Cognitive Artifacts

Donald A. Norman

Norman, Donald A. (1991), "Cognitive artifacts", i John M. Carroll (red), Designing interaction, Cambridge University Press, Cambridge. 20 sidor.

A cognitive artifact is an artificial device designed to maintain, display, or operate upon information in order to serve a representational function.

The distinctive characteristics of human beings as a species are:

1. Their special ability to modify the environment in which they live through the creation of artifacts and

1. the corresponding ability to transmit the accumulated modifications to subsequent generations through precept and procedure coded in human language. (Cole, 1990, p. 1).

Artifacts pervade our lives, our every activity. The speed, power, and intelligence of human beings are drambtically enhanced by the invention of artificial devices, so much so that tool making and usage constitute one of the defining characteristics of our species. Many artifacts make us stronger or faster, or protect us from the elements or predators, or feed and clothe us. And many artifacts make us smarter, increasing cognitive capabilities and making possible the modern intellectual world.

My interest is in cognitive artifacts, those artificial devices that maintain, display, or operate upon information in order to serve a representational function and that affect human cognitive performance. In this chapter I discuss three aspects of cognitive artifacts:

1. Two differing views of artifacts: the system view and the personal view;
2. Levels of directness and engagement: the relationship between those aspects of artifacts that serve the execution of acts and those that serve the evaluation of environmental states and the resulting feeling of directness of control or engagement;
3. Representational properties of cognitive artifacts: the relationship between the system state and its representation in the artifact.

Some History

Despite the enormous impact of artifacts upon human cognition, most of our scientific understanding is of the unaided mind: of memory, attention, perception, action, and thought, unaided by external devices. There is little understanding of the informationprocessing roles played by artifacts and how they interact with the information processing activities of their users.

The power and importance of culture and artifacts to enhance human abilities are ignored within much of contemporary cognitive science despite the heavy prominence given to its importance in the early days of psychological and anthropological investigation. The field has a sound historical basis, starting at least with Wundt (1916), nurtured and developed by the Soviet socialhistorical school of the 1920s (Leont'ev, 1981; Luria, 1979; Vygotsky, 1978; Wertsch, 1985), and still under study by a hardy band of social scientists, often unified by titles such as "activity theory," "action theory," or "situated action," with much of the research centered in Scandinavia, Germany, and the Soviet Union.

In the early part of the 20th century, American psychology moved from its early interest in mental functioning to the behavioral era, in which studies of representational issues, consciousness, mind, and culture were considered, at best, irrelevant to science. These dark ages ended in the mid1950s, but by then, the historical continuity with the earlier approaches and with European psychology had been lost. As a result, American cognitive psychology had to recreate itself, borrowing heavily from British influences. The emphasis was on the study of the psychological mechanisms responsible for memory, attention, perception, language, and thought in the single, unaided individual, studied almost entirely within the university laboratory. There was little or no emphasis on group activities, on the overall situation in which people accomplished their normal daily activities, or on naturalistic observations. Given these biases and history, it is no surprise that little thought was given to the role of the environment (whether natural or artificial) in the study of human cognition.

The field has now returned to pay serious attention to the role of the situation, other people, natural and artificial environments, and culture. In part, this change has come about through the dedicated effort of the current researchers, in part because the current interest in the design of computer interfaces has forced consideration of the role of real tasks and environments, and therefore of groups of cooperating individuals, or artifacts, and of culture.

The birth, death, and now apparent rebirth of the interest in culture and artifacts in thought is reflected in a survey paper by Cole, "Cultural psychology: a once and future discipline?" (Cole, 1990). For Cole, cultural psychology builds on the two major assumptions that stand as the opening quotation to this chapter: (1) the human's ability to create artifacts; (2) the corresponding ability to transmit accumulated knowledge to subsequent generations.

In this chapter I emphasize the informationprocessing role played by physical artifacts upon the cognition of the individual hence the term cognitive artifact. Here, I will not be concerned with how they are invented, acquired, or transmitted across individuals or generations. The goal is to integrate artifacts into the existing theory of human cognition.

The field of humancomputer interaction has pioneered in the formal study of the cognitive relationship between a person's activities, the artifact of the computer, and the task, and this chapter is a result of work in that tradition. However, most of the work has been narrowly focused on the details of the "interface" between the person and the machine. But it has become increasingly clear that the nature of the interaction between the people and the task affects the artifact and its use, with the view and use of the artifact varying with both the nature of the task and the level of expertise and skill of the people (e.g., see Bannon & B0dker, this volume, both for a clear description of this philosophy and also for a general review). I agree that we need a broader outlook upon tools and their use, but we also need better scientific understanding of the role played by the artifact itself, and so the main focus is upon the properties of the artifact and how its design affects the person and task.

It is clear that we are entering a new era of technology, one dominated by access to computation, communication, and knowledge, access that moreover can be readily available, inexpensive, powerful, and portable. Much of what will transpire can be called the development of cognitive artifacts, artificial devices that enhance human cognitive capabilities. As we shall see, however, artifacts do not actually change an individual's capabilities. Rather, they change the nature of the task performed by the person. When the informational and processing structure of the artifact is combined with the task and the informational and processing structure of the human, the result is to expand and enhance cognitive capabilities of the total system of human, task, and artifact.

