In M. Eid & E. Diener (Eds.), Handbook of Multimethod Measurement in Psychology (pp. 353-369). Washington, DC: American Psychological Association. 2006

CHAPTER 24

MULTIMETHOD APPROACHES TO THE STUDY OF COGNITION: THE EVOLUTION OF CONCEPTS IN RESEARCH ON HUMAN MEMORY

Aaron S. Benjamin

Similar to many scientific pursuits within psychology, the study of human cognition is an exercise that is equal parts imagination, deduction, and salesmanship. Theoretical claims are bootstrapped onto the elaborate but typically freewheeling artifices constructed by fellow psychologists who maintain equally fragile footing. Despite the blustery nature of cognitive theorizing, a central question remains unresolved: What constitutes necessary and sufficient evidence for the existence of a psychological mechanism?

Even the earliest theorists encountered situations in which multiple measures of nominally equivalent cognitive processes had different psychometric properties and showed differential effects of a common manipulation. Ebbinghaus (1885) noted, for example, that measures of relearning were much more sensitive to distant prior experience than measures of recall. Much of the history of cognitive psychology can be interpreted in the context of debates about how to reconcile such differences. The purpose of this chapter is to provide an illustration of how modern cognitive psychology deals with the divergences and convergences made apparent by the use of multiple measures and, in doing so, how those effects can be used profitably in the development of theory and the postulation of mental systems.

I will not attempt to address well-developed statistical tools that are the focus of chapters 18 to 21 and others in this volume. Rather, I will concentrate on model-based interpretations of multiple measures and how the application of such techniques has advanced theoretical development in cognitive psychology. In doing so, I will review four topics related to the specific problems addressed by and applications of multimethod approaches to understanding cognition. In the first and largest section, I will examine several modern examples of how measurements that combine systematically related dependent variables can yield functions that are more reliable and more informative than ones that can be derived from single measures. The second section will focus on the evaluation of the theories of cognition, most specifically on the question of how formal models can be tested in such a way that emphasizes their ability to account for extant data patterns without being so powerful that they predict other invalid data sets. Third, we will address the question of how traditional behavioral measurements in cognitive psychology can be meaningfully integrated with brain-based measures assessing electromagnetic properties of cellular material in the brain or hemodynamic properties of blood flow to the brain. Finally, we will examine one domain in which prominent theorists have tried to establish guidelines for what kind of and how much evidence is necessary for the postulation of a mental system.

To tie these sections together, the accompanying examples in each section will draw on current and historical developments in research on memory, with the objective of illustrating to the reader how the judicious combination of different measures has motivated important theoretical developments in that field.

COMBINING MEASUREMENTS TO YIELD GENERALIZABLE PSYCHOLOGICAL FUNCTIONS

Often we wish to measure cognitive performance in a domain in which behavior is strongly and systematically related to some individual difference variable that we are not concerned with. This fact poses two problems from a measurement perspective. First, it adds a source of variability to our sampling distributions. This problem can be annoying and may force us into increasing the sample size of our experiment, but it is hardly fatal. A second, more dangerous, effect is that individual differences may relegate our measurements to a region of parameter space that does not reflect a meaningful or complete range of the behavior in question.

Three general strategies exist to counter the negative effects of individual differences limiting the range of our measurements. First, the researcher can use established theoretical principles in a domain to interpolate or extrapolate to portions of the function that are sparsely occupied by data. Second, the missing data can be inferred statistically by fitting a parsimonious function to the data, such as the lowest order polynomial that accounts for some predetermined proportion of the data. Third, researchers can use a data-collection strategy that ensures sampling across the range of the measurement in question. This can be done by strategically varying the conditions or instructions of an experiment in such a way so as to induce variability along the individual-difference dimension. By doing so, the function relating that dimension to the performance measure can be estimated for each subject. Here I lay out two examples of how this technique is commonly used in memory research. In both of these cases, the solution to the problem of confounding individual differences lies in the elicitation of measures across multiple strategically varied conditions.

Speed–Accuracy Trade-Offs in Recognition Memory

Consider an experiment in which subjects are asked to make a recognition judgment—that is, to decide for each in a list of stimuli whether they believe to have seen that item in a particular earlier study episode. One subject might not care much about the advancement of science, want to get out and get to lunch, and thus zip his way through our task as quickly as humanly possible, making each decision after only the least amount of deliberation. Another subject might feel as though the experimenter will treat her score as a measure of intelligence, character, or trustworthiness and thus pore over each test stimulus to extract every available mote of information from memory before making a recognition decision. Such individual differences are commonplace in decision tasks like this one. Even if we use some between-subject manipulation of learning, for example, we have faith that random assignment will wash away such strategic differences over our sample.

But what if our entire sample was like the first hypothetical subject described earlier? This scenario is not entirely unlikely at many major American universities. Our laboratory might be aesthetically unappealing, or our experimenters might have bad breath; such factors can also influence strategy selection in our subjects.

Hypothetical group means are shown in the top panel of Figure 24.1 and indicate no effect of our learning manipulation. It would be useful to know if there is a restriction placed on our data by an inadequate range of decision speeds. In this case, all subjects performed the task quickly, but we have no way of assessing that fact. Even if we measured decision response time (RT), we would be ill equipped to make any such judgments without a sense of what the "full" parameter range of response speeds should be. The solution to this problem is to create a within-subjects variable along which we manipulate the decision placement along the speed–accuracy trade-off spectrum. We might, for example, use payoffs for different combinations of correct or speedy decisions. We might simply instruct the subjects to make decisions quickly or to take their time. Perhaps most effectively, we can force subjects to withhold their response until a delimited amount of time has elapsed and then force them to make their response within a given time window (Reed, 1973). If we use such a strategy, we ensure the collection of performance data across a reasonable range of decision speeds. We

