Rational Models of Cognition

Edited by

Mike Oaksford
School of Psychology,
University of Wales, Cardiff

and

Nick Chater
Department of Psychology,
University of Warwick

Oxford · New York · Tokyo
OXFORD UNIVERSITY PRESS
1998

Contents

10 *The rational analysis of categorization and the ACT-R architecture*

John R. Anderson and Michael Matessa

As the cognitive revolution psychologists have striven to develop detailed theories of the information-processing steps involved in performing a particular task. The goal is to be able to say moment by moment what is happening in the minds of people as they do the task. While this goal is laudatory there are serious problems with how feasible it is. When one starts to deal with complex tasks (like doing high school mathematics—Anderson *et al.*, 1995) the sheer volume of the detail can be overwhelming. Even in simpler laboratory tasks, the data that one collects is often at a much coarser grain size than the processes one is postulating. There arise identifiability problems about whether very different ways of processing information might not result in the same data. Individual differences are another problem. When one gets to this level of detail it is unlikely that every subject is doing the task in the same way.

In Anderson (1990) I proposed rational analysis as a way of achieving an abstraction which eliminated much of this mechanistic detail. The fundamental idea was that cognitive functions were adapted to the structure of the environment and one might be able to make predictions about cognition that only required minimal assumptions about information processing. That is, one could formulate an optimization problem in which one tried to find the best behaviour given rather general assumptions about information-processing limitations but detailed assumptions about the structure of the environment. This was a potential improvement because one could directly observe the structure of the environment in a way one could not observe the structure of the mind. A number of other researchers have independently come to explore the promise of this rational analysis approach and I think this conference is a testimony to the productivity of the approach. However, the point of this chapter is that by ignoring the information-processing detail one sometimes loses important generalizations.

I will describe in this chapter a cognitive architecture called ACT-R (Anderson, 1993), which is a version of the ACT theory (Anderson, 1976, 1983). ACT-R was developed to incorporate the rational analyses given of memory and choice in Anderson (1990). However, it does not incorporate the rational analysis of

categorization given in Anderson (1990). I will show that within the ACT-R architecture the rationality of categorization behaviour is actually derivative of the rationality of memory and choice. This is the generalization that was missed by the rational analysis but which comes when one commits to the information-processing detail of a cognitive architecture. The remainder of this chapter will: (i) overview the ACT-R theory; (ii) review the rational analysis of memory and its mapping into the ACT-R theory; (iii) review the rational analysis of choice and its mapping into the ACT-R theory; and (iv) review the rational analysis of categorization and how its mapping into the ACT-R theory is derivative of the memory and choice implementations.

The ACT-R theory

The ACT-R theory admits three basic binary distinctions. First, there is a distinction between two types of knowledge—declarative knowledge of facts and procedural knowledge of how to do various cognitive tasks. Second, there is the distinction between the performance assumptions about how ACT-R deploys what it knows to solve a task and the learning assumptions about how it acquires new knowledge. Third, there is a distinction between the symbolic level in ACT-R that involves discrete knowledge structures and a subsymbolic level which involves neural-like activation-based processes that determine the availability of these symbolic structures. We will first describe here ACT-R at the symbolic level. A symbolic-level analysis of the knowledge structures in a domain corresponds basically to a task analysis of what needs to be learned in that domain. However, as we will see, the availability of these symbolic structures depends critically on the subsymbolic processes. It is in the design of these subsymbolic processes that we have borrowed extensively from the rational analysis of Anderson (1990). Therefore, we will describe these subsymbolic processes as a part of our description of the rational analysis of memory and choice.

Declarative knowledge reflects the factual information that a person knows and can report. According to ACT-R declarative knowledge is represented as a network of small units of primitive knowledge called chunks. Figure 10.1 is a graphical display of a chunk encoding the addition fact that $3 + 4 = 7$. A chunk just binds a number of elements (in this case *three, four, and seven*) into a configuration. The roles of the elements in the configuration are encoded by attributes (i.e. *addend1*, *addend2*, and *sum* in Fig. 10.1). In Fig. 10.1 are also quantities like B_i, S_{ji}, and W_j. These are subsymbolic quantities which will be explained when we come to the rational analysis.

Procedural knowledge, such as mathematical problem-solving skill, is represented by a large number of rule-like units called productions. Production rules are condition–action units which respond to various problem-solving conditions with specific cognitive actions. The steps of thought in a production system correspond to a sequence of such condition–action rules which execute one after another. Production rules in ACT-R test for the existence of specific goals in their conditions

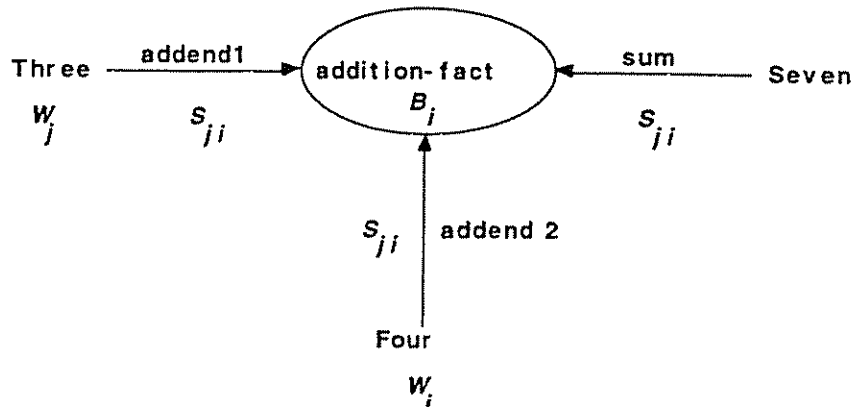

Fig. 10.1 A network representation of a declarative ACT-R knowledge unit

and often create subgoals in their actions. For instance, suppose a child was at the point illustrated below in the solution of a multicolumn addition problem:

$$\begin{array}{r} 534 \\ +248 \\ \hline 2 \end{array}$$

Focused on the tens column the following production rule might apply taken from the ACT-R simulation of multicolumn addition in Anderson (1993):

IF the goal is to add $n1$ and $n2$ in a column and $n1 + n2 = n3$
THEN set as a subgoal to write n3 in that column

This production rule specifies in its condition the goal of working on the tens column and involves a retrieval of a declarative chunk like the 3 + 4 = 7 fact in Fig. 10.1. In its action it creates a subgoal which might involve things like processing a carry. It is many procedural rules like this, along with the declarative chunks, which in total produce what we recognize as competence in domains like mathematics.

