrm{Cov}\,(A, B)$, whence Cov $(A, B) = (\sigma^2_A + \sigma^2_B - \sigma^2_{A-B})/2$. Hence with unit-variance scaling for Y_1 and Y_2 as before,

$$\mathrm{Cov}\,(\dot{Y}_1, \dot{Y}_2) = \tfrac{1}{2}(R_1^2 + R_2^2 - \sigma^2_{\dot{Y}_1-\dot{Y}_2}), \tag{3}$$

while

$$\Sigma d = {}_{\mathrm{def}} = \sigma^2_{\dot{Y}_1-\dot{Y}_2} = \sum_{R=1}^{m} (\beta_{1k} - \beta_{2k})(r_{1k} - r_{2k}) \tag*{(4)[2]}$$

in which β_{ik} is the β-coefficient for predictor X_k in Y_i's linear regression upon the cue variables, r_{ik} is the linear correlation between Y_i and X_k, and "Σd" is Hammond's abbreviation for the variance of the linear-regression difference. Substitution into (1) then yields

$$r_{Y_1Y_2} = \tfrac{1}{2}(R_1^2 + R_2^2 - \Sigma d) + r_{\tilde{Y}_1\tilde{Y}_2} \sqrt{\eta_1^2 - R_1^2}\sqrt{\eta_2^2 - R_2^2}, \tag{5}$$

which, apart from the improvement already noted in (2), is Hammond's first equation (p. 298). The Σd term in (5), however, tends to be misleading. It *seems* from (4) and (5) that in order for achievement correlation $r_{Y_1Y_2}$ to be maximal, Σd should be zero (since it is a variance it cannot be negative), which in turn requires that $\beta_{1k} = \beta_{2k}$ and $r_{1k} = r_{2k}$ for each cue X_k, i.e., that the cue-utilization coefficients exactly match the cues' ecological validities. But in fact, $\Sigma d = 0$ is optimal *only* when there is no error variance in the distal variable's total regression upon the cues. For with the parameters of Y_1's relation to the X_k held constant, $r_{Y_1Y_2}$ is maximal when $\dot{Y}_1$ and $\dot{Y}_2$ are positively collinear (i.e., when $r_{\dot{Y}_1\dot{Y}_2} = 1$) *and* all the variance in Y_2 not needed for an optimal $\tilde{Y}_2$ is invested in $\dot{Y}_2$, i.e. when $\sigma_{E_2} = 0$. But if $\sigma_{E_1} > 0$ under these optimal circumstances, Y_2's projection into linear cue space is longer than Y_1's (i.e., $\sigma_{\dot{Y}_2} > \sigma_{\dot{Y}_1}$), whence the difference-variable $\dot{Y}_1 - \dot{Y}_2$ necessarily has nonzero variance. That is, if $\sigma_{E_1} > 0$, Σd must be positive if the quantity $R_2^2 - \Sigma d$ in (5) is to be maximal.

Moreover, the more that the cues are redundant, the more a good match between $\dot{Y}_1$ and $\dot{Y}_2$ can tolerate large discrepancies between cue-utilization coefficients and ecological validities. To illustrate this by means of an extreme example, suppose that there are just two cue variables and that their correlation is unity. Then $\dot{Y}_1$ and $\dot{Y}_2$ are perfectly collinear regardless of what values the β-coefficients may have, including the case where X_2 has zero weight for Y_1 while X_1 has zero weight for Y_2. This point has considerable significance for triple systems in which two persons judge the same distal variable. For given considerable cue redundancy, the judges could reach close, accurate agreement in their judgments, yet differ markedly in their cue utilizations. This shows how, in real life, persons who have achieved consensus and mutual trust on certain public issues could nonetheless dissipate their accord in acrimonious dispute over the bases for their conclusions. Contrary to the spirit of Hammond's 2-person studies, perhaps, sometimes it doesn't pay to let the right hand know how the left hand is doing it.

Next, it is worth noting the lens model's formal scope. This is in no way limited to cognitive or even psychological systems, for the model applies to any two variables Y_1 and Y_2 whose relation is mediated by one or more variables X_k. In particular, it is *not* requisite that Y_2 be a perception or judgment about distal variable Y_1. Y_2 could just as well be, say, degree of pupillary dilation aroused by miniskirt brevity Y_1. Neither need the X_k be proximal variables in Brunswik's sense, namely, aspects of events at the organism/environment interface. In fact, as is true of Hammond's own work, the X_k can themselves be distal variables or central percepts thereof such that the S first judges the values of $X_1, \ldots, X_m$ (or, alternatively, $X_1, \ldots, X_m$ are his judgments of the distal cues) and from there tries to infer the value of an even-more-distal variable Y_1.

On the other hand, the lens model's capacity to analyze "inductive knowing" has severe limitations, for the only inference pattern it subsumes is the statistical enthymeme:

The value of X_1 on this occasion is —,
the value of X_2 on this occasion is —,
. . . .
the value of X_m on this occasion is —;

---

therefore, the value of Y on this occasion is probably —.

(This argument is enthymematic because it lacks a major premise supplying probabilities for Y given the values of the cue variables.) The lens model does *not* cover even inductive inference of population parameters from observed sample frequencies, much less confirmation of theories by tests of their observational consequences.

