Everyday cognition in adulthood and late life

Edited by
LEONARD W. POON
University of Georgia
DAVID C. RUBIN
Duke University
BARBARA A. WILSON
University of Southampton

1989

*The right of the
University of Cambridge
to print and sell
all manner of books
was granted by
Henry VIII in 1534.
The University has printed
and published continuously
since 1584.*

CAMBRIDGE UNIVERSITY PRESS
Cambridge
New York Port Chester Melbourne Sydney

7 Issues of regularity and control: Confessions of a regularity freak

David C. Rubin

The laboratory offers a high degree of experimental control. That is why laboratories were devised in the first place, and that is why scientists often forsake real-world problems to enter them. This chapter is an argument that, given our present state of knowledge, control is more often a vice than a virtue. It is not experimental control that is now desirable, but rather regularity of results. The time for control will come, but psychology entered the laboratory too quickly. Psychology must first spend time observing and quantifying behavior in its fuller state of complexity. This chapter describes the rationale for this approach, followed by examples of research that has applied it to the topic of human memory.

The best way to start arguing for a flight from the laboratory in the area of memory research is to examine how we entered the laboratory in the first place. Ebbinghaus (1885/1964), in arguing that human memory could be studied in a scientific fashion, provided as clear a definition of the experimental method as exists in the psychological literature. In a section titled "The Method of Natural Science," he writes as follows:

> We all know of what this method consists: an attempt is made to keep constant the mass of conditions which have proven themselves causally connected with a certain result; one of these conditions is isolated from the rest and varied in a way that can be numerically described; then the accompanying change on the side of the effect is ascertained by measurement or computation. (Ebbinghaus, 1885/1964, p. 7)

> When, however, we have actually obtained in such manner the greatest possible constancy of conditions attainable by us, how are we to know whether this is sufficient for our purpose? When are the circumstances, which will certainly offer differences enough to keen observation, sufficiently constant? The answer may be made: When upon repetition of the experiment the results remain constant. (p. 12)

Notice that Ebbinghaus insisted that quantification and control are necessary for science. But how did he know if sufficient control had been obtained? Not by

Support for the preparation of this chapter was provided by NIA grant AG04278 and NSF grant BNS-8410124.

measuring how accurate the controls were, but by measuring their effects on the regularity of the observed behavior. Ebbinghaus looked for evidence of sufficient control in the results, not in the environment. I shall not argue with any of what Ebbinghaus said, except to note that he made one assumption that is almost always correct in the physical sciences and almost always incorrect in studying human behavior. He assumed that increasing control over the external environment increases the regularity of the behavior observed. It is much more common that the opposite is true – a counterintuitive claim that will soon be supported. This one minor flaw in a brilliant work has slowed progress in memory research and in psychology in general. Ebbinghaus did exactly what was needed to demonstrate to a skeptical world that a scientific study of memory was possible. The problem is that his success hid his erroneous assumption for so long.

Less control provides more regularity

The real world, not the laboratory, offers the best chance of observing regularity. For some behaviors and some initial levels of control, my counterintuitive claim must be wrong. However, when the average degree of control that has existed in the laboratory is considered, the claim is basically correct. When the environment is controlled in the laboratory, it tends to become simpler than the environment in which people normally operate. It tends to lack the kind of stimuli to which people normally respond. The simpler, or impoverished, environment of the laboratory is a stimulus that fails to exert stimulus control over behavior. The result is an increase in nonrepeated, seemingly random, behavior. This argument could be made in terms of using stimuli with evolved salience, but it can be made equally well on the basis of past learning. This argument asks for representative stimuli, as Brunswik (1955) or Petrinovich (Chapter 2, this volume) might, but not for reasons of representative sampling.

As a concrete example of the claim that more control yields less regularity, let us return to Ebbinghaus and the domain of verbal material. Consider a thought experiment in which two subjects randomly selected from a large introductory psychology course are both asked to learn and later to recall many lists, each 100 syllables long. Some of the lists are randomized nonsense syllables translated from German and are of known meaningfulness and pronunciability. Some of the lists contain 100 syllables of randomized nouns normed on dozens of different properties by a dull verbal-learning researcher (Rubin, 1980). The stimuli in the first two kinds of lists are presented at a rate of one every 3 sec. Some of the lists contain 100 syllables arranged in simple sentences, with the order of the sentences randomized. Each sentence is presented as a whole, and each subject signals, within set limits, when the next sentence should be presented. Some of the lists are stories 100 syllables long, and, finally, some of the lists are poems 100 syllables long. For the last two kinds of lists, the entire 100 syllables are presented at once, and thus the experimenter has no control over and no knowledge of the encoding time spent on each individual syllable. Presentation times are chosen for the different kinds of lists so that, on the average, each subject recalls 50 syllables from each kind of list.

Our task as psychologists is to predict the degree of agreement between the two subjects. In particular, we are to predict the type of list on which the two subjects will most often recall the same syllables. If we choose, we may also predict the type of list on which the two subjects will most often agree on the order of recall of those 50 syllables, the latency between the recall of particular syllables, or almost any other dependent measure. Agreement between the two subjects is a measure of regularity, of how well the situation controls behavior. In a thought experiment with more than two subjects, the larger the agreement between randomly selected pairs of subjects, the lower the error variance. The lists that are the simplest, most controlled, and least affected by the idiosyncrasies of the past experiences of the individual subjects are the nonsense syllables. The lists that are the most complex, least controlled, and most affected by the idiosyncrasies of the past experiences of the individual subjects are the poems. Would anyone choose the nonsense syllables?

As the lists become more complex, they become more structured, but it probably is not the degree of the structure alone that determines the ability of the stimuli to control behavior. Rich past histories, genetic or environmental, tend to increase a stimulus's control of behavior and make responses more similar among individuals, even when the degrees of formal structure among classes of stimuli do not clearly differ.