It might seem that these chunks and productions are all separate, disjoint pieces of knowledge and that there is nothing in the ACT-R theory to produce the overall organization and structure in cognition. However, this ignores the contribution of the goal structure. So, for instance, in multicolumn addition there is a goal structure that organizes the overall addition into specific column additions. In the case of multicolumn subtraction, as it is typically taught in America, there is also a subgoal of coordinating borrowing especially from zero (Van Lehn, 1990).

What we have described so far is like a rigid computer programming language. If there was only this symbolic level, there would not be the possibility that things might take different amounts of time, that the same person might be variable in how they try to solve a problem, or that people might be variable in the answers they give. However, underlying the ACT-R theory there is an intricate set of subsymbolic computations determining which chunks and productions come to mind (if at all) and how long it takes people to perform steps of cognition. It is these subsymbolic computations that create the continuous, qualitative structure of human cognition.

These computations are not just imperfections added to a cold platonic ideal of perfect cognition. It turns out that these subsymbolic computations play a critical part in enabling ACT-R to adapt to the immense volume of knowledge stored in memory and the variability in the world. These sub-symbolic computations are guided by results from the earlier rational analysis of cognition in Anderson (1990). The next two sections of this chapter will describe how the rational analyses of memory and choice have been mapped into the ACT-R architecture.

The rational analysis of memory and its mapping into the ACT-R architecture

According to the rational analysis of memory, the task of the memory system is to retrieve knowledge that is relevant to the current context. The assumption is that the memory system tends to make memories available as a function of the probability that they are needed in the current situation. This probability is called the ***need probability***. The greater the need probability for a memory the more resources the memory system should devote to retrieving that memory. According to the rational analysis there are two factors that a system can use in estimating the probability that a fact is needed. One is its past history of use where memories that have been used more frequently and recently have higher need probabilities. The second is the information that is being focused upon in the current context. For instance, if one is thinking about dogs it is more likely that memories involving dogs are relevant. Thus, according the rational analysis the probability that a memory is needed is a function of a historical factor and a contextual factor. Calling upon Bayesian theory it was proposed that the odds that a memory is needed is given by the following formula:

$$\frac{P(M \mid H \,\&\, Q)}{P(\overline{M} \mid H \,\&\, Q)} = \frac{P(M \mid H)}{P(\overline{M} \mid H)} * \prod_{j \in Q} \frac{P(j \mid M)}{P(j \mid \overline{M})} \qquad \text{the rational memory equation}$$

where $P(M|H \,\&\, Q)$ is the posterior probability that the memory is needed given its past history of usage H and the current cues Q, $P(\overline{M}|H \,\&\, Q) = 1 - P(M|H \,\&\, Q)$, $P(M|H)$ is the probability just conditionalizing on past history, $P(\overline{M}|H) = 1 - P(M|H)$, the product is over the individual cues j in the context Q, $P(j|M)$ is the conditional probability of cue j being focused upon if memory M is needed, and $P(j|\overline{M})$ is the conditional probability if M is not needed. This equation can be characterized in Bayesian terms as:

$$\text{posterior odds} = \text{prior odds} * \text{likelihood ratio}$$

where history of the memory sets the prior odds and the context sets the likelihood ratio. One can rewrite this in log terms as:

$$\log(\text{posterior odds}) = \log(\text{prior odds}) + \log(\text{likelihood ratio})$$

which is the form from which we derive ACT-R's activation equation. Thus, the basic claim of this rational analysis is that whatever things affected the history factor or the context factor would similarly affect human memory. Lael Schooler in his dissertation (Schooler, 1993) explored how the past history of appearance of a fact in the environment was related to its probability of reappearing. He found effects of frequency, recency, and spacing of presentations which were mirrored by similar effects in human memory. He also studied how effects of associative context interacted with the history factor in the environment (Schooler and Anderson, 1997). Again he found that memory mirrored these effects in the environment. Thus, it does seem that human memory is adaptive in much the way that the rational analysis implies.

When it came time to map this rational analysis of memory into the ACT architecture it seemed that the mapping was fairly straightforward. The memories we were talking about were declarative memories and so it seemed apparent that we were talking about the subsymbolic level of declarative memory in ACT-R. Activation levels controlled the availability of chunks in ACT-R and so it seemed obvious that the rational analysis of memory should be mapped on to a proposal for the setting of activation levels. As activation quantities added rather than multiplied, it seemed obvious to take activation of a chunk as reflecting the log need odds of that chunk. The following became the central equation describing activation levels in ACT-R:

$$A_i = B_i + \sum_j W_j S_{ji} \qquad \text{activation equation}$$

where A_i is the activation of chunk i and reflects the log posterior need odds that chunk i is needed, B_i is the base-level activation of chunk i and reflects the log prior odds, and the strengths of association, S_{ji}, reflect log ratios of conditional probabilities, $log(P(j|M)/P(j|\overline{M}))$, from the rational equation. The source activations, W_j, reflect weighting of the elements j. The use of source activations was something not in the original rational analysis but can be understood in the rational framework as reflecting the validity of j or emphasis on j as a cue.

We also took the prescription of the rational analysis for how these activation levels should map on to behaviour. There was a latency analysis which is not relevant to the current chapter, but the probability analysis is relevant. In particular, if there are a number of memories competing for retrieval the probability that memory i will be retrieved is:

$$\text{Probability of retrieving } i = \frac{e^{A_i/s}}{\sum_j e^{A_j/s}} \qquad \text{Boltzmann equation}^{1}$$

where s reflects the noise in the activation levels and is related to the variance, σ^2, of the noise in the activation levels by the equation $s^2 = 6\sigma^2/\pi^2$. This equation is

identical to one used to describe the probability of a Boltzmann machine being in a state (Ackley *et al.*, 1985; Hinton and Sejnowsky, 1986). In the Boltzmann application the A_i would be the energy of state i and s would be the temperature.