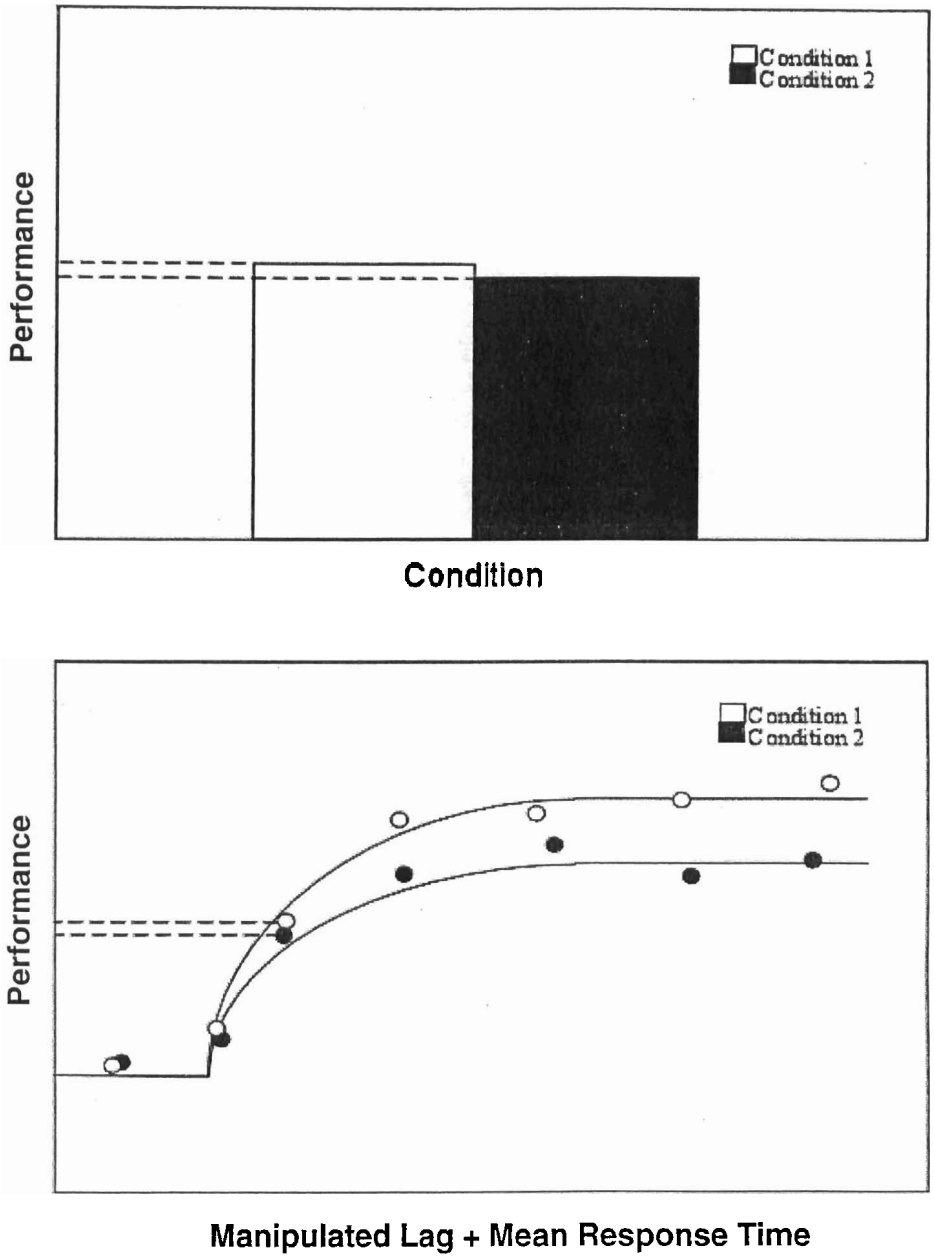


FIGURE 24.1. Group means *(top panel)* and speed–accuracy trade-off functions *(bottom panel)* for two hypothetical conditions.

can also clearly detect those subjects that ignore our manipulation and treat them and their data appropriately.

The data in the bottom half of Figure 24.1 show what such figures look like. The data here have been fit with a shifted exponential function,

$$P = A\left(1 - e^{-R(t-I)}\right) \text{ for } t \geq I \qquad (1)$$

in which A represents asymptotic accuracy, R the rate of approach to the asymptote, I the point at which performance first rises above the floor of chance performance on the task, and t the time point after the onset of the stimulus. One important aspect of such a function is that is can be used to describe behavior for each subject. Whereas an individual mean provides only a scalar value that is some unknown combination of performance and

individual-difference characteristics, this function provides estimates of performance across the entire meaningful range of the confounding individual-difference variable. And, by doing so, we can now see that our failure to detect group differences in the top part of the Figure 24.1 owed in large part to the fact that our subjects, by virtue of their inherent laziness and consequent choice of a particularly speedy decision strategy, placed themselves in a range in which it would have been quite difficult to detect an effect of our learning manipulation.

Figure 24.2 displays some actual results that demonstrate how this technique has proven useful in evaluating important theoretical questions in human cognition. In the top half of Figure 24.2 are empirical speed–accuracy functions for the endorsement of studied and unstudied high- and low-frequency words (Hintzman, Caulton, & Curran, 1994). As is commonly found, recognition is superior for low-frequency words in two ways: the rate of correct endorsement for studied items, or hit rate, is higher, and the rate of incorrect endorsement of unstudied items, or false-alarm rate, is lower, thus yielding a *mirror effect* (Glanzer & Adams, 1990). Most theoretical stances are in agreement about the nature of the difference in hit rate: The presentation of an uncommon word constitutes a distinctive event, and distinctive events are more memorable. However, there are several different extant proposals as to the nature of the difference in false-alarm rate. One suggestion is that the higher false-alarm rate to common words reflects the fact such words enjoy higher baseline levels of familiarity because of the greater number and frequency of exposures to such words, by definition (e.g., Glanzer & Adams, 1985; Hintzman, 1988).

Another suggestion is that recognition decisions are made after two sources of evidence are assessed. First, the word is matched against memory, yielding an overall assessment of mnemonic familiarity. Second, the word is evaluated as to its likely memorability, and recognition standards are set that are commensurate with that assessment (e.g., Benjamin, Bjork, & Hirshman, 1998; Brown, Lewis, & Monk, 1977). That is, after determining how familiar a word is, the subject makes a metamnemonic assessment of how familiar it *would* be, if the word had been studied. Because subjects know high-frequency words to be less memorable, they set lower standards for such words and therefore endorse unstudied high-frequency words at a higher rate (Benjamin, 2003; cf. Wixted, 1992). Central to this suggestion is the idea that this postretrieval assessment is deliberate and should only be evident if enough decision time has elapsed for the subject to incorporate such knowledge.

As can be seen in Figure 24.2, the difference in false-alarm rate appears in each response period, including the very short ones. This result is inconsistent with the concept of a postretrieval assessment. However, if these data had not been collected across a spectrum of decision times, this conclusion would have been impossible to reach.

Now consider the display in the bottom half of Figure 24.2, which depicts results from a different recognition experiment. In that experiment, subjects studied multiple lists, each of which consisted of words that were semantically associated to a single, unstudied "critical" word (cf. Roediger & McDermott, 1995). At test, the distractor set included words that were unrelated to the themes of the study lists and also the critical unstudied high associate mentioned before. An interesting pattern of false endorsement of the critical foils is evident: The rate first rises and then falls with decision time (Heit, Brockdorff, & Lamberts, 2004). Notably, if one assessed only a limited range of the speed–accuracy function here, one could conclude that false-alarm rate to "critical" items either increases or decreases along that range, depending on where one found oneself on that function (Benjamin, 2001).