The thought experiment just described was intended to put in concrete terms the argument that more experimental control often leads to less regularity. Later we shall consider actual research in human memory in which minimal control produced exceedingly regular behavior.

More control provides less knowledge

Regularity is all we really need at our present state of theoretical advancement in most memory research. Control added beyond that necessary to gain repeatable results is undesirable. In addition to yielding decreased regularity, an increase in control has the effects of hiding unexpected results from the researcher and of decreasing the extent to which the results can be generalized. A well-controlled experiment allows the effects of only the independent variables in the design to be observed. Important but unexpected factors have little chance of being discovered if a controlled experiment is designed properly; they simply enter into the error variance. In addition, the predictions about the dependent variable typically are about changes in magnitude across conditions, and so the dependent variable usually is too simple to provide a sufficiently complete description of behavior to facilitate the formation of post hoc hypotheses (Rubin, 1985). "Sloppier" experiments, with less control of the experimental conditions and less control of the responses the subjects are allowed to make, provide greater chances for unexpected factors to be noticed. Moreover, if unexpected factors are not observed, we can have greater confidence that, if present at all, they will have only small effects; see Mook (Chapter 3, this volume) for a counterargument. Similarly, if regularity is sought, but controls are kept to a minimum, then the results of the observation can be generalized at least to the extent to which the conditions of observation varied. For example, in a sloppy study of memory, if the retention interval and the degree of initial learning are allowed to vary across subjects, and if subjects all produce similar behaviors, then we can generalize over variations in retention interval and degree of initial learning. In short, given two studies, if extraneous factors affect the result of interest, then we are more likely to be able to document those factors with the more open-ended, sloppier study, and if the two studies turn out to have equally regular results, then the results of the sloppier study can be generalized to a wider range of situations.

More control requires more knowledge

It has been claimed that regularity, not control, is what is needed at our current state of advancement. But certainly there must be some conditions under which control is the main objective and different conditions under which regularity is the main objective. Regularity should be the goal in a science that does not have theories to predict specific outcomes from experiments. Of course, all scientific inquiry must have some guiding hypotheses, but unless those hypotheses are sufficiently well developed to make predictions that other competing assumptions will not, there is little reason to test them. Rather, what we need are repeatable phenomena at a level of complexity that provides constraints for the formulation of broad classes of theories. Control should be the goal in a science in which theories have been developed that warrant serious testing. For such theories, the proper controls offer elegant tests. In particular, in situations where little control is applied, different theories often make the same predictions, but with more control these same theories can be forced to make different predictions. Whenever we have enough knowledge to form and test such competing theories, control is to be preferred. Unfortunately, such theories are still rare in cognitive research.

An example from developmental psychology may be helpful. General hypotheses led Piaget to discover and document the phenomenon of conservation in the child. Once this phenomenon was shown to be robust, competing theories could be developed and tested. The first step was to demonstrate a repeatable phenomenon that constrained possible theories and that allowed for experimental manipulation. Only after that was accomplished could competing theories be formulated. Progress would have been much slower if complex theories had preceded observation. Numerous analysis-of-variance experiments could have been run, deciding among instances of a class of theories, all of which were flawed.

Some subtle and some obvious implications accompany the goal of finding regularity and the goal of maximizing experimental control. When control is sought, the actual results obtained often are not intended to be generalized. Very specific conditions are obtained for a test of a theory, and these conditions are the only ones for which the interpretation of the results is important (Mook, Chapter 3, this volume). Changing one of the conditions slightly might alter the results completely, but if this condition is not of interest to the theories being tested, or if the theories predict such a change, the change is not of consequence. Unless a strong theory exists to state which of many possible changes in conditions should have large effects, however, such large changes in results with minor changes in conditions will reduce the accumulation of knowledge to chaos, as Petrinovich (Chapter 2, this volume) points out.

The role of theories also changes as one goes from emphasizing control to emphasizing regularity. When control is the major goal, a single theory, or a set of theories, determines what factors to control and manipulate and what questions experiments should be designed to answer. When regularity is the major goal, theories take a more subservient role in relation to the data and are used mostly to guide research in general terms and to interpret the observed data.

Regularity is proving the null hypothesis

The emphasis on regularity, rather than control, changes the role of inferential statistics in experimental inquiry. Current inferential statistical methods in psychology bias against the search for regularity. Inferential statistics note differences, not regularities. Nonetheless, understanding behavior often involves noting regularities, not differences. To note regularities usually is to accept the null hypotheses.

The statistical issue whether or not empirical support can be gained from accepting the statistical null hypothesis is a classic issue that has received considerable attention. In the sixties, it was the subject of a lengthy series of exchanges in the *Psychological Review* and *Psychological Bulletin* (Binder, 1963; Edwards, 1965; Grant, 1962; Wilson, Miller, & Lower, 1967). More recently, Greenwald (1975) has developed a model of how prejudice against gaining support from accepting the null hypothesis produces detrimental effects on the advancement of science. Thus, the view that accepting the null hypothesis cannot provide evidence for an empirical hypothesis is not held by all schools of statistics or by all psychologists. Two issues will be given as examples.

Those who argue that accepting the statistical null hypothesis can offer no support for a theory point to the fact that it is easy to do an experiment with so little power that the null hypothesis will be accepted even though a real effect is present. Those on the other side of the issue argue that because the means of two groups will almost always differ by some small, perhaps infinitesimal amount, it is always possible to reject the null hypothesis by increasing the power of the experiment. "Putting it crudely, if you have enough cases and your measures are not totally unreliable, the null hypothesis will always be falsified, *regardless of the truth of the substantive theory*" (Meehl, 1978, p. 822). There is a parallel between the two approaches. The power of an experiment must be chosen appropriately no matter whether one is seeking to gain support by rejecting or accepting the null hypothesis. Accepting the null hypothesis does not indicate that two means are identical, only that they are equal within the power of the experiment to resolve them.