At the subsymbolic level, ACT-R has a theory of how the B_i and the S_{ji} are learned. This involved Bayesian learning processes which try to calculate the best estimates of these quantities given the experiences so far. For current purposes we do not need to detail these learning processes except to say that their net effect is make the B_i reflect log base odds in experience and the S_{ji} to reflect log likelihood ratios in experience. In the later analysis of categorization we will use the assumption that these activation quantities estimate these statistics describing the experience in the environment.

The rational analysis of choice and its mapping into the ACT-R architecture

According to the rational analysis of choice, subjects make choices among options in decision making and problem-solving situations in order to maximize their expected utilities: each option has an expected probability (P) of achieving the goal and expected cost (C) of achieving the goal. If the value of the goal is G the expected gain associated with that option is PG. Subtracting the cost from this gives us the expected ability $PG - C$. The rational analysis claimed subjects choose the option with the highest $PG - C$. This rational analysis might seem to fly in the face of conventional wisdom that people do not behave to maximize their expected utilities (Dawes, 1988). For instance, in a simple probability-matching situation one might expect subjects to always guess the more probable outcome which would maximize number of correct guesses. In contrast, subjects will often choose the more probable alternative with a probability that approximates the empirical probability (Kintsch, 1970). Lovett (1998) accounted for such deviations by assuming there was some randomness in the estimation of these expected values.

It was obvious where in ACT to map this aspect of the rational analysis. Choice of what to do next is handled by what is called the conflict resolution process which selects among various productions. Productions control the direction of behaviour. Thus, we assumed that in conflict resolution ACT-R chose among competing productions according to their expected utility calculated as $U = PG - C$. As in the case of activations, we assumed that there was noise in the actual evaluation process. The same basic Boltzmann equation applied as in the case of activations and the probability of selecting production i with evaluation U_i was:

$$\text{Probability of choosing } i = \frac{e^{U_i/s}}{\sum_j e^{U_j/s}} \qquad \text{Boltzmann equation}$$

where the summation in the denominator is over all available options. Lovett's dissertation research is a nice example of where this analysis applies. In her dissertation studying strategy selection in problem solving, Lovett (1994) developed

what she calls the building sticks task which is illustrated in Fig. 10.2. Subjects are told that their task is to construct a target stick and they are given various sticks to work with. They can either choose to start with a stick smaller than the target stick and add further sticks to build up to the desired length (called the undershoot operator) or to start with a stick longer than the target and cut off pieces equal to various sticks (called the overshoot operator). This task is an analogue to the Luchins waterjug problem (Luchins, 1942). In terms of production rules, Lovett analysed this as a competition between two alternative productions:

Overshoot:

IF the goal is to solve the building sticks task
and there is no current stick
and there is a stick larger than the goal
THEN start with the stick
and set a subgoal to subtract from it.

Undershoot:

IF the goal is to solve the building sticks task
and there is no current stick
and there is a stick smaller than the goal
THEN start with the stick
and set a subgoal to add to it.

She showed that subjects would tend to choose whichever production was more successful. She modelled this by assuming subjects adjusted their estimates of the probability (P) of success associated with each production. This led to adjusted expected $PG - C$ utilities which resulted in different probabilities of choice according to the Boltzmann equation above. In this ACT-R framework she was able to show that there was value to having subjects choose among alternatives probabilitistically according to this formula. This allowed them to learn about the relative pay-offs of both alternatives and adjust their choice preferences when the probabilities changed.

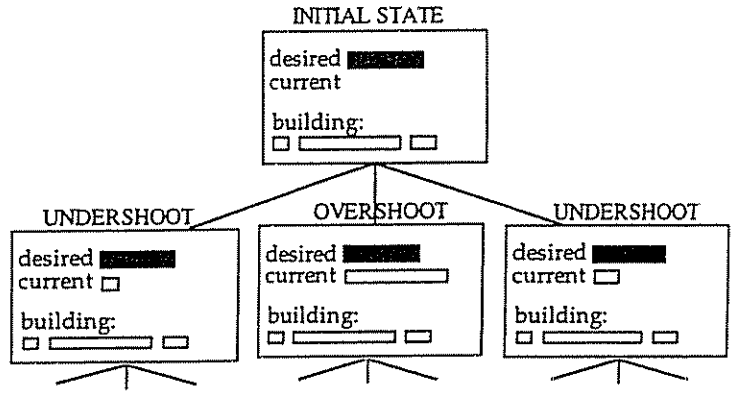

Fig. 10.2 Initial and successor states in the building sticks task

Rational analysis of categorization and ACT-R

Anderson (1990) reported a rational analysis of categorization which was subsequently applied to a broad range of psychological phenomena (Anderson, 1991) and machine learning data sets (Anderson and Matessa, 1992). The basic assumption in the rational model of categorization is that there are real categories out there in the world. Species of animals are described as the paradigm case of such categorization where the animal kingdom can be partitioned into almost non-overlapping groups (defined by prohibitions against interbreeding among species) which share a set of features probabilistically (defined by the gene pool within a species). It is argued that identifying the category of an object is functionally valuable because it enables one to make predictions about the object. Thus, for instance, classifying a striped object as a tiger versus zebra is valuable because it allows one to make predictions about the relative danger of the object.

The rational analysis proposes that the basic task of categorization is predicting the probability that an object will display a specific feature given that it displays certain other features. Thus, if an animal approaches us of a certain size with certain features like striping, we might be interested in predicting whether it is dangerous. Under the rational analysis, this is done by determining the probability that the object is in various categories and the probability that it will display the feature given that it is in each category. Thus, the predicted probability $Pred(i|F)$ that an object with observed features F will display feature i is given as

$$Pred(i|F) = \sum_k P(i|k)P(k|F)$$

where the summation is over the various categories k and $P(k|F)$ is the probability of coming from category k given that features F are observed and $P(i|k)$ is the probability of displaying feature i given that the object comes from category k. Murphy and Ross (1994) argued that rather than taking a weighted average over categories, subjects simply inquire as to $P(i|k)$ for the category with the highest $P(k|F)$.