Another aspect of the lens model which is highly susceptible to misunderstanding is the multiplicity of cue mediation (Brunswik's "vicarious functioning"). Mathematically, the number of cue variables mediating between focus variables Y_1 and Y_2 can always be reduced to two, while more generally, an n-system can be parsed to have no more than n relevant cues. This is because the curvilinear regression of each Y_i is some exact function $\phi_i(X_1, \ldots, X_m)$ of the cues and is hence itself a cue; consequently, if $X' =_{\text{def}} \phi_i(X_1, \ldots, X_m)$ for $i = 1, \ldots, n$, the pair X_i' and X_j' of transformed cue variables suffices to mediate the relationship between focus variables Y_i and Y_j $(i, j = 1, \ldots, n)$. The maneuver I am describing here is a familiar one in multivariate analysis, where for linear transformations it is known as "rotation of axes." Basically, the point is that cue space (curvilinear as well as linear) can be spanned in any number of ways, and how we choose to span it for a given analysis is mathematically arbitrary albeit this very much affects the number of relevant cue variables. Consequently, with one important qualification, use of the lens model to study e.g. how judgment is affected by the number of relevant cues is a meaningless enterprise. The qualification is that some ways to span cue space may well have greater "psychological reality" than do others—e.g., $X_1, \ldots, X_m$ may correspond to S's direct perceptions in a way that rotated cues $X_1', \ldots, X_n'$ do not. (Thus when I simultaneously perceive the height and distance of an object, I do not also perceive e.g. its height-times-its-distance.) What differences in "psychological reality" may in fact exist among transformationally equivalent sets of cue variables is an exceedingly interesting research question which to date has been virtually untouched.[3] However, the theory of this must be *added to* the lens model, not sought within it, even though this issue could well profit from lens-modeled research on how the accuracy and ease of acquiring distal/central correlations vary as a function of the particular axes in cue space along which input information is distributed.

The point just made about the number of cue variables also holds for linear vs. nonlinear cue utilization. The extent to which the relation between cues and focus variables is linear rather than curvilinear is very much an artifact of how we choose to span cue space. For example, the rotation from $X_1, \ldots, X_m$ to $X_1', \ldots, X_n'$ described above guarantees that all cue/focus relations are linear (though of course it does not also insure that the X_k' themselves are related only linearly), while it is an old and much practiced tradition in sensory psychology to reduce nonlinearities in the system by scaling input intensities as decibels. Admittedly, some of the nonlinearly alternative scalings of a given cue variable are intuitively more "natural" than are others, but until we learn more about what underlies this intuition and how to assess it empirically, it is hard to know how seriously to take recent work on linear vs. nonlinear cue utilization (cf. Goldberg, 1968).

Finally, under what circumstances does the lens model yield interpretively *significant* parsings of multivariate data? This occurs, I propose, when and only when the particular parameterization chosen for the model's application to a given phenomenon corresponds to the latter's second-level sources of variation, i.e., when the parameters most directly reflect factors in the phenomenon's underlying mechanism. What I mean by this can best be clarified by a highly oversimplified example. Consider a single-system with one cue variable X and judgment variable Y; specifically, suppose that X is distance-in-inches between eyebrow and hairline in a series of life-sized facial photographs, that Y is the S's estimate of IQ for a person whose photograph he is shown, and that the experimental design restricts X to only three values, 1 inch, 2 inches, and 3 inches. Then the regression of Y upon X for S at any given moment can be described by three parameters, two alternative choices for which are

$$\text{Parameterization } A\text{: } M_{Y|X_i} = a_1 + a_2X_i + a_3X_i^2,$$

$$\text{Parameterization } B\text{: } M_{Y|X_i} = b_i \quad (i = 1, 2, 3),$$

where X_i is value i of X and $M_{Y|X_i}$ is the contingent mean of Y given X_i, i.e. the average IQ which S guesses for photographs with an i-inch forehead. Suppose also that S has previously been trained (by methods which need not concern us here though in practice this would be an important detail) to have the judgment function $M_{Y|X_i} = 80 + 20X_i$; i.e., for parameterizations A and B, respectively,

$$a_1 = 60,\ a_2 = 20,\ a_3 = 0;$$

$$b_1 = 80,\ b_2 = 100,\ b_3 = 120;$$

but that now, working *only* with photographs having one-inch foreheads, S is retrained to give the response $Y = 85$ to X_1-stimuli. Our touchstone question now is: How does this retraining on X_1 modify S's responding to stimuli with cue values X_2 and X_3? In terms of the B-parameterization, one S—call him "linear"—might have post-retraining response parameters of

$$\text{Linear } S\text{: } b_1 = 85, b_3 = 110, b_3 = 135,$$

whereas another S might have post-retraining parameters of

$$\text{Hullian } S\text{: } b_1 = 85, b_2 = 97, b_3 = 119.$$

I call the second S "Hullian" because I intend him to generalize more or less according to the Hull-Spence model under which a new response (here the judgment $Y = 85$) learned to stimuli with feature X_1 should transfer in some degree to other stimuli to the extent they have features similar to X_1, but that apart from primary stimulus generalization, reconditioning on one stimulus leaves responding to other stimuli basically unaltered. Consequently, parameterization B, which has no built-in connections between S's response tendencies to the various stimuli, is most appropriate for the Hullian case. In contrast, the linear S generalizes by a pattern best characterized by parameterization A; namely, his cue-utilization function tends to maintain an invariant linear form whose slope coefficient is the primary manifestation of S's learning experiences, in this case changing from $a_2 = 20$ to $a_2 = 25$.

In short, a phenomenon's parameterization should be chosen to reflect the nodes at which it is modulated by changes in background constancies, for this is when the parameters give inductive access to the phenomenon's underlying sources (cf. Rozeboom, 1961). Since Hammond's work with the lens model has until now emphasized linear parameters (as shown e.g. by his parameterizing curvilinearity only as a residual), it would be desirable to determine whether his S's really do tend to generalize linearly in these situations. And if they do, what then is the theory—so strongly at odds with traditional models of learning—which explains *how* Ss are able to profit from past experience in this way?

NOTES

[1] The difference lies in my having analyzed the part of Y_i linearly unaccounted for by the cue variables into Y_i's curvilinear-residual regression $\hat{Y}_i$ upon the cues plus its component E_i entirely unrelated to the latter, whereas Hammond does not separate these. The improvement is important for the model's application to study of nonlinear systems, for the original version confounds inefficiency of curvilinear cue utilization with the distal variable's intrinsic unpredictability.