The second criticism against gaining support from accepting the null hypothesis is that the null hypothesis can be confirmed because a sloppy experiment was performed, introducing random error. The counter to this is that the null hypothesis can be rejected because a sloppy experiment was performed, introducing nonrandom error. Again, there is a parallel. The quality of an experiment must be checked no matter whether the null hypothesis is to be accepted or rejected. Experimenter error can lead to the null hypothesis being either falsely accepted or rejected.

The preceding arguments were not intended to overcome all the biases present in psychology today against gaining support from accepting the null hypothesis (Greenwald, 1975). It can be argued that there are some reasons to prefer rejecting the null hypothesis (Wilson et al., 1967). All that should be made clear is that gaining support from accepting the null hypothesis is not prohibited by most statistical theory and that it may be the proper way of couching certain questions. Conceptually, support for regularity can be gained most directly through inferential statistics by acceptance of the null hypothesis, because usually it is a claim of no noticeable difference, rather than a claim of greater similarity than would be expected by chance.

The resistance of some to gaining support for a theory by accepting the null hypothesis may just be an aspect of the near monopoly held by inferential statistics in the study of human memory, as well as in most other areas of psychology. Most papers published in psychology make use of inferential statistics. Most undergraduate and graduate students in psychology are required to take a course in inferential statistics. In fact, if all the courses required for a graduate or undergraduate psychology degree in all psychology departments were listed, statistics probably would be the mode. For some, it is hard to imagine how it could be otherwise; yet this is not the case for all other sciences.

Given the current state of the advancement of psychology, often it is necessary to demonstrate that results cannot be attributed to chance. As the field advances, or as more regular data are obtained, such demonstrations should no longer be necessary. Thus, in physics, for example, the use of inferential statistics is exceedingly rare (Binder, 1963; Meehl, 1978). The null-hypothesis question (Is the speed of light constant in a vacuum?) is asked, rather than the more primitive psychological form (Is the speed of light different from zero at the .05 level?). One knows that results could not have occurred by chance. The same is true in certain areas of psychology. In psychophysics and in animal learning, it is possible to publish papers without using inferential statistics. There are questions of interpretation of the results, but the interpretations do not include the possibility that the results occurred by chance. In general, it is likely that as psychology advances, the role of inferential statistics in psychology will decrease. I, for one, look forward to the day when psychologists studying cognition will be embarrassed by the period in which they repeatedly had to report in journal articles that their results did not occur by chance.

The differences between the approaches of seeking regularity and seeking control may have been exaggerated slightly in order to make them clearer. In actual practice, the two extremes always mix and always share some common properties. Both can be done well or poorly. Both can be done with or without intelligence and creativity. Both need results that are repeatable under their respective conditions of observation. Both are of little interest if such repeatable results have no implications for theory development. Neither approach can be done to the exclusion of the other. Nonetheless, seeking regularity and seeking control remain different approaches to psychology.

Searching for regularity

Arguments about the way science should be considered are pleasant diversions from research. It is the research itself, however, that is the real determinant of which arguments are best. If regularity, and not control, should be our current goal, then I should be able to demonstrate progress from this approach. If the research presented here convinces psychologists to alter their research behavior, then the arguments will have been superfluous. If the research fails to impress anyone, then the arguments, at best, will have been mere curiosities. In research, we put in our effort, and we take our chances.

Having wandered around the "real world" for some time now without a proper laboratory to protect me, I have stumbled across more than my fair share of regularity. By examining what works and what does not work, I have developed some heuristics that I believe increase the chances of finding interesting regularities.

Heuristic 1

Use as little control as possible in the beginning. Do not worry about the subjects' prior exposure to the material they will learn, or even where, when, or how well the subjects initially learned the material they will recall. No effort should be made to simplify or control the material being used or the context in which it is being presented and recalled.

Heuristic 2

Do not decide on a dependent measure until after looking at pilot data. Examine the data accumulated for possible regularities, and then adapt the dependent measure to what the data show. Psychologists tend to use "amount" measures as their default, such as amount recalled or amount of time taken. These usually do not show the greatest regularity in nonlaboratory situations. Such amount measures are very sensitive to the variables that are not controlled outside of the laboratory, such as motivation level (Weiner, 1966), retention interval (Wickelgren, 1972), and type of encoding used (Craik & Lockhart, 1972). This is one reason why amount measures are so popular in the laboratory. Of course, there are exceptions, but, usually, dependent measures based on relations among which items or which aspects of items are recalled are more stable in nonlaboratory settings. Such measures might include the order of recall, or which, rather than how many, parts of a stimulus are recalled.

Heuristic 3

Do not be tied too strongly to any one theory or hypothesis when starting to collect data. Rather, try to think of the greatest possible number of theories that could be used to explain possible findings (Broadbent, 1973). This heuristic is so obvious that I would leave it out if my reading the journals did not convince me it needed to be mentioned. No one but the author of an article knows what was the true course of events leading to published research, but research usually is written as if one specific hypothesis and, except for confounding factors that were controlled, only one hypothesis was considered. Broadbent (1973) argues for the folly in this approach. First, a researcher often tests a pet theory by predicting an observation that would also be predicted by a host of other theories (Meehl, 1978). Thus, rather than testing the theory of interest, a whole set of theories is tested, and little support is gained for the pet theory itself. Second, and more seriously, if the predicted observation fails to materialize, the researcher is in a difficult position. If an alternative theory does not exist, the researcher may be tempted to search for explanations for the failure, without discarding the pet theory itself (Petrinovich, Chapter 2, this volume). That is, the researcher may not find it easy to be as impartial a judge as would be possible if several alternative theories had been postulated. Efficient progress in science will be fostered by serious consideration of sets of alternative theories and by emotional detachment from any one particular theory.