Under either the original rational analysis or the Murphy and Ross analysis it becomes critical to be able to calculate $P(k|F)$—either to weight the various categories or to determine the most probable category. This is where the heart of our analysis of categorization behaviour lies. The following Bayesian equation relates this to the prior probability, $P(k)$, of an object coming from category k and the probability of displaying features F given that the object comes from category k.

$$P(k|F) = \frac{P(k)P(F|k)}{\sum_k P(k)P(F|k)}$$

The rational categorization equation

As many Bayesian approaches, this predicts that we should see behaviour as being sensitive to both the base rates, $P(k)$, and the conditional probabilities $P(F|k)$.

Gluck and Bower data set

In the past we have applied this model to data collected by Gluck and Bower (1988) and in this chapter we will be considering applications of ACT-R to this experiment. Gluck and Bower had subjects classify patients as having a rare or a common disease given a set of symptoms. In the experiment we will be considering, Gluck and Bower's third experiment, subjects saw 250 patient descriptions consisting of binary features on four dimensions. Patient descriptions were generated by first randomly selecting the common disease with probability 0.75 and with the rare disease with probability 0.25. (This corresponds to the prior probability of the category, $P(k)$, in the rational categorization equation.) Then symptoms were then randomly generated to describe the patient with probabilities conditional on the disease category. (These are the conditional probabilities, $P(F|k)$, in the rational categorization equation.) They found that subjects are sensitive both to base rates of the disease and to the conditional probabilities of those symptoms for that disease in just the way that the rational categorization equation would imply. In fact in these experiments subjects come close to probability matching, which is to say they diagnosed a symptom combination as common with a probability that closely matched its actual probability of being common.

Table 10.1 describes the structure of the Gluck and Bower material. There we have the 16 possible stimulus patterns defined by the values (symptoms) on the four dimensions. The table also gives the probability $P(F)$ of the pattern overall, the probabilities of the pattern given the common $P(F|c)$ or rare disease, $P(F|r)$ and the probability $P(c|F)$ of the common disease given the pattern. It is this last quantity which should be directly relevant to the subjects' decision making. Note that there are only four patterns where $P(c|F)$, the probability of the common disease given the pattern, is less than 0.5. These are the four cases where there is a + symptom on dimension 1 and a – symptom on dimension 4. Thus, subjects need only look and the first and fourth dimensions in order to make their classification behaviour. The

Table 10.1 Structure of Gluck and Bower material

dim 1	dim 2	dim 3	dim 4	P(F)	P(F|c)	P(F|r)	P(c|F)
1+	2+	3+	4+	0.014	0.014	0.014	0.750
*1+	2+	3+	4–	0.022	0.010	0.058	0.333
1+	2+	3–	4+	0.025	0.022	0.034	0.659
*1+	2+	3–	4–	0.044	0.014	0.134	0.243
1+	2–	3+	4+	0.031	0.034	0.022	0.824
*1+	2–	3+	4–	0.038	0.022	0.086	0.438
1+	2–	3–	4+	0.050	0.050	0.050	0.750
*1+	2–	3–	4–	0.076	0.034	0.202	0.333
1–	2+	3+	4+	0.046	0.058	0.010	0.947
1–	2+	3+	4–	0.038	0.038	0.038	0.750
1–	2+	3–	4+	0.070	0.086	0.022	0.920
1–	2+	3–	4–	0.066	0.058	0.090	0.659
1–	2–	3+	4+	0.104	0.134	0.014	0.966
1–	2–	3+	4–	0.082	0.090	0.058	0.824
1–	2–	3–	4+	0.160	0.202	0.034	0.947
1–	2–	3–	4–	0.134	0.134	0.134	0.750

*Cases where the rare disease (r) is more probable.

'optimal' behaviour in a traditional sense would be to classify the disease causing the symptoms as rare only if the first dimension has the positive value and the fourth dimension has the negative value. However, there is at least one important caveat on the degree to which we can expect subjects to display this optimal behaviour. As stimuli are randomly generated and there are only 250 trials, the odds are low that subjects will get enough experience with each pattern to get a representative sample so they can know how to classify that pattern. Thus, to some degree subjects' experience with these patterns will fail to be representative.

Figure 10.3 shows, for the last 50 trials, the probability of classifying each of the 16 feature combinations as the common disease. This is plotted as a function of the actual probability that this feature combination would display the common disease. The figure displays the data from Gluck and Bower and the predictions of two ACT-R models that we will describe. As can be seen the Gluck and Bower subjects come close to probability matching which indicates that they were sensitive to both base rates and the conditional probabilities. Thus, the experiment shows that subjects are quite in tune with the experienced probabilities as the rational model would imply. The rational model does not have a response function and thus does not actually predict the probability that subjects would choose the common disease. It only predicts that only that their choice probability would be correlated with the experienced probability.

There is another set of data in this experiment which is difficult to accommodate in the rational model directly, which is problematical for many other models of