[2] *Proof:* Let $D =_{\text{def}} \dot{Y}_1 - \dot{Y}_2$ and assume unit-variance, zero-mean scales for all the Y_i and X_k. Then $\dot{Y}_i = \sum_{k=1}^{m} \beta_{ik} X_k$ $(i = 1, 2)$, so $D = \sum_{k=1}^{m} (\beta_{1k} - \beta_{2k}) X_k$. Also, since any linear combination of the X_k has zero covariance with any component of Y_i orthogonal to the X_k, $\mathrm{Cov}(D, Y_i) = \mathrm{Cov}(D, \dot{Y}_i)$. Hence $\sigma_D^2 = \mathrm{Cov}(D, D) = \mathrm{Cov}(D, \dot{Y}_1 - \dot{Y}_2) = \mathrm{Cov}(D, Y_1 - Y_2) = \mathrm{Cov}\left[\sum_{k=1}^{m} (\beta_{1k} - \beta_{2k}) X_k, Y_1 - Y_2\right] = \sum_{k=1}^{m} (\beta_{1k} - \beta_{2k}) \mathrm{Cov}(X_k, Y_1 - Y_2) = \sum_{k=1}^{m} (\beta_{1k} - \beta_{2k})(r_{1k} - r_{2k})$. To complete the proof, note that σ_D is invariant under all arbitrary linear rescalings of the variables so long as Y_1 and Y_2 retain unit variance.

[3] To my knowledge, nearly all the extant research bearing on this lies in multidimensional psychophysical scaling, where non-euclidian distance metrics introduce anisotropies in perceptual space. (See Garner, 1970.)

COMMENTS ON PROFESSOR METZGER'S PAPER

William W. Rozeboom

There is so much I like about Professor Metzger's paper that I am loath to say anything critical about it. Yet it perpetuates a philosophic error which invites total disaster upon any theory of cognition which makes it. Insomuch as nothing significant in Metzger's contribution rests on this blemish, I can best show my respect for the former by attempting to free it from the latter.

Although the error to which I refer is prime contender for epistemology's Original Sin, it is certainly not original with Professor Metzger. In fact, it so thoroughly saturates the mother's milk of his intellectual heritage—the brilliant Germanic tradition of act-psychology—that he many never have had occasion to reflect that an alternative is conceivable. Consider, for example, the seminal views of Brentano and Köhler:

"Every mental phenomenon is characterized by ... the intentional (and also mental) inexistence of an object, and ... reference to a content, a direction upon an object (by which we are *not* to understand a reality in this case), or an imminent objectivity. Each one includes something as object *within itself*, although not always in the same way. In presentation something is presented, in judgment something is affirmed or denied, in love [something is] loved ... The hypothesis that a physical phenomenon like those which exist intentionally *in us* exists outside of the mind [is not logically self-contradictory]. It is only that when we compare one with the other, conflicts are revealed which show clearly that there is no actual existence corresponding to the intentional existence in this case ... We will make no mistake if we quite generally deny to physical phenomena any existence other than intentional existence." (Brentano, 1874, pp. 50, 55; italics added.)

"The Behaviorist tells us that observations of direct experience is a private affair of individuals, whereas in physics two physicists can make the same observation, for instance, on a galvanometer. I deny the truth of the latter statement ... If somebody observes a galvanometer, he observes something different from the galvanometer as a physical object. For the object of his observation is the result of certain organic processes, only the beginning of which is determined by the physical galvanometer itself. In a second person, the observed galvanometer is again only the final result of such processes, which now occur in the organism of this second person. By no means do the two people observe the same instrument then, although physically the processes in one and the other are started by the same physical object." (Köhler, 1929 p.20.)

With such illustrious precedents as these, it is scarcely surprising to find Metzger asserting that

> ... "behind the world of the immediately given, behind the world of percepts, the presumed reality of the naive realist, there exists another world that to the phenomenal world has the relation of the original to its image but in itself is metaphenomenal or transphenomenal. That means that by its very nature it evades every direct observation and is therefore excluded from scientific thinking by positivism" (Metzger, p. 252 above).

Despite Metzger's labeling of his position here as "strict critical realism," it is in fact an orthodox phenomenalism. (I would have liked to call it a "crypto-phenomenalism," but there is nothing at all crypto about it.) A *real* critical realist would hold that what we observe directly is (in general) not mental phenomena but objects in Metzger's transphenomenal world.

It might seem a bit arrogant of me to stigmatize phenomenalism as a pure-and-simple error when so many first-rate thinkers have held this view and my earlier arguments against it (p. 62 above) are so skimpy. So I shall merely point out that *if* one distinguishes a mental act's content from its object sufficiently well to see that what intends the object most directly is not the act's nominal subject (i.e., a person) but its content, *then* it becomes evident that phenomenalism is both gratuitous and strongly counterintuitive. The realistically natural view here—the only one which now makes any sense to me although clarity in this matter was for me no simple overnight attainment— is that when physicist *o* observes galvanometer *g*, this analyzes as *o*'s having a mental content m_g (in this instance a percept) such that m_g is about (represents, signifies, is *of*) object *g*, whatever the latter may ontologically be. But if the property of *having* m_g is misconstrued as an *experience of* m_g—and note that the familiar verb-form "to experience *x*" is treacherously ambiguous between "to have an experience of *x*" and "to have *x* in experience"— the result is a phenomenalism which sees m_g as the *object* of *o*'s mental act, behind which may (Metzger) or may not (Brentano) lurk a corresponding "real" but unobserved entity *g*.[1]

If I wished to amplify my objections to phenomenalism here, I would probe with such questions as why having a percept should require being aware of that percept when e.g. having a brain tumor in no way requires awareness of that tumor, and what in the analysis of '*o* perceives object *x*' should necessitate that *x* be something within *o*'s mind. But I am content just to note that if Professor Metzger can be tempted to try on a genuine critical realism for size, he will discover that nothing in his paper needs amendment beyond a few labels and phrasings. He will still want to recognize two realms of being, the outer physical vs. the inner experiential, but what was formerly called the "apparent environment" or "phenomenal world" is now seen as a configuration of representations, or meanings, which are generally *of* the outer physical world. Similarly, his "world of percepts" (cf. quotation above) remains just as before except for the assumption that a person perceives his percepts. Instead, the latter are the *means* by which one perceives something else.