I shall draw examples almost completely from my own work so that I can report the actual steps that went on in the research. For each example given, I shall show how the three heuristics functioned. Where possible, I shall include data not previously reported.

Example 1: The tip-of-the-tongue phenomenon

When a person can almost, but not quite, recall a word after being given its definition, many facts about the word often are known, such as how many syllables are in the word and the first letter of the word (Brown & McNeill, 1966). Heuristic 1 applies to the conditions of learning in this situation. The word that is being partially recalled was learned without laboratory control. The experimenter has no knowledge of the context or spacing of the learning, the amount of prac-

Table 7.1. *Number of correct letters*

	Target word										
Direction of scoring	p	h	i	l	a	t	e	l	i	s	t
Left	27	24	16	16	2	0	0	0	0	0	0
Right	0	0	0	0	0	0	0	0	18	18	18

tice, the potential interference from other words learned at approximately the same time, the retention interval measured from the word's last use, or a host of other factors known to affect memory. In this study, heuristic 1 also applied to the recall situation used (Rubin, 1975). Subjects were simply asked to recall the letters of the word, using blanks to indicate letters they could not recall. Brown and McNeill found that subjects tended to know the first letter of a word by asking subjects if they knew the first letter. In contrast, I asked subjects to record all they knew, hoping to find out whatever it was that they in fact did know. In sum, as little control as possible was applied to both the learning and the recall situations.

A response measure was needed to capture the regularity present in the recall. Subjects in pilot studies appeared to recall clusters at the beginnings and ends of words. That is, they not only knew the first letter that Brown and McNeill had asked them to report but also knew the first and last few letters. The subjects, however, were not especially accurate on how many letters they did not know in the middle of a word. A dependent measure was therefore chosen that scored letters as correct if they were in the correct positions counting from either the beginning or the end of the word. Starting at the beginning of a word, a letter was scored correct only if it and the letter before it were correct. Starting at the end of a word, a letter was scored as correct only if it and the letter after it were correct. This measure was an adaptation of the common transitional error probability; so it, like the rest of the measures to be introduced here, was far from novel. Only the contexts in which the measures were used were different. Table 7.1 shows the results of applying this transitional-probability type of measure to the recall of 37 subjects who were in the tip-of-the-tongue state when presented with the definition of the word *philatelist.*

The theory used to describe the results was not the idea that guided the research. Brown and McNeill reported that their subjects, when in the tip-of-the-tongue state, knew that the letter *p* began the word *philatelist.* It seemed likely that the subjects who knew *p* also knew at least the phoneme indicated by the first two letters, *ph.* My hunch was partly correct, but it also appears that subjects in the tip-of-the-tongue state actually know the entire first morpheme, *phil.* The hypothesis of morpheme-like recall was driven by the data. The hypothesis did not lead me to do the experiment. I tried to find out something about the tip-of-the-tongue phenomenon by applying as little control as possible and by fit-

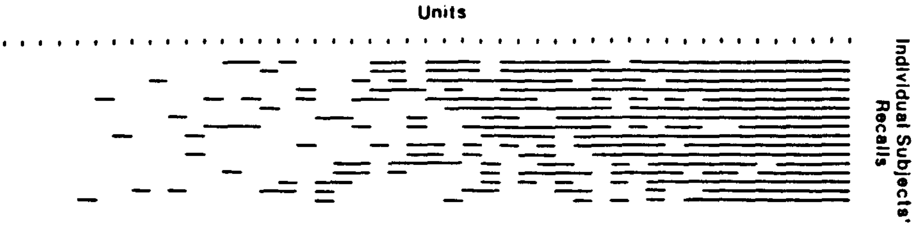


Figure 7.1. A scalagram analysis of the recalls of 16 older men for the 47 units of the Lincoln story. Both the subjects and the units are rank-ordered by probability of recall.

ting the dependent measure to the regularity observed. The particular hypothesis served only to describe the results. I had a whole set of possible hypotheses based on single-letter, phoneme, syllable, and morpheme clusters that I was ready to accept, given different possible outcomes. The interpretation was then checked using a different method and a different sample of subjects.

Example 2: Prose memory

When I started research on prose memory in 1970, such materials were considered by many as not controlled enough for laboratory study. Heuristic 1 was applied, in that I was working in the least controlled situation in which I thought that lawful results could be found. While scoring some pilot recalls, I noticed that subjects tended to recall different amounts of material, but tended to recall, or not recall, the same units. It was almost as if some parts of the passage were always recalled, some parts were never recalled, and some parts were recalled only by those subjects who recalled most of the passage. Applying heuristic 2, dependent measures were chosen to capture this regularity. Scalagram analysis (Guttman, 1947; Kenny & Rubin, 1977) was the first measure to be adapted to describing the regularity observed. In scalagram analysis, the complete data matrix of individual subjects recalling or failing to recall individual units is displayed. Both the subject axis, which runs horizontally, and the units axis, which runs vertically, are rank-ordered by amount recalled. Figure 7.1 displays the data for 16 subjects in a normative aging study (mean age = 68 years) who recalled the Lincoln story after a 10-min retention interval (Rubin, 1978). Perfect scalability would result in solid lines for those in the upper part of the figure, followed by uninterrupted blanks, and would imply that for all subjects there exists a single rank order of units from most to least likely to be recalled. Violations of this ideal are counted, and a normalized index is reported. The coefficient of reproducibility for Figure 7.1 is .84, indicating that if I were told exactly how many units each subject recalled and if I were given the group's overall rank order, then I could predict exactly which units each subject would recall and be correct an average of 84% of the time.