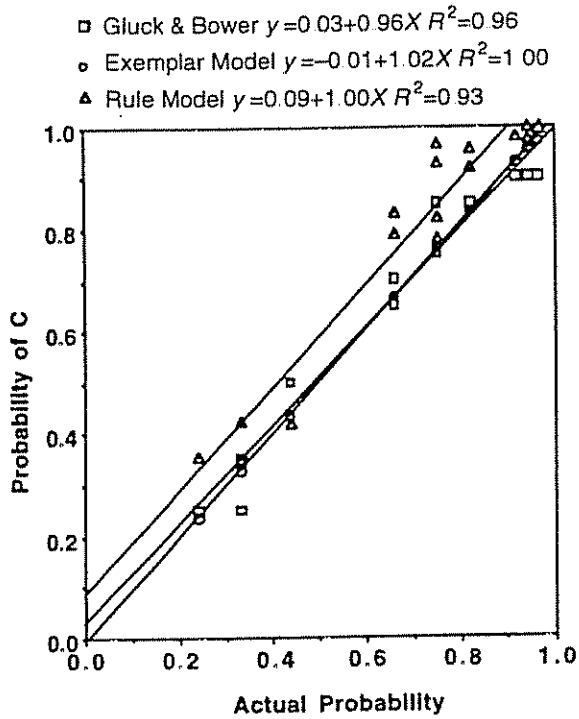

Fig. 10.3 Probability of a diagnosis of the common disease Data is from Gluck and Bower (1988) and the predictions are from two ACT-R models

categorization, and which is frequently thought of as the most interesting aspect of the experiment. After completing 250 trials of categorization subjects were shown each of the eight symptoms separately and asked to rate the probability of the common disease and the rare disease given that symptom. Figure 10.4 shows the results from Gluck and Bower and the predictions of two ACT-R models. Because of the difference in base rates it turns out that none of the symptoms individually was associated with the rare disease in a majority of the training trials. That is to say, for no symptom is the conditional probability of the disease given the single symptom greater than 0.5. There is one symptom for which the conditional probability is 0.5. This is the positive value on the first dimension (denoted 1+ in Fig. 10.4) for this symptom subjects show a greater than 50% choice, demonstrating a base-rate neglect in their probability estimates. It is difficult to predict this result from a number of models including exemplar models. In exemplar models there will be an equal representation in memory of instances with the symptom and the rare disease and instances with the symptom and the common disease.

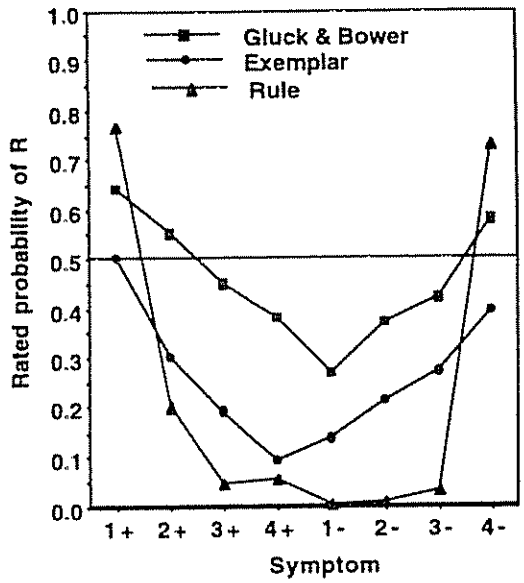

Fig. 10.4 Rated Probability of a disease given a symptom. Data is from Gluck and Bower (1988) and the predictions are from two ACT-R models

If one assumes that subjects' reports reflect strengths of associations between the symptoms and the diseases, Gluck and Bower show that a greater than 0.5 rating of symptom 1+ follows from competitive learning algorithms such as these involving the delta rule (Rescorla and Wagner, 1972). This is basically because competitive learning algorithms need to increase the strength of association to the symptom most predictive of the rare disease so that the subject has a basis for responding 'rare' to those symptom combinations where rare is the more common diagnosis. There have been a number of other efforts to account for this base-rate neglect and the related inverse base-rate effect demonstrated by Medin and Edelson (1988). Many of these are connectionist models including Gluck (1992), Schanks (1992), and Kruschke (1996). In all cases the fundamental idea is that in order to compensate for a strong association of some symptoms to the common disease, the learning system

has to give special emphasis to symptoms that are distinctive of the infrequent category.

The rational model predicted that subjects should be able to respond to single dimensions with probabilities that reflected their actual association with the category. Thus, it predicts no base-rate neglect. However, this failure to predict base-rate neglect depends on its independence model for combining dimensional information made for analytic convenience rather than anything fundamental about the rational model. Elsewhere (Anderson and Fincham, 1996) we have showed that this independence assumption is neither valid empirically nor desirable from an adaptive perspective. However, without the independence assumptions the rational model cannot predict how subjects will generalize from training on multidimensional stimuli to tests on single-dimensional stimuli. The correct approach from a rational analysis perspective would be to inquire of the environment what the relationship is between the statistics that map multisymptom presentations onto categories and the statistics that map single-symptom presentations onto categories. Such data are probably available in domains like medical diagnosis but we have not mined it.

ACT-R's analysis of categorization

One might hope that we could go to an architecture like ACT-R to answer the question of how multidimensional experience should be mapped into single-dimension predictions. However, unlike memory and choice, there is no architectural primitive in ACT-R that corresponds to the rational analysis of categorization. This is not to say that ACT-R cannot account for categorization behaviour. It certainly can model tasks like the Gluck and Bower experiment. In such models it would view subjects as trying to solve the problem of labelling the stimuli. There are at least two ways for ACT-R to try to solve this problem—corresponding to the two major types of theories in the categorization literature. First, ACT-R can try to retrieve an example similar to the test stimulus and categorize the test stimulus with whatever category is stored with the example. This is an exemplar model and it will rely on ACT-R's theory of declarative memory which stores past examples. Alternatively, ACT-R can also take a rule-based approach and use production rules to process the values on the dimensions as voting for one category or another. This will rely on ACT-R's theory of procedural memory. Thus, the ACT-R position is that categorization is not an architectural primitive but rather in studying categorization we are studying a problem-solving task which ACT-R can solve by different configurations of its architectural primitives. We will discuss in detail our efforts to use the exemplar-based and the rule-based approach to model the Gluck and Bower task in ACT-R.