NOTES

[1] Such a view is obviously going to have trouble separating intentional contents from objects. With evident reluctance to make much of it, Brentano construed the distinction as that of a proposition vs. the nominative term therein—e.g., that "if I make the judgement 'A centaur does not exist', then ... the object is a centaur [while] the content of the judgement is that a centaur does not exist" (Brentano, 1874, p. 71 f.)—so that the content "includes the [object] within itself, and likewise exists within the subject" (Brentano, 1874, p. 71). Köhler, on the other hand, ignores the content/object distinction altogether.

COMMENTS ON PROFESSOR WILSON'S PAPER

William W. Rozeboom

I am delighted by this opportunity to root around in Wilson's pea patch, for the formal properties of his memory model illustrate why information-processing, cybernetic, systems-theoretical or computer-oriented approaches to psychology—call these "automatistic" theories for short—are both my joy and my despair.

The brief history of automatistic psychology nicely demonstrates how a movement founded on naivete, bad metaphor, and word magic can nonetheless evolve into a powerful and legitimate force within its discipline. First came post-war advances in control-systems engineering (notably, signal transmission and computer theories) which, needing verbal labels for new technical concepts, expropriated commonsense cognition talk for this purpose. The resulting mechanistic marvel with its spray-on cognitive complexion was promptly embraced by psychonomically frustrated onlookers as the Lochinvar who could at last breach the mind's maidenhead to inner mysteries, and from this seduction was born automatistic psychology in the back alley of psychological science. Initially, automatistic theorizing was little more than a revelling in the licence to speak cognition words out loud once more, not knowing or caring whether this touched any substantive issue not already well-assimilated in other terms by the older behavioristic/associationistic traditions. By the late '50s, however, its awe-eyed panting after systems engineering to disclose the essence of human cognition was giving way to simulation programs built upon genuinely psychological if still ingenuously introspective hypotheses about problem-solving processes. Thereupon it found congenial companionship in the re-cognitization underway in most orthodox sectors of psychology, especially math models, concept formation and psycholinguistics, until today its concepts have become familiar throughout much of the psychological mainstream.

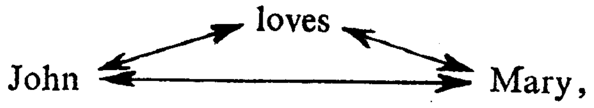
Automatistic theories have two major strengths. One is their emphasis on explanatory models that really work, i.e. which do in fact have the data implications ascribed to them, unlike so many past theoretic proposals especially in the S-R tradition. The other is their avid willingness to acknowledge the detailed complexity of inner events, both in diversity of process stages and nonlinearity of the functions by which one leads to another. (In retrospect, we can see how unbelievably empoverished—though in part deliberately so—orthodox behavioristic and associationistic theories have been in this regard.) But offsetting these virtues are two equally serious debilities. One is a strongly hypothetico-deductive[1] outlook which takes the main automatistic goal to be creation of computer programs (or programmable models) simulating commonsense human competences with indifference to whether organic systems work in at all the same way. Even worse—because it is more insidious—is that while the intended scope of automata theory includes all reactive systems, organic as well as artificial, its past development has been massively preoccupied with computer programming, thus restricting its repertoire of technical concepts largely to structures and functions practical for computer engineering. Consequently, if neural action has a fundamentally different organization from the unit-by-unit discrete serial activation schematized by flow diagrams—as we have good reason to suspect—it is a moot question whether the basic formal properties of higher organisms can be effectively captured by current styles of automatistic thinking. For automatistic theories to make the serious psychological contributions now within their grasp, they must learn how to conceive of system structure in terms dictated wholly by psychological considerations, unconstrained by the zeitgeist in computer-theoretic software. It is from the perspective of this latter point that I want to discuss automatistic models of memory.

The automatistic use of memory words, though often an outrage to this concept's cognitive core (cf. Rozeboom, 1965), nonetheless addresses an important general feature of adaptive systems likewise central to memory phenomena proper, namely, re-activation of processes in a system by stimuli which would be ineffective for this had not these processes or something like them been active in the system previously. More specifically, the matter at issue is "information storage and retrieval," for analysis of which we may usefully think of the organism's (system's) properties as being of two kinds, *states* and *process stages* (Rozeboom, 1965, p. 339ff.). Process stages are those conditions of the organism which vary as a function of input and hence share the latter's moment-to-moment instability, notably sensations, ideation, and behavior—i.e., psychological *activities*. In contrast, an organism's state properties—habits, preferences, traits, and other dispositional attributes—are stable though by no means unchanging characteristics which are relatively independent of the organisms's momentary process condition. The organism's moment-to-moment process activity is governed by process laws whose parameters are set by the organism's state properties, while the latter in turn are determined by state laws whose independent variables generally include certain features of the organism's process history. (Thus in traditional association theory, how likely it is that arousal of idea x reminds a person of idea y is given by the strength of his $x \rightarrow y$ association, where the latter is a state property determined in part by his past frequency of thinking x and y jointly.) Idealistically speaking, moreover, once a process m becomes activated in a system s, it often occurs that s thereby acquires a state property μ whose presence subsequently enables m to be activated in s by process antecedents ("recall cues") not previously capable of this. In automatistic jargon, such an m is an item of "information," formation and retention of μ is "storage" of m, and subsequent re-activation of m through μ's agency is its "retrieval."