A second, even simpler dependent measure was formed for the same data: the number of subjects recalling each unit (Rubin, 1978, 1985). This measure, which

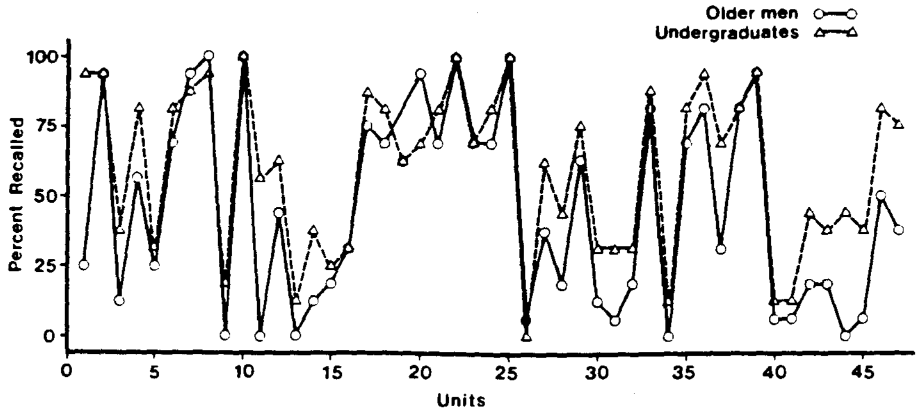


Figure 7.2. The probability of each of the 47 units of the story shown in Figure 7.1 being recalled by the subjects whose data are shown in Figure 7.1 and by 16 undergraduates.

is equivalent to adding across scalability figures such as Figure 7.1 to obtain one value for each unit of the story, was used to describe another regularity noticed in the data. Figure 7.2 displays the same data as Figure 7.1, as well as recalls from 16 undergraduates. The horizontal axis displays the units of the Lincoln story in the order in which they occur in the story. The vertical axis displays the percentage of the 16 subjects in each group who recalled each unit. The older subjects recalled only three-quarters as much as the undergraduates (45% versus 60%), but they tended to recall the same units. The correlation between the two groups on which units they recalled, calculated over the 47 units of the Lincoln story, is .85, compared with an average reliability (Cronbach's alpha) of .90 (Cronbach, 1951). This indicates that the recalls of the two groups correlate with each other almost as well as the recall of each group would be expected to correlate with those for new groups of subjects drawn from the same population.

Heuristic 3 suggests that one should not be strongly tied to any one theory, but rather should think of as many theories as possible. A dozen theories for why specific units would be recalled were considered. The factors in these theories varied from serial position to the contribution of each unit to the overall image produced by the story. If reliable differences had been observed between the two age groups, they could have been probed using all of the dozen theories.

Example 3: Very long term memory

The same techniques that were applied to prose learned in the laboratory can be used to study material learned without the benefit of laboratory control. As might be expected from the arguments made earlier, such material often can

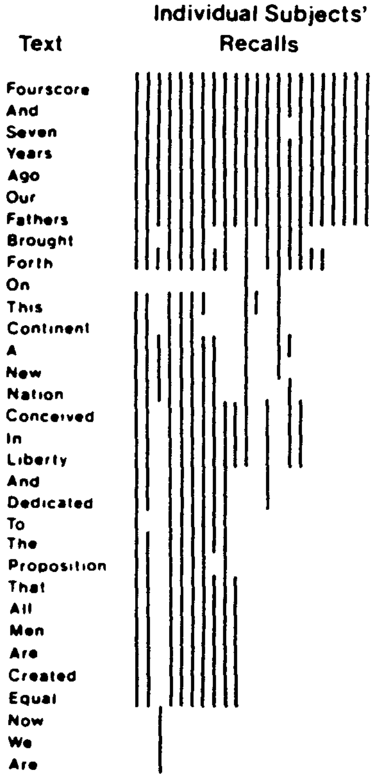


Figure 7.3. Recalls of the Gettysburg Address. The subjects are rank-ordered by the amount they recalled.

demonstrate greater regularity than that learned under more controlled conditions. Figure 7.3 is a scalagram like that of Figure 7.1, except that the units have been left in their original sequential order instead of being rank-ordered by amount recalled. Figure 7.3 displays the recalls of 26 subjects for the exact words of the Gettysburg Address (Rubin, 1977). The only errors that were scored as correct were spelling errors and the substitution of *forefathers* for *fathers*, a substitution made by 12 of the 22 subjects who recalled the word. The units in Figure 7.3 were not reordered by amount recalled because the actual order of the words in the passage is such a good predictor. Using the rank order of units of the group as a whole and the number of units each individual recalled, exactly which word each individual recalled can be predicted 97.6% of the time. Using primacy, instead of the empirical rank order, results in only a slight drop to 97.0%. Thus, not only do all subjects tend to recall the same words, they also tend to start at the beginning, recalling as much as they can until they stop. For this example, competing hypotheses were not considered. The regularity observed was strong enough to rule out most reasonable alternatives before they could even be considered. It should be noted that these data, collected without any laboratory control of learning or retention interval, are considerably more regular than the laboratory data of Figure 7.1.

Example 4: Semantic domains

The organization of semantic memory and its effect on the retrieval of information learned in and out of the laboratory have a long history in experimental psychology (Bousfield, 1953; Bousfield & Sedgewick, 1944; Kausler, 1974). Following heuristic 1, no control was exerted over when, how, and where the instances recalled were learned. For a semantic domain such as *animals*, the learning began very early and probably progressed at a slower rate through high school to college. For a semantic domain such as *all the faculty members at the university*, the learning had a much later onset and was of shorter duration. The recall situation was also as uncontrolled as possible, given that the subjects had to be informed to recall a specific semantic domain. Subjects were simply asked to recall as many instances of a semantic domain (e.g., animals) as they could. The open-ended nature of the recall situation allowed the subjects rather than the experimenter to define the instances of the semantic domain.