This method thus has three major advantages. First, we minimize the risk of individual difference variables colluding in such a way so as to restrict our measurements to a range in which effects are not easily detected. Second, when we reparameterize our accuracy data as the terms of the function that we fit them to, we hopefully increase the reliability and validity of our data. I say "hopefully" because such an outcome depends critically on the correctness of the function that we choose to summarize our data. The question of how to evaluate

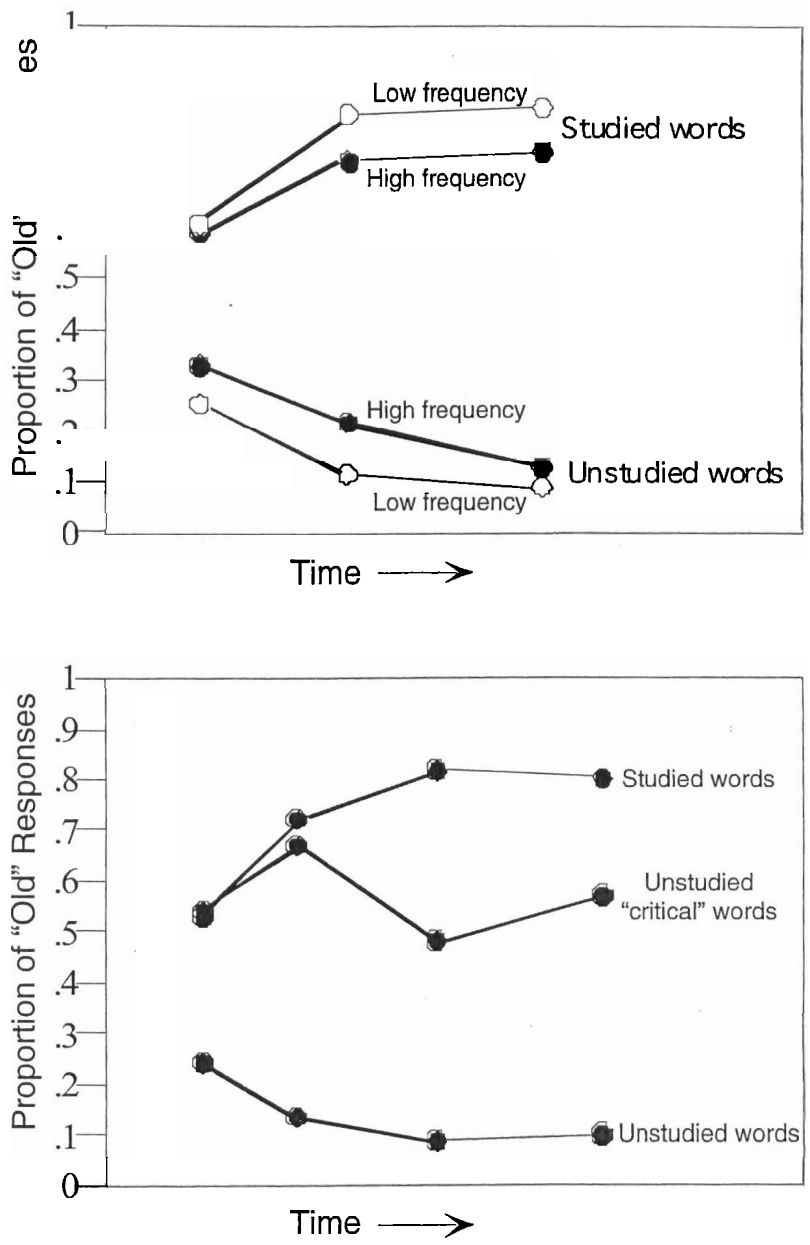


FIGURE 24.2. *Top panel:* Proportion of endorsements to old and new high- and low-frequency words across a range of decision times. *Bottom panel:* Proportion of endorsements to old, new, and new "critical" words across a range of decision times.

the correctness of a model is addressed in the next major section of this chapter. A final advantage is that the derived functions allow us to evaluate hypotheses that would be unaddressable were we to deal with single data points, for example, questions about the rate of information accrual.

Response Bias in Recognition Memory

In the previous example, I portrayed decision time as a potential individual-difference variable influencing recognition performance. Similarly, individuals can differ in the amount of evidence they demand before making a positive recognition

response. If a test word is only somewhat familiar, how is that uncertainty translated into a response? Clearly, different people bring different evidential standards to the table, and aspects of our experimental situation also influence how subjects make their decisions. Subjects might want, for example, to maximize the proportion of correct responses to old items—thinking that such a measure more validly reflects memory ability—and thus set a low recognition criterion: If a test item looks even vaguely familiar, they choose to endorse it. This somewhat arbitrary choice can influence our results: In the top part of Figure 24.3 are hypothetical group means, again corresponding to performance as a function of some manipulation of learning. Here the comparison of conditions is complicated by large differences in the overall "agreeability" of our subjects: Subjects in the left condition say "yes" more often than does the other group—to both old *and* new items. This fact reveals that our manipulation affected the decision strategies associated with recognition, but it is unclear whether it also influences memorability. To answer this question, we need to implement an experimental strategy similar to the one discussed earlier and gain experimental control over response criterion placement.

The lower part of Figure 24.3 shows performance across a wide range of response biases, plotted on axes corresponding to hit rate and false-alarm rate, yielding a receiver-operating characteristic (ROC). Such data can be elicited by, for example, having subjects complete multiple recognition tests under different payoff conditions. More commonly, subjects are asked to indicate a degree of subjective confidence along with the recognition decision; performance is then plotted as a cumulative function of the hit rate and false-alarm rate at a given confidence level and below. This technique allows for the construction of a ROC from two related but fundamentally different measures: the yes/no recognition response and subjective confidence.

In such a display, differences between subjects or between conditions that reflect differences in criterion setting for the decision component of the recognition judgment are virtually eliminated, and regularities in the form of the ROC are evident. In our example, we can see that the dots, corresponding to the data in the top half of the figure, lie on an isodiscriminability curve. In other words, no differences in memorability are apparent. Yet we could only reach this conclusion by uniting multiple measures and constructing an ROC that fits the data points. Different tasks yield different functional forms, and qualities of the ROC can be directly tied to psychological parameters, given a well-specified theory of the recognition decision.

For example, the Theory of Signal Detection (TSD), which has evolved into a theory of recognition (Banks, 1970; Egan, 1975; Lockhart & Murdock, 1970) by virtue of analogy with problems of discrimination in psychophysics (Green & Swets, 1966) and engineering (Peterson, Birdsall, & Fox, 1954) suggests that all stimuli—studied and unstudied—elicit some degree of mnemonic evidence, and the task for the subject is to set a decision criterion at some point on the spectrum of potential evidence values.