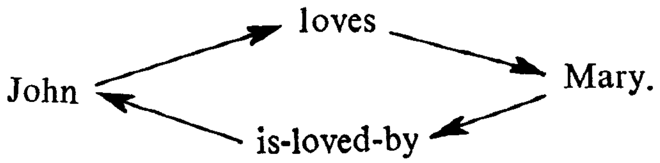
The problem most explicitly confronted by past theories of memory has been mechanisms for efficient storage and retrieval of information. Crucial to any such theory, however, is its implicit conception (scarcely ever examined critically) of what *logical kinds* of items are to be stored and retrieved. The Quillian-Wilson model makes an important advance in the latter respect, and I shall speak to this first.

Quillian (1967) and Wilson make clear that their model is specifically intended to handle *propositional* information, i.e. to store, retrieve, and make derivations from input in the form of declarative sentences. This would seem only natural for work on cognition were it not for the fact that virtually all past theories of psychological mechanism, traditional and automatistic alike, have treated process stages as unstructured aggregates of units lacking internal composition relevant to the system's function, so that a system's process condition at any given moment can be expressed by a simple list of *terms* naming which process elements are currently active. In contrast, processes which carry propositional information must be described by well-structured configurations of terms able to differentiate e.g. the process complex {*John loves Mary, John plays football*} from {*John loves John, Mary plays football*} even though the set of process elements is the same in both, namely {*football, John, loves, Mary, plays*}. Since Wilson does not detail how his model embodies and exploits this propositional structure at the process level (it should, for example, be able to extract {*a football player loves Mary*} from the first but not the second of the information complexes just mentioned), I cannot evaluate its success at this. From what I know of Quillian's version (wherein process structure is represented by a "tag" on each process element noting where in the state structure its activation came from), it should be possible to show that processing of propositions has important limitations in this model dues specifically to its psychonomically unnecessary flow-diagram construction (see below). Neither does the model make any provision for degrees of belief, much less for other dimensions of propositional attitude. But modern psychology has elsewhere recognized the propositional aspects of cognition in scarcely any way at all, and I cannot find it in my heart to fault a theory for not having attained mecca when it is struggling to get leg up on the highway thereto which most other pilgrims have never even thought to tread.

At first glance, information storage in the Quillian-Wilson model appears to be accomplished by a more-or-less orthodox associative structure whereby if the organism's state properties include an associative linkage from element *x* to element *y*, arousal of process *x* interacts with state property $x \rightarrow y$ to bring about activation of process *y*.[2] But Quillian-Wilson memory differs from true associationism in three fundamental respects. One is that the Q-W system does not form unmediated associations among all co-experienced process elements, but only those which reflect the grammatical structure of input sentences. Thus where classical association-theoretic principles imply that input of *John loves Mary* should produce the associative network



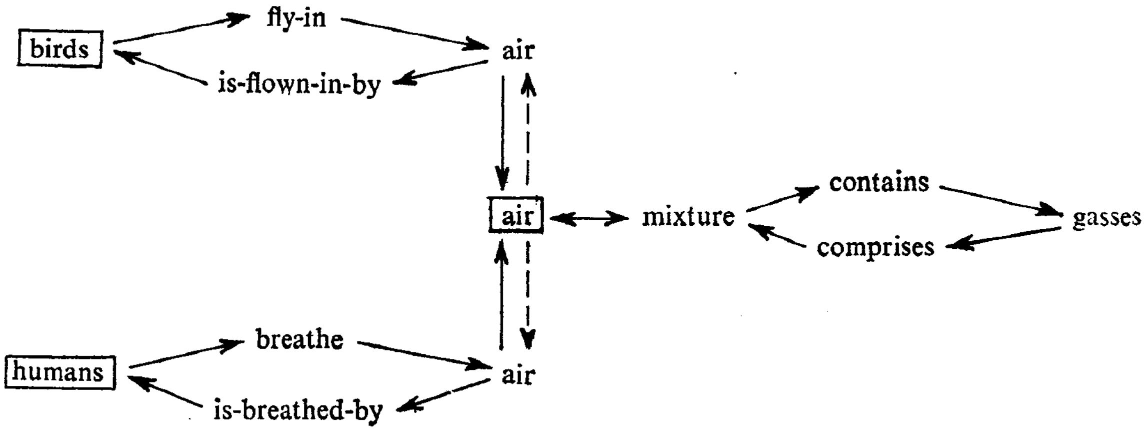
Wilson's version of the Q-W model converts this into the state structure



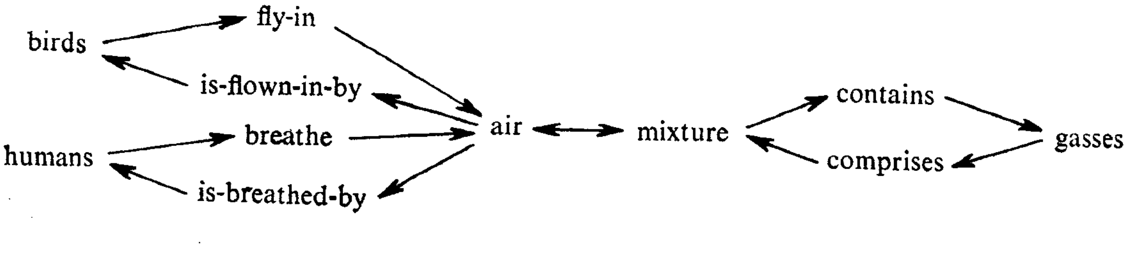
(Wilson does not say how his model manages to parse received sentences correctly, and to insert the verb's passive transformation, but it should not be difficult for an auxilliary input-processing routine to do this so long as the grammar of the input strings is carefully standardized. How such a routine differs from traditionally conjectured perceptual mechanisms, and in what respects humans might really work like this, is an instructive bit of analysis for another occasion.)