Following heuristic 2, the dependent measure was adapted to fit the regularity observed. Certain words tend to be recalled together in clusters. Because these clusters usually are small (Gruenewald & Lockhead, 1980), a measure based on local ordering was adopted. The similarity between any two items was defined as the number of subjects recalling the two items next to each other. This measure had the added advantages of a clear theoretical interpretation under most associative and search models of memory and a historical tie to the measure of clustering introduced by Bousfield (1953). Similarity matrices of the most commonly recalled instances of a domain were then constructed. The subjects thus determined which items would be considered as the central members of a domain and how the items would be related. The results are exceptionally stable and provide similarity spaces that are in good agreement with those obtained from other techniques such as similarity ratings (Rubin & Olson, 1980). Figure 7.4 shows an example of the similarity space that resulted from 20 recalls of the *animal* domain by a single subject. The figure shows clear clusters for American wild animals, African wild animals, farm animals, and small animals, including *mouse* and the pets *cat* and *dog*.

Of course, what is being studied both here and in the more traditional laboratory list-learning studies of clustering is the preexperimental, unobserved, and uncontrolled processing that led to the memory of the semantic domain being organized in the first place. Thus, the acquisition and retention of the material that is recalled in the method just described or that leads to the clustering of lists learned under experimental control are equally uncontrolled in both methods. In one case, however, a veneer of experimental control is applied to the experiment

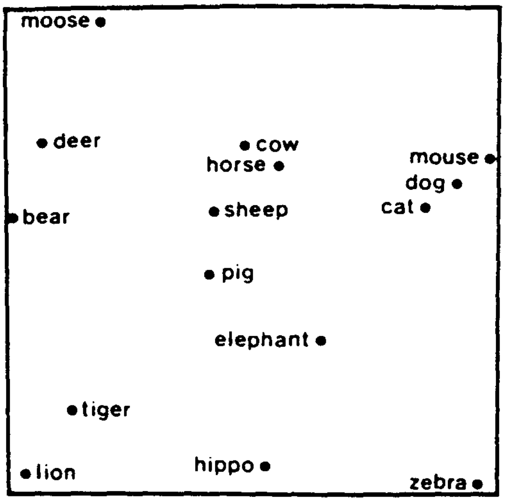


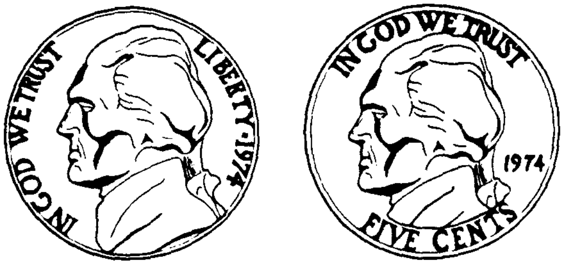
Figure 7.4. Similarity space produced from 20 repeated recalls of one subject. (From Rubin & Olson, 1980)

reported, at the cost of producing "noisier" data and less opportunity for uncovering unexpected results. This point is most obvious for this example, but it also holds for the other examples presented here.

Example 5: Recall of coins

At this point, the application of the heuristics should be predictable. For the recall of coins, little control was exerted over the conditions of encoding. Information about coins is learned over a long period of time, with repeated exposures occurring daily for most subjects. The exact amount and spacing of each subject's exposures and the amount of attention given to coins at each exposure were uncontrolled and unknown. Some subjects probably collected coins, studying in great detail their dates, mint marks, and other characteristics that affect their numismatic value. Other subjects did not. Minimal control was exerted over the conditions of recall. Subjects were simply asked to draw, in empty circles, the two sides of common coins. This was the most open-ended question that could be devised that would elicit information about the recall of coins. Recognition could have been used to assess memory for coins, but the foils and the target presented to the subject provided considerable information, and in our experience the particular foils used greatly affected the results (Kontis, 1982).

Following heuristic 2, a dependent measure was chosen to capture the regularity noted. Subjects' recalls were partial, but they tended to recall the same items in the same locations. As with the earlier examples, the amount recalled for a particular unit, rather than simply the total amount recalled, was chosen as a measure. In particular, the number of subjects who recalled each inscription in



Actual Coin

Modal Recall

Figure 7.5. An actual nickel and the nickel recalled by averaging over subjects' recalls. (From Rubin & Kontis, 1983).

each possible location on a coin was scored (Rubin & Kontis, 1983). Tables of such item–location pairings were reported, and modal coins were constructed by filling in the coin sequentially with the most frequently recalled item–location pair that did not already have its item or location used. Figure 7.5 shows the actual nickel and the modal nickel.

Following heuristic 3, theoretical preconceptions were kept to a minimum. The modal nickel shown in Figure 7.5 was identical with the modal penny, dime, and quarter drawn by subjects, except for the particular value and the particular profile drawn. The discussion centered on the role of schematic knowledge and the conditions necessary to have people store and recall detailed information not covered by a schema. The discussion was a way to describe the regularity observed and to suggest further experiments; it was not used to guide the initial formulation of the research. Once the regularity was observed, competing explanations were considered in an attempt to understand what had been found.

Example 6: The autobiographical memory of college students

For this example, the regularity to be discussed was described in an article by Crovitz and Schiffman (1974). The regularity was so impressive that it begged for further understanding. Crovitz and Schiffman presented 98 subjects with 20 common nouns and asked each to record the event from his or her life that each word evoked. The subjects were then asked to date their memories in terms of how long ago the incidents had occurred. Figure 7.6 shows the number of memories per hour that were reported as a function of hours since the incidents. Both axes are logarithmic; so the straight line is a power function. The points plotted are the common time markers of English ranging from 1 to 23 hours earlier to 1 to 17 years earlier. The straight line is of the same form as would be expected from laboratory studies of retention, and it supports the interpretation that the line represents the retention function for autobiographical memory (Rubin, 1982).