Certain versions of this theory posit that the probability distributions for evidence are Gaussian in form. This theory has implications for the form of the ROC. Specifically, underlying Gaussian probability distributions imply that a plot of the ROC on binormal axes should yield a straight line. More formally,

$$Z(HR) = \frac{1}{\delta_S} Z(FAR) + \frac{\mu_S}{\delta_S} \qquad (2)$$

in which δ_S represents the variability of the evidence distribution for studied items, and μ_S represents its mean. This function is superimposed on the two conditions in Figure 24.3 (on probability axes). Distributions of equal variance thus imply that that line should have unit slope. Figure 24.4 shows actual ROC and zROC functions from a representative experiment on recognition memory. The similarities among the *Z*-transformed functions are striking: they do indeed appear to be linear and have a slope of ~0.8 (Ratcliff, Sheu, & Gronlund, 1992). These functions thus reveal that the underlying probability distributions may well be normal, but they are apparently not of equal variance. This particular result suggests that the variance of the

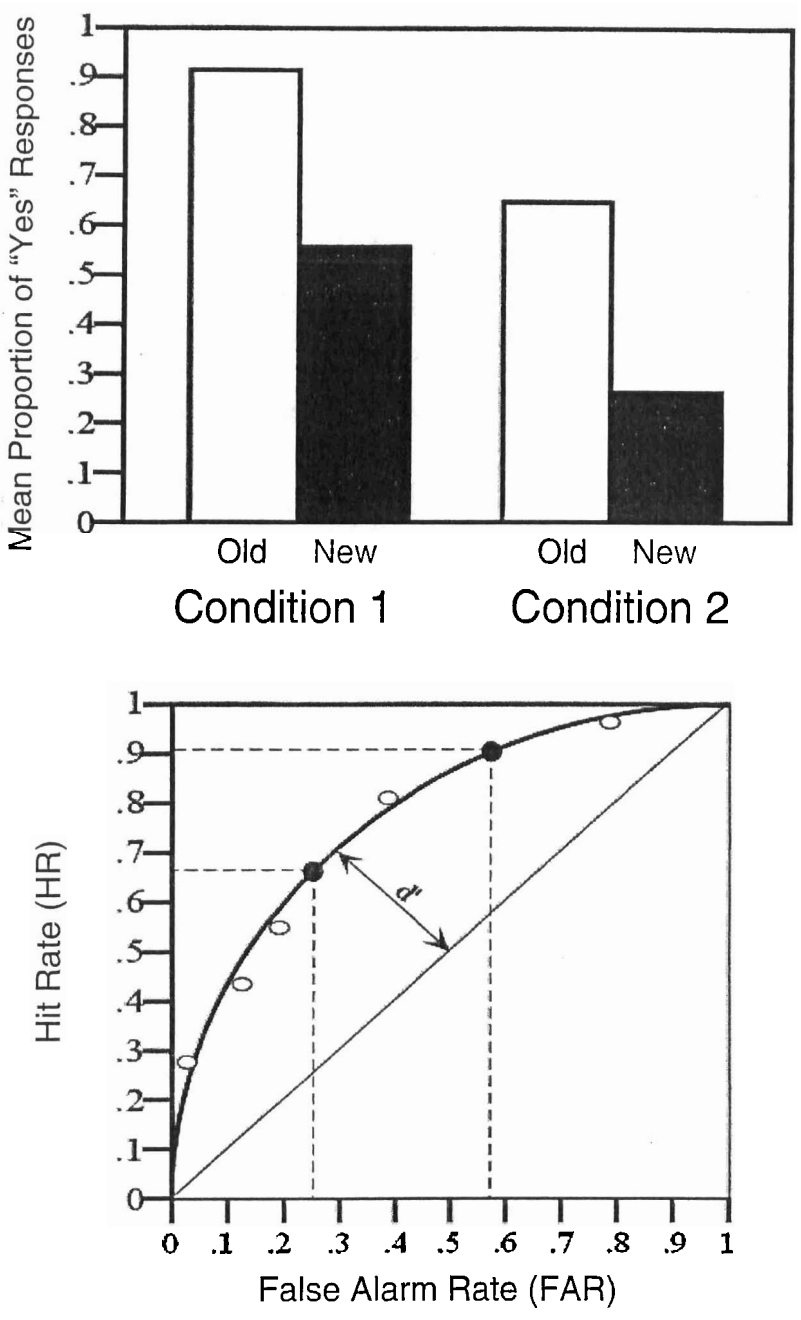
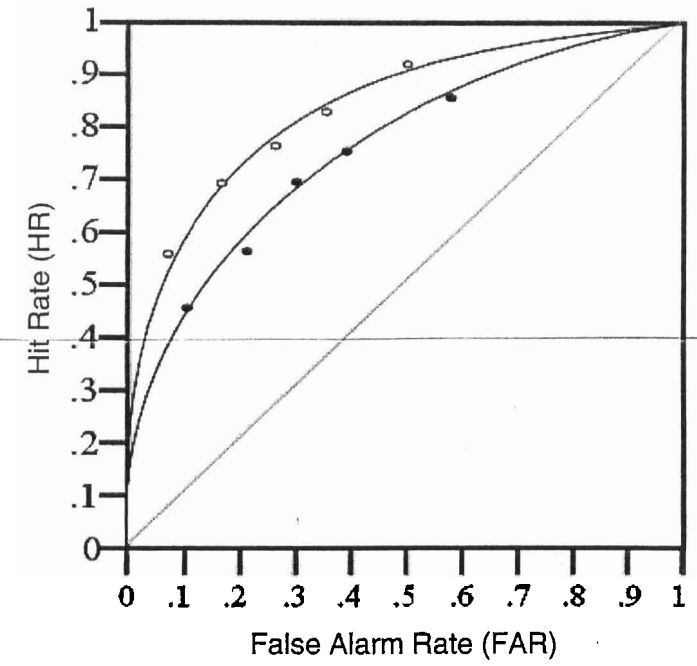


FIGURE 24.3. Hit rates and false alarm rates from two hypothetical conditions (*top panel*); hit rates and false alarm rates coplotted across a range of response criteria, as a receiver-operating characteristic (ROC; *bottom panel*). d′ indicates the discriminability of studied and unstudied stimuli.

distribution of evidence for studied items is approximately 1.25 times larger than for the distribution for unstudied items.

The form of ROC curves has also been brought to bear on the question that we introduced earlier, namely, what processes underlie the mirror effect in recognition? Consider the relationship between word frequency and recognition, as discussed in the previous section. The evidence from speed–accuracy trade-off functions was equivocal as to the question of whether a slow-acting deliberative process combines with general memory familiarity to produce



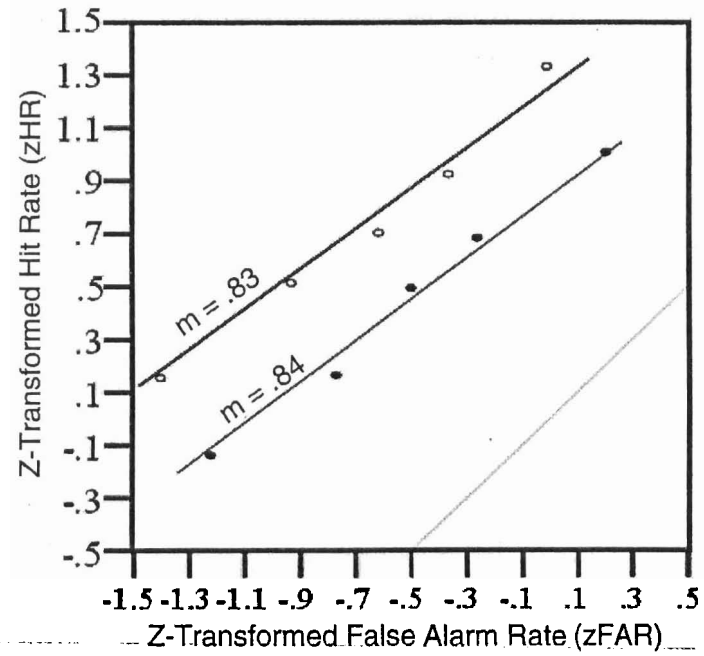


FIGURE 24.4. ROC and normalized ROC (zROC) functions from an experiment on recognition memory. The slope of the line is indicated by *m*.