Secondly, the Q-W model contains no provision for generalization and graded arousal. Orthodox association theory draws heavily upon the principle that an association $x \rightarrow y$ will also interact with a process *z* to evoke *y* in strength which is an increasing function of *z*'s similarity to *x*. In contrast, activation in the Q-W model is all-or-none, and *z* can directly arouse *y* only if a z-node has been specifically linked to a y-node, regardless of how similar z may be to other processes linked directly to y.

The Q-W model's third critical departure from association-theoretic orthodoxy is that the elements coupled by activational linkages are not themselves process elements (or state surrogates thereof), but something else which might be called "containers" of process elements. Since this point touches upon the model's most basic structural properties, it is worth reproducing a portion of Wilson's Figure 1 (p. 368) augmented by additional information planes not made explicit there. For Wilson, the input information {*Air is a mixture of gasses, Birds fly in air, Humans breathe air*} is stored in a network something like Memory Structure *A*, in which the dotted arrows are between-plane connections which Wilson has added to Quillian's model. In this structure, *air* occurs in two token nodes and one type node; what "*air*" represents in the diagram is not itself joined to other process terms by association arrows, but is *carried by*, and can hence be common to more than one of, the entities ("nodes") which the arrows connect. In contrast, were process elements themselves to be the system's



Memory Structure A



Memory Structure B

memory nodes, as is true of orthodox association theory, the state diagram most like A would be Memory Structure B. (There are no boxed nodes in B because the type/token distinction here lacks significance.)

What functional differences are there between structures A and B, and why should Quillian and Wilson have proposed the first rather than the second? Regarding the latter, A is the legacy of computer-oriented thinking. Programming concepts are understandably geared to the practicalities of computer hardware; and to date the latter require routing of activity from one *place* to another in the system, while each separate item of information is stored at a different location which must be reached before this item can be acted upon. As for the difference between structures A and B, this depends very much on whether Quillian's or Wilson's version of A is at issue. For Quillian (1967), the class of between-plane connections comprises only one-directional links from token nodes to type nodes with the same content (e.g., from *air* in the top and bottom planes of A to $\boxed{\textit{air}}$ in the second), so that a plane can be entered only through its type node. Consequently, starting with activity in the top plane of A, Quillian can reach the information plane whose type node contains *air*, but cannot retrieve the non-typal information about *air* in the bottom plane. In contrast, because Wilson's between-plane links are bidirectional, any two planes tokening the same content x are mutually accessible through x's type node; hence activation of a given plane P permits retrieval of all information stored elsewhere about all process elements tokened in P.

More generally, any two nodes which are n pulses of activation apart in structure B are at most $n + 2$ pulses apart in Wilson's version of structure A. While this still leaves some minor differences in formal potential between B and Wilson's A, we have insufficient detail about the intended functioning of the latter to tell whether A would be appreciably superior to B for this purpose. In short, then, Wilson has labored to make the location-addressible memory structure presupposed by automatistic theories yield content-addressible memory function. But *psychological* theories of memory have always assumed content addressibility at the outset, without much hang-up over how this occurs in the organism. And if, as I am inclined to believe, the most pressing task for the psychology of memory is to learn more about the *functional* intricacies of recall (for only then will we know what our conjectured mechanisms are supposed to *do*), automatistic struggles to devise more efficient shuttle circuits for retrieving information scattered throughout a maze of locations are for psychology (*contra* computer theory) largely waste motion. I hasten to add, however, that the psychological relevance of such models would be greatly enhanced by careful comparative study of the formal differences among e.g. structure B and the two versions of A to lay bare what corresponding differences they entail for discernable memory phenomena which empirical psychology has not yet thought to research.

Another automatistic conceptual bias more likely to obscure than to illuminate the bases of organic behavior lies in the essentially *seriatim* character of computer operations. That is, computers still do only t things at a time, where t is seldom greater than unity. If Wilson's model concurs in this (unlike Quillian, he does not explicitly commit himself to it), then problems of selection arise whenever a type node is activated. For if the total memory structure contains n token nodes for process element x, each of which is linked bi-directionally with type node $\boxed{x}$, there are then $n + 1$ different exits from the latter. If only one of these exits can be followed at a time, is the choice made randomly or is there some logic of selection? Whichever exit is initially chosen, does activation immediately press onward from the new node thus reached, or are all exits from $\boxed{x}$ somehow scanned before action is propagated unconditionally; and if the latter, what determines the final choice? The technical points at issue here cannot be made clear without more detail about the model's intended functions and their manner of execution; but it is abstractly evident that if only a small fixed number of nodes can be activated at once, then the greater the average number of exits per node, the smaller should be the probability that the system will accomplish a given task within a specified period of time. For temporal efficiency, it should be possible for all exits from an activated node to be followed simultaneously, but I doubt that this is compatible with the Q-W model's projected routines for processing the information so activated.