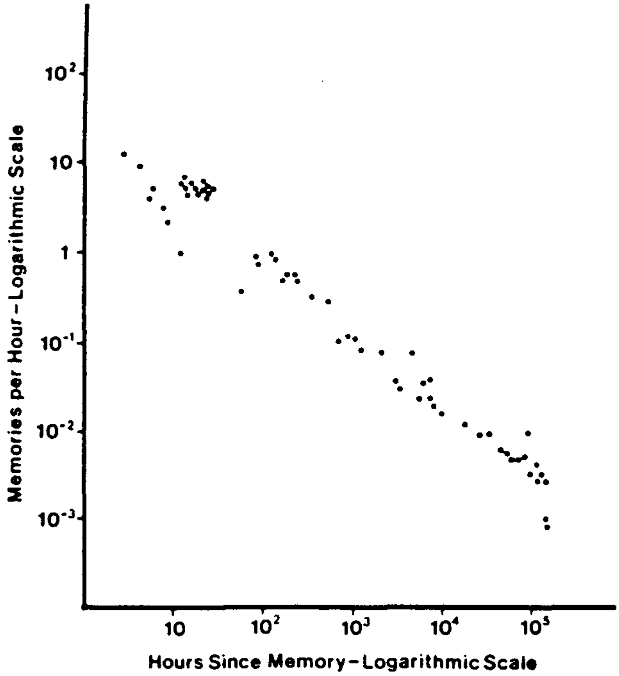


Figure 7.6. The distribution of autobiographical memories over time. (Adapted from Crovitz & Schiffman, 1974).

Crovitz and Schiffman had no control or knowledge of the learning, rehearsal, or other aspects of the encoding of the memories. They provided only the barest of constraint during recall. Moreover, the same results are obtained if even less constraint is applied by asking subjects to recall autobiographical memories in the absence of any cue words (Rubin, 1982). Thus, heuristic 1 was followed. Heuristic 2 was applied in that the plot shown in Figure 7.6 was formed to capture the regularity present in the data. In my work, I have attempted to follow heuristic 3 by formulating several possible theoretical explanations for the phenomenon and then attempting either to rule them out or to provide support for them.

Example 7: Autobiographical memory in older subjects

One extension of the Crovitz and Schiffman study involves the use of older subjects. Older subjects have had more years to build an autobiography to report than have younger subjects and therefore can provide information about changes over long time periods that undergraduates cannot. The same uses of heuristics 1, 2, and 3 apply here as in the younger subjects, and the results are the same for the most recent 20 years of both groups' lives. For periods further back than 20 years, the older subjects' data do not continue to decrease, but

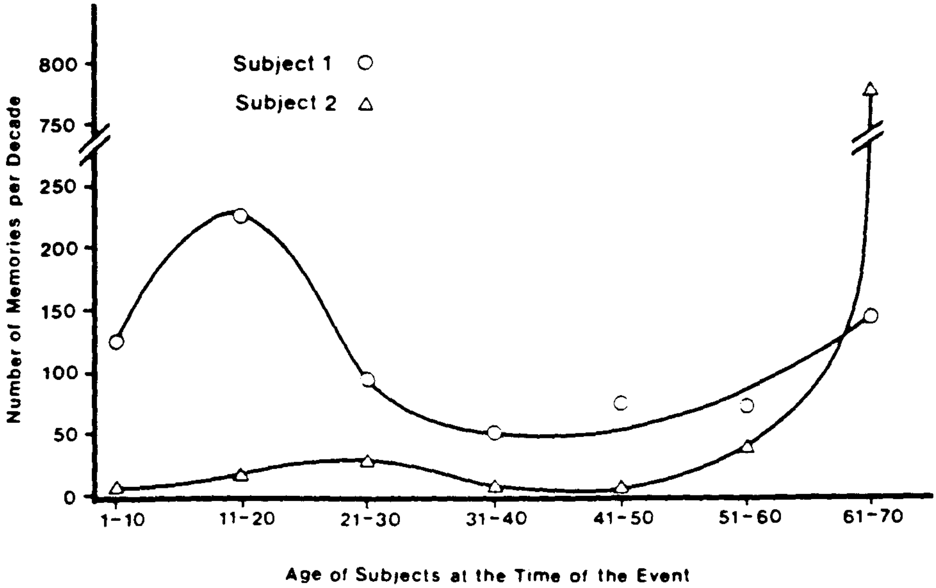


Figure 7.7. The distribution of autobiographical memories from two 70-year-old subjects.

rather show an increase for some periods in their youth. These findings hold for the reanalysis of data from several laboratories (Rubin, Wetzler, & Nebes, 1986). To describe these results, a simple histogram, rather than the more complex log-log transformation, is best. Figure 7.7 shows the histograms for two 70-year-old subjects who each provided dated memories to 921 cue words. The data are grouped into decades of life. It should be noted that it was the search for an understanding of an observed regularity that led to the reanalysis of existing data from other laboratories and that in turn led to the search for theoretical explanations.