The exemplar approach in ACT-R

Figure 10.5 represents how a particular symptom pattern could be represented in ACT-R's declarative memory. It is a chunk associating the symptoms with the disease diagnosis. Basically, there would be 32 such chunks learned in the experiment—one to represent each of the 16 symptom patterns associated with each

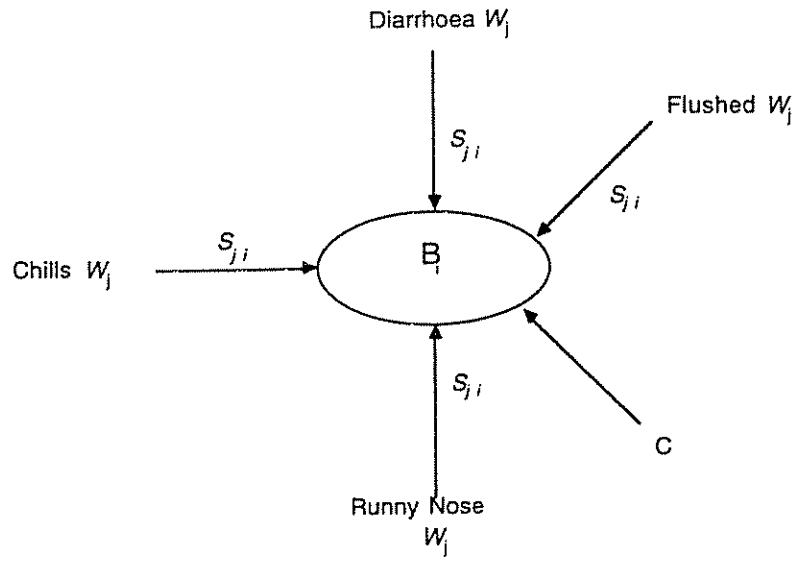

Fig. 10.5 ACT-R's representation of a symptom pattern associated with a disease

of the two diseases. These chunks would vary in terms of their base-level activations which in turn would reflect their frequencies of presentation which can be calculated from Table 10.1. When a particular pattern is presented for judgement we assume that the following production rule would apply to classify it:

IF the goal is to classify a patient with symptom pattern W, X, Y, & Z
and there is an instance of disease D associated with symptom pattern
W, X, Y, and Z
THEN say the disease is D.

For any test pattern there would be two chunks that might be matched to the second clause in the production above. Depending on which chunk was retrieved either the common disease or the rare disease would be assigned as the value of D. The relative probability of retrieving these two diseases would depend on the relative activation of the two chunks which is given by the earlier activation equation. It turns out that the strengths of associations S_{ji} from the symptoms to the two competing chunks will be the same[2] and the only difference in activation levels comes from the base-level activations, B_i, which will reflect the frequencies, N_i, of the chunks. According to the ACT-R theory these base level activations will reflect the log frequencies. More specifically, in recent applications we have assumed that the base-level activation grows half as fast as the log frequency. Thus,

$$B_i = 0.5^* \ln(N_i)$$

Substituting into the earlier Boltzmann equation we can now convert these base-level activation into a probability of the common disease:

$$P(c) = \frac{N_c^{0.5/s}}{N_c^{0.5/s} + N_r^{0.5/s}}$$

where N_c and N_r are the frequencies of the common and rare chunks for that disease pattern.

The only free parameter in fitting this model to the data is s which reflects the noise in the activation levels. We fit the model to the classification data from Gluck and Bower (the 16 points in Fig. 10.3). The estimate of the parameter s was 0.705 which is in keeping with the sub-1 recommendation of Anderson and Lebiere (1998). The R^2 between predicted and observed data is 0.965. As the subjects, with this value of s, ACT-R is basically probability matching.

There remains the question of how to relate the model to the ratings data in Fig. 10.4. The task subjects have been trained on is to classify the disease patterns of four symptoms and not to assign a number to a single symptom. The production rule above could be easily generalized to a single symptom. In this case the rule would just be:

IF the goal is to classify a disease with symptom W
and there is an instance of disease D associated with symptom W
THEN say the disease is D

However, this is just a rule for saying common or rare. It does not provide a basis for assigning a number to reflect a probability. To deal with the probability assignment task we took a very concrete solution. We assumed that the subject set up a loop to classify the disease a number of times and converted the proportion of rare classifications into a probability. This is a mechanism to enable the subject to assign a number whose expected value is equal to the proportion of common disease classifications. Other researchers have asked subjects to classify the diseases and found that classification proportions are similar to the ratings (Estes *et al.*, 1989; Shanks, 1990; Nosofsky *et al.*, 1992; Cobos *et al.*, 1993; Myers *et al.*, 1994). Thus, our prediction is that the proportion of rare diseases assigned to a symptom would be:

$$Prop(R) = \frac{\sum_{i \in R} N_i^{0.5/s}}{\sum_{i \in R} N_i^{0.5/s} + \sum_{j \in C} N_j^{0.5/s}} \qquad \text{exemplar equation}$$

where the summations are over the instances that have that symptom and the rare disease (R) and the instances that have that symptom and the common disease (C). This is the basis for the predictions in Fig. 10.4. We used the same s parameter that was estimated for Fig. 10.3. Thus, this is a parameter-free prediction.

As can be seen, this mechanism generally underpredicts the rated probabilities of the rare disease. In particular, for the critical 1+ stimulus it predicts almost exactly the experienced value of 0.50 while subjects in the Gluck and Bower experiment are giving 64% rare disease ratings. Thus, in essence, we have replicated the standard failure of the exemplar-based models to account for the rating data in the Gluck and Bower experiment. While the ACT-R equation provides for a non-linear mapping of frequency on to probability it provides little basis for escaping the logic that if there has been an equal number of examples mapping the symptom on to both diseases, the probability of classification should be about 0.5.

The rule-based approach in ACT-R

An alternative approach to this task is to assume that the subject is following some set of rules for classifying the stimuli into categories. There are many rule systems subjects might follow but we assumed they followed the rule system sketched out in Table 10.2. Essentially, the subject looks at each value on each dimension and counts that value as either voting for the common or the rare disease. The subject totals the counts and then chooses the disease with the higher count. If there is a tie, we let the system choose R, which is a bit arbitrary but gives it a basis for responding appropriately in the four rare cases in Table 10.1.[3] There are 16 processing rules corresponding to the two ways of categorizing each of the two values on each of the four dimensions. There are two decision rules for responding to the count.