Perhaps the best way to highlight the logical suppositions of automatistic views on storing and retrieving propositionally structured information is by contrast with the most natural psychonomic approach to this—"natural" in being an old intuition of classical psychology albeit one never well developed technically. This is the notion that activation of a process $R(x_1, \ldots, x_n)$, wherein elements $x_1, \ldots, x_n$ stand in relation R, strengthens a relatively permanent "memory trace" τ of $R(x_1, \ldots, x_n)$ given which the probability, intensity, and/or latency with which another process $S(y_1, \ldots, y_m)$ revives ("redintegrates") the structured complex $R(x_1, \ldots, x_n)$ is a function jointly of τ's strength and the extent to which $S(x_1, \ldots, x_m)$ resembles the process $R(y_1, \ldots, y_m)$ of which τ is the trace. (The detailed nature of this "resemblance" needs to be worked out by future research, but its primary determinants are presumably (a) the proportion of elements in $R(x_1, \ldots, x_n)$

and $S(y_1, ..., y_m)$ common to or, more weakly, similar in both, and (*b*) structural similarity whereby, e.g., $R(x_1, ..., x_n)$ is more similar to itself than it is to an elementwise identical process $R(x'_1, ..., x'_n)$ in which the x'_i are a permutation of the x_i.) At this level of the trace model's conception, an organism's memory state is characterized simply as a set $\{\tau_i\}$ of memory traces. Nothing is said about linkages or other relations among the τ_i because there is so far no work for between-trace connections to do (which of course in no way precludes later postulation of these if need arises). Moreover, nothing in the trace model's initial conception suggests that memory traces are differentially *accessible* to various recall cues. That is, the basic postulate concerning how an active process S interacts with a trace τ_i to revive the latter's process counterpart does not view this as dependent upon whatever additional traces are also present (though it is entirely open with respect to whether τ_i's strength is influenced by other traces). Hence in this first approximation to whatever more sophisticated version of the theory may eventually evolve, a given input S is conceived to operate upon all traces simultaneously, with a corresponding propensity to concurrent revival (in degrees respectively appropriate to the individual traces) of all processes from whose traces S can get any action. Finally, in light of this press to simultaneous arousal of indefinitely many processes, some principles of process concatenation are needed (e.g., formation of a composite by superimposition of constituents), the details of which again remain open for future research but wherein concepts of "competition," "summation," and others long exploited to this end in the verbal learning and behavior-theoretic literature may be expected to figure prominently.

I do not suggest any inherent incompatibility between trace theory and automatistic approaches to memory, for there is no reason why functional properties envisioned by the former cannot be reasonably well approximated by some ingeniously contrived computer-theoretic mechanism. My point is that those functions which are most basic in trace theory's initial conception are still alien to automatistic thinking and will undoubtedly remain so until automatistic models shed their conceptual dependency on computer programming or computers become designed around physical principles vastly different from their present "digital" construction. For now, the physical analogies most appropriate to trace theory are not switching circuits with all-or-none seriatim action and discrete channels of arousal established apart from the contents of their termini, but wave phenomena in which state structures are swept by a complex wavefield to which these resonate in degrees determined by the intensity of field components in or near the bands to which the resonators are tuned and whose joint emissions modulate the wavefield's character by cancellation and enhancement. To be sure, analogies are merely heuristic for scientific theory, and as memory research progresses we may well discover phenomena more readily modelled by switching circuits than by wave physics (or, more likely, not well modelled by either). But it would be unfortunate if the pre-packaged technical sophistication of contemporary computer programming were to occlude our access to those explanatory concepts which extrude most naturally from empirical work on memory phenomena.

NOTES

[1] See Rozeboom, 1970, pp. 90ff., for arguments against hypothetico-deductivism as a proper mode of scientific inference.

[2] More precisely, when x and y are process elements, what the terms "x" and "y" refer to in the associative concept "$x \rightarrow y$" are state surrogates of x and y in the way, e.g., the wiggles on a phonograph record are state surrogates of the acoustic processes they help activate. This distinction is a fine point which I will not try to keep verbally explicit here.

COMMENTS ON PROFESSOR PRIBRAM'S PAPER

William W. Rozeboom

There are few academic sports spectaculars quite so exhilarating as the sight of playmaster Pribram finger-tipping the ball in full sprint downfield. Yet if the game is not to degenerate into a shambles, someone must take responsibility for blowing the whistle on fouls.

Actually, my whistle chirps here will be rather timid, for while I have deep suspicions about much of the action in Pribram's performance, it all happens too fast for me to tell exactly what is going on. According to Pribram, the general sequence of cognitive events in an organism is for stimulus input to be first *coded* by the nervous system and then recoded into patterns of neural activity called *Images-of-Events*. Meanwhile, internal physio-chemical conditions give rise (via coding?) to *Monitor-Images* while images of a third kind, *Images-of-Achievement*, are representing *Actions* (i.e., external accomplishments). When these images-of-achievement interact with images-of-events on the one hand and with monitor-images on the other, *signs* and *symbols* respectively result. Finally, linguistic knowledge results when "man manipulates Symbols as Signs." All of this seems very profound—too much so, unfortunately, for me to understand very clearly. I do, however, find myself noting possible inconsistencies and wondering if Pribram has really addressed the definitive issues of cognition.

His theory of action, for example: I think I am safe in construing this to be very similar to Metzger's account (p. 244ff. above). Certainly Pribram's statement that "Images-of-Achievement guide movement ... by tuning the reflex" (p. 455), i.e. that these set the equilibrium points in homeostatic lower-level motor processes, well fits this conception. But then I am at a loss to interpret his claim that images-of-achievement "are composed of signals [from muscular force fields] initiated by forces external to the organism" (p. 455), for this seems to imply that the reflex tuning so brought about is determined blindly by the organism's recent history of muscle events rather than by superordinate control from his cognitively intended goals. Very likely a simple rephrasing or word of clarification would allay my doubts on this point (as the final draft of Pribram's paper has already done for certain other qualms I had originally raised here). Considerably more than that, however, seems necessary to make public the substantive insights which I trust underlie the pyrotechnic dazzle of Pribram's account of cognition's afferent stages:

Consider, for example, his concept of "coding". Does this have any *psychological* implications beyond recognizing the obvious fact that since neural propagation of input signals cannot literally copy physical events at the receptor surface, central sensory processes must be transformations of their input precursors? I grant that Pribram is working towards a specific theory concerning what aspects of CNS activity are correlated in what way with input patterns, but he hasn't suggested what import this may have for a psychology which abstracts the functional properties of cognition from its neurophysiological substratum.