Some lessons from the examples

Heuristic 1 suggests that experimenters should use as little control as possible. This was clearly followed in the examples presented here. In most cases, no control over or even knowledge of the host of parameters used to describe learning in the laboratory was used. This not only allowed a wide range of conditions of learning to be sampled in one experiment but also indicated that such a variety in conditions would not lead to a corresponding variety in results. Thus, whatever results were obtained could be generalized to a wide range of learning situations. The procedures used to obtain recalls also exerted as little control as possible, allowing the subjects to reveal what memories they had in a way that was as unbiased by the experimenter as was possible. In this way, subjects were free to present organization that the experimenter did not know to be present when the investigation was begun. For example, subjects demonstrated morpheme-like clustering in the tip-of-the-tongue phenomenon, marked primacy in very long term memory for sacred material, specific kinds of clustering in semantic domains, location-specific recall of items on coins, and interpretable distributions of memories over retention intervals in autobiographical memory. Nonetheless, the data observed were more regular than those usually obtained under strict laboratory control. That is, the variability among subjects on the dependent measures used usually was smaller than would be expected in the laboratory using standard measures. In fact, in many examples, the data were so regular that the results from individual subjects could be displayed (Figures 7.1, 7.3, 7.4, and 7.7), allowing for the form and degree of individual differences among subjects to be specified in detail (Bruce, Chapter 4, this volume).

Heuristic 2 suggested that the dependent measure to be used be tied to the regularity observed in the data. The dependent measures varied from example to example. In all cases they were more complex than the total amount recalled, but not much more complex. The measures were always the recall of a particular unit relative to other units of recall. The measures used may, at times, have been novel in the domain in which they were applied, but they were not novel measures. All the measures presented here have been used before, and most have known statistical properties. The main difference between the work presented here and more hypothesis-driven experimental work using the same measures is that here the measures were chosen to describe regularities noticed in the data, rather than being chosen a priori to test a hypothesis. Once the regularities were described in a quantitative fashion, they could be used to constrain and, through experimental manipulation, to test theories. In fact, the slightly more complex measures used here offer more of a challenge to theories than does the simpler amount-recalled measure. For instance, predicting the relative frequency of recall of units from a prose passage read or recalled under various conditions allows for a more efficient test of a complex theory than does predicting the amount of recall from those conditions (Rubin, 1985).

Heuristic 3 suggests that researchers should not be tied to one theory to try to explain the regularity they observe. As the examples demonstrate, this often leads to research that appears more data-driven than theory-driven. Theory, especially the kind of theory that has motivated laboratory research, however, is not ignored. It determines the general areas in which regularity is sought and what kinds of regularity are worth pursuing, and once theoretically interesting regularities are found, it shapes the experiments that are performed. If one were to criticize the collection of examples used to argue my points, one could say that as a collection they fail to cumulate into a body of knowledge in the way that a series of theory-driven experiments would. However, each example was an attempt to add to the cumulative knowledge base of an area of research at a very basic level. Rather than explicitly questioning the existing assumptions of these areas, an attempt was made to avoid as many of the assumptions as possible. The attempt was made by providing little control or constraint over the subject and the data analysis, thereby allowing the structure present in memory to become apparent.

How, when, and *why* should memory be studied outside of laboratory controls? *How?* By using the three heuristics discussed in the second half of this chapter. *When?* When a theory does not yet exist in an area of research that makes predictions worth the effort of testing. *Why?* Because there is more regularity and more chance of finding theoretically interesting results in data collected using the naturally occurring learning, the wide range of conditions, and the minimal control that are more common outside of the laboratory.

References

Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 70,* 107–115.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology, 49,* 229–240.

Bousfield, W. A., & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30,* 149–165.

Broadbent, D. E. (1973). *In defense of empirical psychology.* London: Methuen.

Brown, R., & McNeill, D. (1966). The "tip-of-the-tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5,* 325–337.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Crovitz, H. F., & Schiffman, H. (1974). Frequency of episodic memories as a function of their age. *Bulletin of the Psychonomic Society, 4,* 517–518.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology.* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover. (Original work published 1885)

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin, 63,* 400–402.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review, 69,* 54–61.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Gruenewald, P. J., & Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 225–240.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement, 7,* 274–279.

Kausler, D. H. (1974). *Psychology of verbal learning and behavior.* New York: Academic Press.

Kenny, D. A., & Rubin, D. C. (1977). Estimating change reproducibility in Guttman scaling. *Social Science Research, 6,* 188–196.

Kontis, T. C. (1982). *In search of the schema for a common object.* Unpublished senior honors thesis, Duke University, Durham, NC.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Rubin, D. C. (1975). Within word structure in the tip-of-the-tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior, 14,* 392–397.

Rubin, D. C. (1977). Very long-term memory for prose and verse. *Journal of Verbal Learning and Verbal Behavior, 16*, 611–621.

Rubin, D. C. (1978). A unit analysis of prose memory. *Journal of Verbal Learning and Verbal Behavior, 17*, 599–620.

Rubin, D. C. (1980). 51 properties of 125 words: A unit analysis of verbal behavior. *Journal of Verbal Learning and Verbal Behavior, 19*, 736–755.

Rubin, D. C. (1982). On the retention function for autobiographical memory. *Journal of Verbal Learning and Verbal Behavior, 21*, 21–38.

Rubin, D. C. (1985). Memorability as a measure of processing: A unit analysis of prose and list learning. *Journal of Experimental Psychology: General, 114*, 213–238.

Rubin, D. C., & Kontis, T. C. (1983). A schema for common cents. *Memory & Cognition, 11*, 335–341.

Rubin, D. C., & Olson, M. J. (1980). Recall of semantic domains. *Memory & Cognition, 8*, 354–366.

Rubin, D. C., Wetzler, S. E., & Nebes, R. D. (1986). Autobiographical memory across the lifespan. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 202–204). Cambridge University Press.

Weiner, B. (1966). Effects of motivation on the availability and retrieval of memory traces. *Psychological Bulletin, 65*, 24–37.

Wickelgren, W. A. (1972). Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology, 9*, 418–455.

Wilson, W., Miller, H. L., & Lower, J. S. (1967). Much ado about the null hypothesis. *Psychological Bulletin, 67*, 188–196.