This approach assumes that there are rules to classify each symptom in either of two ways. Which way is taken will depend upon our theory of choice sketched out earlier. So, for instance, a subject will have one rule to classify stuffed nose as voting for the rare disease and another rule as voting for the common disease. Adapting the Boltzmann equation, the relative proportion of choices between the two will depend on the expected gain of the two productions according to the equation:

$$P(C) = \frac{e^{U_c/s}}{e^{U_c/s} + e^{U_r/s}} = \frac{e^{(U_c - U_r)/s}}{1 + e^{(U_c - U_r)/s}} \qquad \text{rule equation}$$

where U_c is the expected value of the common disease and U_r is the expected value of the rare disease and s is the noise parameter. Thus, the subject will choose among common or rare as a function of the relative pay-off of these two rules.

Table 10.2 Production rules for the Gluck and Bower task

Processing rules (a set of four such rules for each dimension)
IF the value of dimension *X* is +
THEN increment the count for *C*
IF the value of dimension *X* is +
THEN increment the count for *R*
IF the value of dimension *X* is –
THEN increment the count for *C*
IF the value of dimension *X* is –
THEN increment the count for *R*
Decision rules
IF the final count is greater for *C* than *R*
THEN choose *C*
IF the final count is greater for *R* than *C* or equal to *C*
THEN choose *R*

Altogether there are eight pairs of productions rules in competition—two for each of the eight symptoms. Their utilities are basically measures of how often the firing of

that production is associated with a successful classification. That is, the Us are measures of the probability of correct classifications when that production fires. There is an interesting recursive assessment problem here in that the utilities will determine the probabilities of the productions firing and the probabilities of their firing will determine their utilities. Moreover, this is compounded by a credit-assignment problem because successful classification does not depend on any one production's firing but rather on the whole sequence of production firings together with the decision rules. None the less, it is the case that for most values of the noise parameter, s, there is a stable state that the system will converge upon such that the probabilities for all 16 processing productions will result in a set of utilities which in turn produce these production probabilities. Thus, as in the case of the exemplar-based ACT-R model, we can search for a value of s which gives a best fit to the Gluck and Bower data.

Table 10.3 displays the set of utility differences and probabilities that result when we estimate s to be 0.027. This is the value of s which results in the best fit to the classification data in Fig. 10.3. The utility differences in Table 10.3 are between the probabilities of success when the common production fires for that symptom and when the rare production fires. The probabilities can be derived from the utility differences according to the rule equation given earlier. It is not so easy to derive, but it can be shown that these utility differences result when the probabilities in Table 10.3 are in force. Looking at Figure 10.3 it can be seen that this model gave a pretty good approximation to the classification data, although it gave about 10% more C ratings than did the Gluck and Bower subjects. The R^2 between theory and data is 0.899, which is somewhat worse than the exemplar model.

Table 10.3 Probabilities and utility differences for rules

	Utility differences among productions $(U_c - U_r)$	Probability of common productions
1+	−0.032	0.235
1−	+0.158	0.997
2+	+0.037	0.799
2−	+0.128	0.991
3+	+0.081	0.953
3−	+0.092	0.968
4+	+0.077	0.946
4−	−0.027	0.271

With respect to the single-symptom rating data we assumed that subjects gave their number, as in the exemplar model, to reflect the number of repeated classifications that resulted in a common disease categorization. In this case, the probability just reflects the probability that the subject will classify that symptom as rare. As can be seen from Fig. 10.4 this rule model, unlike the exemplar model, has

no difficulty in producing over 50% rare classifications to the 1+ symptom. This is because the process of learning utilities in ACT-R is in effect a competitive learning algorithm, such that it learns that it will optimize its overall classification by treating 1+ as a symptom of the rare disease. In fact the problem with the predicted ratings is generally that they are more extreme than the observed ratings. It would be possible to correct this problem if we assumed subjects introduced some attenuation in the extremity of the ratings that they gave.

Conclusions

Both the exemplar-based and rule-based ACT-R models have done reasonable jobs of accounting for the subject's classification in the training phase of the experiment. However, the exemplar model reflected the standard deficit of exemplar models in accounting for the 'base rate' neglect in the judgement data. The rule-based model does not have this difficulty. Although its predictions are too extreme, extra assumptions about attenuation could get them quite close. However, we would not want to conclude that the rule-based model is the better account. It is extremely dubious that all subjects respond to the Gluck and Bower task in the same way. Probably some mixture of exemplar and rule models provides the best account of the data. Such mixture models also reflect one current direction of Nosofsky and Palmeri's thinking about categorization (Nosofsky *et al.*, 1994; Palmeri and Nosofsky, 1995). Indeed, a virtue of ACT-R is it allows a mixture of models to be expressed in one framework. The thing we want to emphasize, rather, is that with just one estimated parameter both models provided reasonable fits to the classification data by just responding to the statistical structure of the experiment. In both cases, that one parameter was a measure of how much noise there was in the mapping of the experimental structure on to the subsymbolic quantities of the system.

One might think this noise in the underlying subsymbolic quantities reflects some sort of imperfection in the system. It might seem that we have case of a system being as good as it can given the imperfections of its hardware. This might be an accurate characterization of the situation but there are other ways of understanding what is going on. Rather, it could be that this noise is actually adaptive because it causes the system to consider possibilities other than the one that currently appears optimal. As Lovett (in press) develops, this allows the system to be able to better learn information about the environment and particularly to be more responsive when some environmental change reorders the relative merit of the choices. In Boltzmann machines (Ackley *et al.*, 1985; Hinton and Sejnowsky, 1986) such stochasiticity is used to search and discover the structure of the environment. Our central equation, while it derives from assumptions other than those in the Boltzmann machines, is the same equation and also reflects the idea that variability can be useful in enabling a system to sample the structure of its environment.

Categorization is not an architectural primitive in ACT-R in the same way as memory and choice are. Still we have produced two separate implementations of categorization in ACT-R which are almost equivalently capable of responding

adaptively to the statistical structure of their training environment (although they generalize differently to the rating task). The adaptiveness of the categorization behaviour they display derives from the adaptiveness of the architectural implementations of memory and choice. Thus, we see that the rational analysis offered in Anderson (1990) missed a significant generalization. It treated categorization as a separate optimization problem when in fact its adaptive performance could be derived either as a consequence of adaptive procedures for retrieving memories or adaptive procedures for choosing among rules for behaviour.