the empirical dissociation seen between hit rate and false-alarm rate as a function of word frequency. In the preceding case, the argument concerned whether subjects made a postmnemonic assessment of the normative familiarity of the stimulus, thereby deriving a value against which to compare the actual experienced familiarity of the word.

Another argument is that two different processes can contribute to the endorsement of an item on a recognition test. The first is the same as that por-

trayed in the earlier argument: Stimuli enjoy some temporary boost in familiarity as a function of exposure, and this familiarity value provides some evidence of the recency or probability of past encounters with this word. Notably, however, the familiarity itself conveys nothing about the specific nature of the previous experience, so it can lead to spurious false alarms to other recently exposed but contraindicated stimuli (Jacoby, 1999) or even to unstudied stimuli that are systematically related to studied materials (Roediger & McDermott, 1995).

Familiarity is hypothesized to be augmented by an additional process, often called *recollection*, that serves to retrieve specific aspects of the prior encounter with the stimulus. One might recollect that a word was presented in italic typeface, or that a recommendation regarding life insurance came from a particularly disreputable agent, or that an author's name is familiar only because of a well-publicized tawdry scandal. Obviously, the details of a recollective experience can alter the way in which we engage a stimulus: We might choose to interact differently with a well-respected member of our field than with a convicted felon. With respect to word frequency, it has been suggested that the advantage that studied low-frequency words enjoy owes to a greater rate of recollection for such words, and that the lower false-alarm rate for unstudied low-frequency items reflects lower baseline familiarity (Reder et al., 2000).

Whereas familiarity is presumed to reflect a continuum of mnemonic evidence, recollection is typically thought to be a finite-state process. That is, recollected evidence directly implicates a specific prior experience as the locus of familiarity for an item, and that evidence specifies conclusively the status of the stimulus in question: It was experienced in the appropriate, sought-after context, or it was not. This process is finite-state in the sense that the evidence either promotes or discourages a response, with no degrees in intervening uncertainty. Finite-state models imply psychological thresholds: There is a point (or multiple points) at which there is an abrupt transition from "no evidence" to "evidence." This stands in contrast to the evidence continuum that familiarity provides, in which no amount of familiarity perfectly implicates prior study; similarly, a complete absence of familiarity does not unequivocally imply the lack of prior exposure.

Unlike the ROC functions described for Gaussian-based evidence distributions, thresholds do not imply ROCs that intersect the origin and the point (1, 1) in probability space, nor are they necessarily linear in binormal space throughout the function. Thus, departures from linearity in the form of the zROC can be taken as evidence for the contribution of threshold-based evidence to the recognition decision.

To use this logic to address the question of how familiarity and recollection contribute to recognition, and how they can be related to the word-frequency mirror effect, Arndt and Reder (2002) estimated ROCs for the recognition of low- and high-frequency words under special conditions designed to promote the use of recollection-based recognition. Under these conditions, subjects were asked to discriminate between studied items and the plurality-reversed complements of previously studied items. Researchers have presumed that a plurality-reversed distractor should elicit approximately equal familiarity to that of the original studied item, thus leaving recollection as the only basis for correct discrimination (Hintzman & Curran, 1994; Hintzman, Curran, & Oppy, 1992). In contrast to the standard ROCs elicited by recognition, as described earlier, ROCs elicited from this task are nonlinear in Gaussian coordinates (Rotello, Macmillan, & Van Tassel, 2000) as are ROCs from other tasks thought to emphasize the contribution of recollection (Yonelinas, 1997, 1999).

In comparing these functions for high- and low-frequency words, Arndt and Reder (2002) reported nonlinear zROCs for plurality-reversed recognition and linear zROCs for standard recognition, thus replicating prior findings. More importantly, the low-frequency zROC was more convex than the high-frequency zROC, a result that suggested that a threshold recollection process played a larger role in low-frequency item recognition then in high-frequency item recognition, consistent with the interpretation of Reder et al. (2000).

More generally, it is important to note that ROC functions can be derived from theories that cannot predict raw hit rates or false-alarm rates. Thus,

only by combining the two and generalizing across different levels of decision bias can such functions be derived. I hope to have shown here that the evaluation and comparison of such functions is central to progress in understanding recognition memory.

Memory Inclusion and Exclusion

For our final example of how the combination of multiple measures can inspire theoretical advances that would otherwise be purely speculative, consider the general problem of how to purify a measure of memory so that our assessment is minimally confounded by factors that look like remembering, but are in fact simply nondeliberative influences of memory. For example, consider a memory experiment in which subjects learn semantically or associatively related pairs of words such as *bread–butter* or *wishing–well*. If we test later memory by presenting the first term of each pair and attempting to elicit the second (*bread–?*), it is an impossible task to discern whether a response of *butter* reveals mnemonic retrieval of the previous study episode or simply temporary enhanced access to that word by virtue of automatic effects and influences of memory. Even more dastardly, the response might indicate nothing more than the prelearned nature of the association—through a lapse in attention or perhaps strategic yawning, the subject may have never even seen the study pair. How can we tease out the deliberative recollective aspect of memory in such a data set?

Jacoby (1991) provided a clever solution to this problem that involves the use of multiple measures. In his experiments, subjects provided their responses under two different conditions. The first replicated the typical memory experiment, in which they were told simply to remember the target word if possible and report it. In the other condition, subjects were told explicitly to produce any word *except* the target word. The combination of these conditions allowed Jacoby (1991) to specify a theory of how deliberate and automatic influences of memory interact to produce responses in this type of cued recall paradigm. He claimed that, in the standard (henceforth, *inclusion*) condition, a response that matched the prior study item could reflect either form of memory and assumed that their contributions were independent of one another:

$$p(target|inclusion) = R + A - RA. \quad (3)$$

Here R indicates the probability of correct recollection of the study episode, and A indicates the probability of automatic nonrecollective influences leading to a correct response. In the condition in which subjects are told not to produce the previously studied pair word, the sources combine differently:

$$p(target|exclusion) = (1 - R)A. \quad (4)$$

That is, if the target word were to be recollected, it would not be produced. Thus, a target response in this condition indicates a lack of such recollection. Under such conditions, the target might nonetheless be produced if automatic influences of memory lead that word to be particularly accessible. The difference between performance in these two conditions is thus equal to R and provides a model-based estimate of the recollective memory contribution to performance in the task. Given this estimate, it is easy to derive the estimate for the parameter A, which reflects the automatic nonrecollective memory influence on the task.