Again, we are told that the first stage of neural coding passes over into images[-of-events] through "a further coding process by which the neural process can represent fully its origin," (p. 453). I am unsure whether this is meant to imply that the pre-Image stage of coded input does not represent its origin as fully as does the Image, or merely that the two coding stages both fully represent their origin. Either way, Pribram's claims about "representation" remain gratuitous at best (and beguiling at worse) until he clarifies what sort of representation is at issue here and faces up to the more important logical problems which remain for his account in this sense of the term. Does he really mean just that variable X "represents" variable Y when X-events are isomorphic to or statistically correlated with Y events? If so, then the Image can represent its origin no better than does the pre-Image stage of coding (since when the relation between variables X and Y is mediated entirely by variable(s) M, Y can be no more highly correlated with X than is M and will be less so if there is any error variance in the system); while by virtue of the reflexivity, transitivity, and (more roughly) symmetry of isomorphisms and correlations, the Image, pre-Image, and environmental origin all mutually represent one another as well as—most accurately of all—themselves. Surely Pribram intends "representation" to be more selective than this, so that an Image represents its external source *rather than* (instead of in addition to) itself or the pre-Image Coded input. Surely in an essay whose theme is the epistemic act of knowing and which purposefully makes free use of classical psychology's major cognitive concepts, the *of*-ness ascribed to Images-of-Events is intended to be the *cognitive* relation whereby an image Y represents an originating event (or between-event relation) X when Y is referentially *about* X. But then which among the events (or relations among events) in the causal sequence leading to Y is the one that Y represents, and by what analysis of aboutness can it be claimed that Y represents *that* particular X rather than some other one of its causal precursors? For example, if a photograph presents a viewer with retinal stimulation that arouses first-stage coded neural activity which in turn produces a recoded Image, is the originating event represented by this Image (1) the pre-image neural coding, (2) the retinal pattern, (3) the configuration of pigments on the photographic print, (4) something in the negative from which the print was made, or (5) the original scene to which this negative was first exposed? If Pribram elects (3) or (5), as I hope would be his preference, on what grounds can he argue that the viewer's Image represents the distally external event rather than its retinal or post-retinal consequence? Since he speaks of "resemblance" several times in this context, would he propose that the Image is literally more *like* (i.e., similar to) its distal origin than it is like mediating events at the sensory interface?

I am similarly uneasy about Pribram's treatment of "signs". We are told that these are produced by "decoding" or "indexing" images-of-events by much the same mechanism that produces images-of-achievement. Just how this occurs is not clear to me, for at one point (p. 456) the achievement-mechanism produces signs by modulating receptor action, which would control which images-of-events are formed in the first place rather than how the latter are subsequently Indexed; later, however, it is said that indexing "derive[s] when Images are Acted upon" (p. 459), while the "interdigitating" of images-of-events and images-of-achievements sounds more like an amalgam of these two image types than like a receptor bias on the first. But more important is what Indexing is conceived to accomplish. I interpret this to be a categorizing (*á la* Bruner) of images-of-events, that is, an abstractive identifying of their distinctive features. For this to be a genuine cognitive operation, however, the Image must have its identified attributes predicated of it in a propositionally structured process; whereas so far as I can make out, Pribram's Signs are simply reactions (central or otherwise) elicited by the Images so indexed. If so, his account of sign processes is nothing more than a neurophysiologically flavored paraphrase of traditional association-theoretic models (*á la* Staats and Kendler) which treat concept formation, abstraction, judgment, and other cognitive phenomena as convergent associations, i.e., as common labeling responses becoming attached to a variety of stimuli. I know Karl well enough by now to feel sure that he has something considerably more interesting than this in mind, but what that something-more may be remains at present a tantalizing mystery.

Pribram's use of the word *Symbol* to denote those "expressions of feeling" which derive from classifying Monitor-Images is strongly at odds with what most philosophers understand by this term, but I suppose that he is keying into the usage under which "symbols" (i.e., the Flag, Hamlet-seen-as-Everyman, firearms-seen-as-phallic, etc.) have an artsy-gutsy subjective/existential orientation *contra* the semantically pure external outlook of "signs." But are hormonal balances and the like then "the events [which symbols] symbolize" (p. 459)? If so, what is the nature of the relationship by which an indexed monitor-image is a symbol *of* a hormonal event? (Pribram emphasizes that it is not an isomorphism, but what then *is* it?)

Finally, to lessen the prospect of rotary agitation within Charles Peirce's grave, a *caveat* should be filed against the view that for Peirce, *abductive* reasoning is hypothesis formation by analogy (a claim which Pribram has now softened considerably since his original presentation but still not entirely abandoned). Peirce used the term "abduction" to describe whatever processes are responsible for a person's first thinking of a hypothesis prior to its subsequent confirmation or disconfirmation in one way or another (see Peirce, *Collected Papers* Vol. VI, p. 358). "Analogy" for him was a form of inference which *contrasted* with reasoning by hypothesis, while "abduction" was an aspect of the latter. In his own words,

> Argument is of three kinds: Deduction, Induction, and Abduction (usually called adopting a hypothesis). (*Collected Papers* Vol. II, p. 53.)

Peirce's concept of "argument" is broader than that of "inference," for it includes the acquiring of hypotheses in ways other than inference, namely, by abduction:

> Abduction must cover all the operations by which theories and conceptions are engendered. (*CP* V, p. 414)

For deriving conclusions from premises, on the other hand,

> non-deductive or ampliative inference is of three kinds: induction, hypothesis [whose premises may be given by abduction], and analogy. (*CP* VI, p. 31),

while

> analogy ... is a type of inference having all the strength of induction and more besides. (*CP* V, p. 411; the logical form of analogical argument is given in *CP* II, p. 310.)

Since Peirce treats analogy as distinct from though similar to induction, he should probably have included Analogy as a fourth kind of argument in the first quotation above.