It might seem that both the exemplar and rule-based approaches cannot be equally adaptive because they generalized differently to the judgement task in Fig. 10.4. However, the thing to recognize about that task is that there is not really a definition of correct behaviour in the rating task. Subjects have not experienced single symptoms before and have no real basis for knowing how to assign numbers. Saying that subjects should assign probabilities that correspond to conditional probabilities is a classic example of 'theorist error' where it is assumed that people should imagine their task in the terms that the theorist has in mind. What is relevant to the decision of adaptivity is whether the two models can respond appropriately to the statistical structure of the task on which they are trained.

Finally, as a background comment we find that ACT-R is capable of providing good models of a task with essentially estimating nothing but a noise parameter. This ability to deploy an ACT-R model without any search through possible models and get reasonable predictions in a nearly parameter-free way is testimony to the power of rational analysis for driving the design of a cognitive architecture at the subatomic level.

In summary, we would argue that the best research plan is to work jointly on cognitive architecture and rational analysis. The rational analysis provides guidance for the design of the architecture. With the architecture in place we are in a better position to understand the meaning of these rational constraints.

Acknowledgements

The research reported in this paper was supported by grant N00014-96-I-0491 from the Office of Naval Research and SBR-94-21332 from the National Science Foundation. We would like to thank Chris Schunn for his comments on the paper.

Notes

1. This equation can be used to predict recall failure by assuming there is some activation threshold which just serves as an A_i in the equation above. If the threshold is the largest activation (i.e. all chunks have activation below the threshold) then nothing will be recalled.
2. In the ACT-R model, the strength of association between cue j and chunk i is defined as $\ln(P(j|i)/P(j))$. In these experiments $P(j|i) = 1$ as the symptom j is always present when a disease pattern i is to be retrieved. Thus, $S_{ji} = -\ln(P(j))$ which does not depend on the chunk i being retrieved.
3. The four R cases in Table 10.1 are tie situations because they occur when the two disease symptoms (1+ and 4−)with strongest association to R are present with two other disease symptoms that will have stronger associations to C.

References

Ackley, D. H., Hinton, G. E., and Sejnowsky, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, **9**, 147–69.

Anderson, J. R. (1976). *Language, memory, and thought*. Lawrence Erlbaum, Hillsdale, NJ.

Anderson, J. R. (1983). *The architecture of cognition*. Harvard University Press, Cambridge, MA.

Anderson, J. R. (1990). *The adaptive character of thought*. Erlbaum, Hillsdale, NJ.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, **98**, 409–29.

Anderson, J. R. (1993). *The rules of the mind*. Lawrence Erlbaum, Hillsdale, NJ.

Anderson, J. R. and Fincham, J. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 259–77.

Anderson, J. R. and Lebiere, C. (1948). *The atomic components of thought*. Lawrence Erlbaum, Mahwah, NJ.

Anderson, J. R. and Matessa, M. (1992). Explorations of an incremental, Bayesian algorithm for categorization. *Machine Learning*, **9**, 275–308.

Anderson, J. R., Corbett, A. T., Koedinger, K., and Pelletier, R. (1995). Cognitive tutors: lessons learned. *The Journal of Learning Sciences*, **4**, 167–207.

Cobos, P. L., López, F. J., Rando, M. A., Frenández, P., and Almaraz, J. (1993). Connectionism and probability judgment: suggestions on biases. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pp. 342–6. Lawrence Erlbaum, Hillsdale, NJ.

Dawes, R. (1988). *Rational choice in an uncertain world*. Harcourt, Brace, and Jovanovich, New York.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., and Hurwitz, J. B. (1989). Base-rate effects in category learning: a comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 556–76.

Gluck, M. A. (1992). Stimulus sampling and distributed representations in adaptive network theories of learning. In *Essays in honor of William K. Estes* (ed. A. Healy, S. Kosslyn, and R. Shiffrin), Vol. 1, pp. 169–99. Lawrence Erlbaum, Hillsdale, NJ.

Gluck, M. A. and Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General*, **117**, 227–47.

Hinton, G. E. and Sejnowsky, T. J. (1986). Learning and relearning in Boltzmann machines. In *Parallel distributed processing: explorations in the microstructure of cognition*, (ed. D. E. Rumelhart, J. L. McClelland, and the PDP Group), Vol. 1. *Foundations*, pp. 282–317. MIT Press, Cambridge, MA.

Kintsch, W. (1970). *Learning, memory, and conceptual processes*. Wiley, New York.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 3–26.

Lovett, M. C. (1994). *The effects of history of experience and current context*. Doctoral Dissertation Pittsburgh, PA: Carnegie Mellon University.

Lovett, M. C. (1998). Modeling choice behavior in humans and animals. In *The atomic components of thought*, (ed. J. R. Anderson and C. Lebiere) Lawrence Erlbaum, Mahwah, NJ.

Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monographs*, **54** (248).

Medin, D. L. and Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, **117**, 68–85.

Murphy, G. L. and Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, **27**, 148–93.

Myers, J. L., Lohmeier, J. H., and Well, A. D., (1994). Modeling probabilistic categorization data: exemplar memory and connectionist nets. *Psychological Science*, **5**, 83–9.

Nosofsky, R. M., Kruschke, J. K., and McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**, 211–33.

Nosofsky, R. M., Palmeri, T. J., and McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53–79.

Palmeri, T. J. and Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal or Experimental Psychology: Learning, Memory and Cognition*, **21**, 548–68.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology*, **42A**, 209–37.

Schanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science*, **4**, 3–18.

Schooler, L. J. (1993). *Memory and the statistical structure of the environment*. Doctoral Dissertation Pittsburgh, PA: Carnegie Mellon University.

Schooler, L. J. and Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology*, **32**, 219–50.

Van Lehn, K. (1990). *Mind bugs: the origins of procedural misconceptions*. MIT Press, Cambridge, MA.