In one striking example of how the combination of inclusion and exclusion memory tasks yields results that would otherwise be unobtainable, consider an experiment reported by Jacoby, Toth, and Yonelinas (1993). Subjects were exposed to two lists of words, the first of which subjects were told to remember and was presented aurally. The second list was presented visually, and subjects were told to read the words aloud. During this second list, some subjects performed an additional attention-dividing task and others did not. The final recall test consisted of presenting word stems (e.g., mer—) and, in the inclusion condition, asking subjects to recall a word from either list that completed that cue; in the exclusion condition, they were instructed to specifically avoid completing the cue with a word that had been presented in either earlier study list. Table 24.1 shows the raw data for the inclusion and

TABLE 24.1

Raw Performance and Model Estimates for Previously Read Words on Tests of Recall Inclusion and Recall Exclusion as a Function of Attentional Condition

	Raw performance		Model estimates	
Attention	**Inclusion**	**Exclusion**	***R***	***A***
Full	0.61	0.36	0.25	0.47
Divided	0.46	0.46	0.00	0.46

Note. *R* is an estimate of the contribution of recollection to performance and *A* is an estimate of the automatic contribution of memory to performance.

exclusion of words that were presented in the visually presented (second) list as a function of the attention manipulation. It also shows the values of *R* and *A*, as reparameterized by Equations (3) and (4). Evident in those parameters is a very clear effect of attention on *R* but not *A*. It is from such results that we can conclude that the automatic effects of memory are relatively impervious to manipulations of attention, but that the deliberative, conscious contribution of recollection is not.

To once again sound the drum that is the theme of this volume, certain conclusions are made possible only by the theoretically motivated combination of multiple measures. Multimethod psychology refers to more than convergent and divergent validity; in each of the examples outlined here, studying individuals under different conditions or in different situations afforded a rich, multifaceted view of their behavior. Just as psychologists include multiple subjects in experiments to be able to generalize across individual differences and to examine effects owing to those differences, multiple methods or experimental circumstances allow the researcher to tease out effects that underlie differences between conditions (as in the final example given earlier) and additionally reduce the risk of being led astray by single oddball conditions that don't generalize well to the naturalistic circumstances that they are intended to simulate.

ASSESSING THE ADEQUACY OF FORMAL MODELS OF COGNITION

In each of the examples outlined in the previous section, I have attempted to illustrate how the theoretical gain obtained from the combination of multiple measures was greater than the sum of the parts (the individual measures). Lurking within this apparently free lunch is a cost, however. In each case, we needed to specify a theory about the relationships among our measures before we could combine them. The cost of combining measures is measured in the assumptions that we make in specifying that theory. In particular, if our theory is wrong, the parameters that we derive from its application may be meaningless or even misleading.

In addition, more accurate theories are often derived from a careful evaluation of the specific points at which prior attempts fell short. Thus, it is critically important to subject such theories to evaluation and cull the herd appropriately. This section briefly reviews recent advances in and discussions of our understanding of how such evaluations can be conducted.

Probably the most common application of model testing involves the logic of goodness-of-fit statistical tests. Such tests assess the extent to which a specified model can handle a particular set of data. One familiar application of such a procedure involves the comparison of obtained frequencies of events to a set of predicted frequencies. The predictions come from a model that can make any number of assumptions about the relationships between the event types to one another (often, that they are independent). The sum of squared differences between the expected and obtained frequencies is the building block for a test statistic that can be compared to an appropriate chi-square distribution.

A more complex model's ability to account for a pattern of data can be summarized with a similar measure, such as Root Mean Squared Error or Percent Variance Accounted For. Such measures provide a good basis for ruling out a model: If no

combination of parameters within a model can allow it to predict a result that is commonly obtained, then something about that model is clearly wrong. To draw on an earlier discussion, if z-transformed ROC functions for recognition memory were typically curvilinear, then we would want to reconsider the assumption that the evidence distributions are Gaussian in form.

Unfortunately, unlike theories in physics, psychological theories are typically quite flexible—so much so, in fact, that there is probably a greater utility in using tools that rule out models not on what they fail to predict, but rather how much they can predict for which there is no evidence (Roberts & Pashler, 2000). If our theory of the form of the zROC was so general that it could not rule out any functional structure, we should be considerably less impressed by its ability to account for the correct linear form. Thus, more appropriate model-testing mechanisms emphasize not only the ability of the model to account for a pattern of data, but also its ability to do so simply, efficiently, and without undue flexibility. These mechanisms deal with such concerns by incorporating factors such as the number of free parameters (Akaike Information Criterion [Akaike, 1973]; Bayesian Information Criterion [Schwartz, 1978]) or even the number of free parameters and the range of function forms that the model can take (Bayesian Model Selection [Kass & Raftery, 1995]; Minimum Description Length [Hansen & Yu, 2001]). These approaches have clear advantages over simple goodness-of-fit tests, on which more complex models have an inherent fitting advantage simply by virtue of their ability to overfit data that in psychological experiments typically include a large amount of sampling error (Pitt & Myung, 2002).

What Makes Theory Useful?

So far, this discussion has emphasized accuracy and flexibility as the principal bases for model evaluation. We want our theories to predict events that happen and not to predict things that don't; if our theory does so with a reasonable degree of success, then we covet it and attempt to defend it against outside claims of inadequacy.

I want to propose a slight amendment to such a system, however. I believe that models can also be tremendously useful when they fail to provide an account for certain data. Models—particularly well-specified mathematical ones—are useful in part because they are putative isomorphisms for the system under investigation. Consider, for example, the question of how to compare the weights of objects. Masses of objects can only be directly compared with an accurate balance. Yet if I want to know whether this APA-produced tome outweighs other recent books in this domain, I don't need to truck my library over to a chemistry lab to use their balancing scales. Rather, the mass of each object is represented as a real number, and I know that the set of ordinal operators in mathematics (including > and <) correspond to "weighing more than" and "weighing less than." To return from this tortured analogy back to the original diatribe, models are useful in part because they provide a different representational system with which to talk about the components of the theory. As discussed early in this chapter, cognitive components are notably vague; grounding a theory in a more formal representational system, such as mathematics, allows us to use the sophistication of that system to derive relationships beyond what our intuitions would have provided us with—even when that formal system is not a fully accurate representation.

One excellent example of how model accuracy and model utility occasionally diverge is provided by the Rescorla-Wagner model of learning (e.g., Rescorla & Wagner, 1972). That theory was itself an attempt to address shortcomings of previous views of associative learning that postulated that contingency of events in time and space was a sufficient (and necessary) precondition for the learning of an association between the events (e.g., Bush & Mosteller, 1951). A number of important results were obtained in the late 1960s that demonstrated the inadequacy of this view by demonstrating conditions in which animals apparently did not learn an association between two stimuli despite highly contingent presentations of the stimuli. One illustrative and fundamental phenomenon is that of *blocking*, in which an organism first learns that A predicts B and later that the compound AC also

predicts B. Blocking is revealed by the fact that the organism does not engage in typical behaviors preparatory for the onset of B when exposed to C alone (Kamin, 1969). The Rescorla-Wagner model explains this result by assuming that an organism learns about the relationships between events only to the degree that outcomes are unpredictable: When an event is expected on the basis of alternative cues (e.g., A predicts B), then nothing is learned about the relationship between additional cues and that outcome (e.g., C and B). Formally, the model can be stated in a reduced form as

$$\Delta A_i = \beta(\lambda - \sum A_i) \quad (5)$$

in which ΔA_i represents the change in the strength of the learned association between two stimuli on Trial i, β represents a learning parameter related to the intensity and associability of the two stimuli, λ represents an asymptotic learning parameter related to the outcome event, and most importantly, ΣA_i represents the summed associative strength between all available stimuli and the outcome event in question. When this value is close to λ, the term inside the parentheses approaches 0; thus learning is weak or nil.

It would be no exaggeration to state that this model has been the single most influential theory of learning since its publication. It has been imported into (or coevolved with) many other domains, including human contingency learning and causality judgments (e.g., Chapman & Robbins, 1990; cf. Cheng, 1997) and artificial learning in neural networks (as the influential delta rule; Rumelhart, Hinton, & Williams, 1986; Widrow & Hoff, 1960). It can account for a huge number of basic phenomena in associative learning (Dickenson & Macintosh, 1978; Miller, Barnet, & Grahame, 1995; Walkenbach & Haddid, 1980) and consequently has been the primary vehicle for the discussion of phenomena in animal learning in introductory textbooks.

These successes notwithstanding, there are numerous examples of how the model fails to account for behavior in the very paradigms it was designed for. To draw again on the example of blocking, as described earlier, remember that the model explains blocking as a deficit in learning—the animal fails to respond to the blocked stimulus because nothing was learned about the relationship of that stimulus to the outcome. Certain phenomena indicate that this assumption is almost certainly false. For example, additional training following the traditional blocking procedure that presents the blocking stimulus (A, in the preceding example) paired with the *absence* of the outcome stimulus (C) can lead to *retroactive unblocking*, in which responding increases to the B stimulus, even though there were no additional presentations of that B stimulus (Arcediano, Escobar, & Matute, 2001; Blaisdell, Gunther, & Miller, 1999).

From a model-evaluation perspective, such data should lead us to cast out the Rescorla-Wagner Model as outdated and unsatisfactory. However, this approach ignores critical aspects of the scientific process; namely, the discovery of phenomena like retroactive unblocking was motivated in large part by the strong (and ultimately incorrect) predictions of the model. In other words, widespread understanding of the model led researchers to devise paradigms that tested its limits. In addition, certain generalities among the phenomena that contradict the model are only apparent in context of how the model deals with them inadequately (Miller et al., 1995). Thus we see that models serve not only as isomorphisms for the systems we study, but also as motivating and organizational tools that enhance our progress toward understanding the mechanisms they purport to represent—even when they do so incorrectly. This approach to model-based psychological science is well reflected in the quip that models should be your friends, not your lovers (Dell, 2004). You maintain them because of what they offer you, but you keep many of them and don't demand too much of any single one.

INTEGRATING COGNITION AND COGNITIVE NEUROSCIENCE

So far we have limited our discussion to (a) how the field of human memory has evolved because of the integration of multiple behavioral measurements, and (b) how the models that serve that function should be evaluated. Here I briefly confront the question of how to integrate behavioral measures with the types of data provided by

research in cognitive neuroscience. Let me warn the excitable reader that I offer no good answers to this question. I am not alone in that regard, but I do offer a few suggestions that might help guide future advances on this front.

In particular, advances in medical imaging have brought to the forefront questions about the integration of physiological data into cognitive theorizing. The issues themselves are quite old, in fact; researchers have used the electroencephalogram (EEG) and galvanic skin response (GSR) to address cognitive-like issues for about a century (Berger, 1929; Féré, 1888; Tarchanoff, 1890). The advances alluded to refer primarily to measures that allow greater spatial precision in viewing the morphological structure of the brain, as well as the transient electrical, chemical, and hemodynamic events that occur during brain function. These techniques—both the new and the old—allow the construction of spatial and temporal maps of activity during the performance of different cognitive tasks. One tack to integrating cognition and neuroscience is a primarily *exploratory* approach. Using cognitive theory to compare tasks that differ in a single putative cognitive component, either parametrically or otherwise, allows the inspired cognitive neuroscientist to compare maps of brain activity and postulate a brain region or regions that are related to the manipulated cognitive component.

Hidden within this approach is the notion that the brain is likely to have divvied up cognitive functions in the same manner as experimental psychologists have. I fear that we have not had that kind of insight, but the approach is valuable nonetheless, for it allows for the evolution of cognitive neuroscience into a second, more mature phase of theoretical development. Using a *hypothesis-testing* approach, specific neural signatures known to accompany cognitive events are sought in paradigms in which there is theoretical debate about the contribution of those cognitive components to the behavior in question. For example, changes in blood flow are apparent in areas in Broca's area 17 during mental imagery (Le Bihan et al., 1993). In addition, "small" mental images elicit greater activation in posterior visual cortex, corresponding to foveal input, whereas "large" mental images elicited greater activation in anterior visual cortex, an area that represents input from the periphery of the eye (Kosslyn et al., 1993). In each of these cases, the researchers used established knowledge about brain function—in this case, that regions of occipital cortex code visual input from the eye—to address the question of whether visual imagery is spatial or propositional in format (Finke, 1980). The evidence revealed that imagery engaged visual areas of the brain and is thus likely spatial in representation. Other recent research has used this approach to address whether people learned an association between visual and auditory stimuli by examining blood flow in visual cortex following presentation of an auditory stimulus that had previously been paired with a visual stimulus (McIntosh, Cabeza, & Lobaugh, 1998). Many other examples exist in the domains of perception, attention, memory, and language.

As results from exploratory cognitive neuroscience increase the number (and validity) of known relationships between neural signatures and cognitive components, the more scientists interested in cognitive phenomena will be able to exploit that knowledge for the purpose of furthering cognitive theory. The back-and-forth between exploratory and hypothesis-testing approaches illustrates one way by which to integrate measures from the two domains. But it is worth noting that the distinction between brain-based and behavioral measures is at least partly artificial. If we measure a button press or a verbal output from a subject, we consider that measure behavioral. Yet at multiple physiological levels, events occur during that press or vocalization that are unique to that output. Muscular events in the arm or larynx, as well as neuronal events in motor cortex, control those very actions that we measure behaviorally. Other neural events combine to derive that pattern of efferent control given the input from sensory organs. No matter what the task, a continuum of events guides the physical input (in the form of light or sound waves, for example) into physical representations in the brain into physical output (in the form of muscular contractions). Whether we

measure those behavioral endpoints or the physical events that precede and determine them—inside or outside the brain—the logic for the combination of multiple measurements remains the same.

The endpoints of this continuum will always be critical measures, however, no matter how precise our measurements of the intervening processes become. Just as it would be impossible to draw any meaningful conclusions about psychology without knowing anything about the physical stimulation to the subject, it is also quite difficult to do so without actually examining behavior. Many behavioral measurements carry with them an inherent dimension of performance quality that other intervening measures do not. If a manipulation enhances the speed or accuracy with which subjects perform a task, we are licensed to attribute to that manipulation an interpretation of quality—that it improves learning, or problem-solving speed, or attentional focus, for example. There is nothing inherently "better" from a cognitive perspective about more blood flow to a particular brain region, greater skin conductance, or higher levels of chemical uptake, even though such effects may well accompany behavioral effects that do allow such an interpretation.

On the other hand, experimental tasks often suffer from a failure to approximate real-world circumstances that elucidate the contribution of the cognitive capacity under study. In part, this may be because of the contrived nature of the chosen behavioral measure. Researchers interested in language comprehension, for example, often measure the rapidity with which subjects can identify probe stimuli as words or nonwords as an index of the degree to which previously read sentences or heard utterances (related to those words) have been comprehended. Clearly, this artificial task makes the laboratory study of language comprehension quite unlike naturalistic language comprehension. Cognitive neuroscience methods provide an opportunity to reduce the reliance on such tasks by allowing measurements in the absence of an overt behavioral task. For any given experimental situation, the choice between behavioral and brain-based measures involves trade-offs, and as the astute reader might suspect, the combination of multiple types of measures across and within single studies often proves the most fruitful approach.

EMPIRICAL EVIDENCE AND THE POSTULATION OF MENTAL SYSTEMS

Recall that we began this chapter with a series of pithy comments about the ways in which cognitive psychologists derive evidence for theoretical entities. That task begins with an analysis of empirical data and proceeds to a theoretical interpretation only through the lens of a particular model. Although we have not emphasized it here, it is important to remember that any comparison of conditions or measures assumes some underlying model, and that those comparisons that are simple do not necessarily reflect simplicity in that underlying model.

Through our short tales in the first section, we discussed the theoretical interpretations of model-based analysis only as necessary. In this section, I outline rules that other researchers have used to guide the relation between theoretical parameters and theoretical entities. Consider the final example from the first section, in which performance from multiple recall tasks was combined to yield estimates of the contribution of deliberative recollection (*R*) and automatic memory retrieval (*A*) to cued recall (Table 24.1). The manipulation of attention had opposite effects on inclusion and exclusion probability, which made the raw data difficult to interpret. However, the model parameters told a very clear story: Attention affects recollection, but not automatic memory. This dissociation provides a first step toward the postulation that these two bases for responding actually represent different memory systems or different memory processes. What else is necessary?

The primary basis for such postulation is the existence of converging multiple dissociations (Schacter & Tulving, 1994). The evidence that aging, for example, selectively impairs recollective but not automatic memory strengthens the case that the two are separate entities (e.g., Benjamin & Craik, 2001; Jacoby, 1999). In the context of animal

learning, Lorenz (1970) argued that imprinting was a fundamentally different process than that of normal learning and pointed to various dissociations between the two, such as the presence of a critical period for the former, but not the latter (cf. Shettleworth, 1993).

Tulving (1984) argued that memory systems should be distinguished in large part on the basis of the information they store and the operations they perform on that information. Thus, *procedural memory*, which governs the executions of actions and skilled performance, can be distinguished from *declarative memory*, which contains verbalizable knowledge. Procedural memory contains information about the rapid coordination of limb movements and thus maintains a unique information store. Declarative memory maintains information in sufficiently flexible form to allow inferential processes to act on propositions in memory and thus allows unique operations unavailable to procedural memory. These differences do indeed play out as a number of dissociations in both animals and humans (Squire, 1992).

In addition, Tulving (1984) suggested that memory systems be defined in part by their neural substrates. This is an important point, given the renaissance of cognitive neuroscience briefly remarked on earlier, and I wish to offer an alternative viewpoint as a final remark. The denouement of the argument is that there is no reason why brain systems and cognitive systems should be one and the same.

But do not all the functions of cognition lie in the brain, and therefore shouldn't the structure of the brain be a reasonable playground for the construction of cognitive theories? The answer is no, for the same reason that neither protein strings, nor molecules, nor atoms, nor quarks should be the building blocks of a cognitive theory. Theoretical entities in cognitive psychology are only useful insofar as they allow a handy categorization of experimental results. Thus, despite the fact that habituation in the eye and in the ear take place in different brain regions, we nonetheless recognize a unifying concept that unites the two forms of learning.

A trickier question, however, is whether we are justified in postulating multiple cognitive components that exist in a single brain region. Consider the granddaddy of all distinctions in human memory, that between episodic and semantic memory (Tulving, 1983). Episodic memory stores events from an autobiographical perspective; semantic memory stores facts and knowledge and contains no information about specific past episodes. This distinction has been among the most useful in modern memory research and makes sense out of a huge number of empirical phenomena. Yet, numerous influential theories propose that the information underlying these two memory "systems" is one and the same. For example, Hintzman (1986) showed that a memory system that stored nothing more than specific individual events—in other words, its memory was exclusively episodic—could yield behaviors that were hallmarks for the postulation of semantic memory. Does such a demonstration imply that the distinction is no longer useful? Of course not. Although it may well turn out the brain does not honor this distinction, there is no reason why a cognitive theory should not. Similarly, we can build a reasonable model out of integers and logic components of the way in which our desktop computer performs some computational task, despite the fact that the computer's own representation is binary, and its logic components are nothing more than the arrangements of binary operators. There is no doubt that knowledge about the structure and function of brain regions can and should inform cognitive notions about memory, but there is a danger is failing to recognize additional appropriate levels of abstraction beyond the physical substrate and inappropriately besmirching theories that have desirable qualities.

SUMMARY

In this chapter, I have provided several examples of how measurements can be combined via models to yield results that are more informative and reliable than the original measurements themselves. This technique must always be accompanied by rigorous model evaluation, lest the interpretation of the parameters be misled by incorrect assumptions about their relation to one another. These same techniques apply to measurements obtained from physiological properties of the brain; doing so will

allow the burgeoning field of cognitive neuroscience to accommodate more readily to the theories of cognitive psychology. Finally, model-based interpretations provide a particularly useful way of seeking dissociations that are the fundamental building blocks of cognitive systems. A dissociation may only become evident when the correct model is imposed on the data. These dissociations should not be taken to imply dissociations at the level of the brain, nor should different brain systems necessarily influence cognitive theorizing.