Two Views of Artifacts: The System View and the Personal View

The most obvious analysis of an artifact is that it enhances human ability. According to this analysis an artifact such as a pulley system makes us stronger, a car makes us faster, and paper and pencil make us smarter. By this analysis, artifacts such as written notes, books, and recordings amplify the cognitive power of human memory and artifacts such as mathematics and logic amplify the power of thought. The notions that artifacts enhance or amplify may be natural, but as Cole and Griffin point out in their essay "Cultural amplifiers reconsidered" (1980), they are badly misleading.

Artifacts may enhance performance, but as a rule they do not do so by enhancing or amplifying individual abilities. There are artifacts that really do amplify. A megaphone amplifies voice intensity to allow a person's voice to be heard for a greater distance than otherwise possible. This is amplification: The voice is unchanged in form and content but increased in quantity (intensity). But when written language and mathematics enable different performance than possible without their use, they do not do so by amplification: They change the nature of the task being done by the person and, in this way, enhance the overall performance.

Artifacts appear to play different roles depending upon the point from which they are viewed. When a person uses an artifact to accomplish some task, the outside observer sees the system view, the total structure of person plus artifact (Figure 2.1) in accomplishing that task. The person, however, sees the personal view: how the artifact has affected the task to be performed (Figure 2.2).

Figure 2.1. The system view of a cognitive artifact. Under this view, we see the entire system composed of the person, the task, and the artifact. Seen from this perspective, the artifact enhances cognition, for with the aid of the artifact, a system can accomplish more than without the artifact.

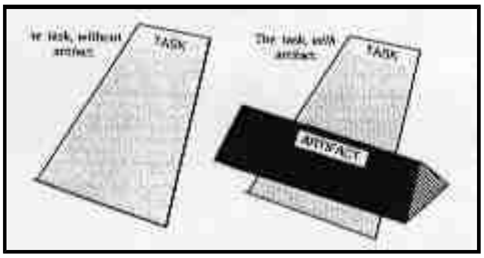

Figure 2.2. The personal view of a cognitive artifact. Under this view, that of the individual person who must use the artifact, the view of the task has changed: thus, the artifact does not enhance cognitionit changes the task. New things have to be learned, and old procedures and information may no longer be required: The person's cognitive abilities are unchanged.

The System View of an Artifact

The two views of artifacts, and an illustration of how cognition is distributed across people and technology, can perhaps most easily be illustrated by example. Consider the everyday memory aid, the reminder or "todo" list, or in industrial contexts, the checklist for a task (e.g., the checklists used by pilots before each critical phase of flight in a commercial aircraft). From the system point of view, checklists enhance memory and performance; from the personal point of view, they change the task.

At first, the checklist or todo list may appear to be a memory aid. It can be seen to help us remember what to do during the course of our activities. In fact, there can be no question that checklists change our behavior and prevent some kinds of forgetting: They are so effective in industrial and aviation settings, that their use is often required by regulation. It is tempting to say that a list extends or enhances our memory. After all, with it, we can perform as if we had a perfect memory for the items on the list. Without it, we occasionally forget to take important actions. When we think of the todo list in terms of what the personpluslist system can do, we are looking at one view of the artifact. This is the view of the artifact from afar, looking at it in the context of the person and the task to be performed: This is the system view. The system view of the list is as a memory enhancer.

The Personal View of an Artifact

The checklist or todo list has another view, the view it presents to the task performer: this the personal view. From the point of view of the user of the artifact, using the list is itself a task. Without the list, we must remember or plan all of our actions. With the list, we need to do very little remembering and planning: The planning and "remembering" were done ahead of time, at the time we made up the list. At the time we perform the individual actions we need not repeat the planning and remembering. The use of a list instead of unaided memory introduces three new tasks, the first performed ahead of time, the other two at the time the action is to be done:

1. The construction of the list;
2. Remembering to consult the list;
3. Reading and interpreting the items on the list.

The fact that the preparation of the list is done prior to the action has an important impact upon performance because it allows the cognitive effort to be distributed across time and people. This preparatory task, which Hutchins calls "precomputation" (E. Hutchins, 1989, personal communication), can be done whenever convenient, when there are no time pressures or other stresses, and even by a different person than the individual who performs the actions. In fact, precomputation can take place years before the actual event and one precomputation can serve many applications.

In the aviation setting, flight checklists are prepared by the chief pilot of each airline, approved by the Federal Aviation Authority, and then passed on to the pilots who use them for many years and many thousands of flights without further modification: This is both precomputation and a distribution of the cognitive task of planning across people and time. To the aviation system, the checklist enhances memory and accuracy of action; to the individual pilots, the checklist is a new task inserted into the daily routine, and at times it is apt to be viewed as extraneous to the main goals of the day. As such it is a nuisance and it can lead to new classes of errors: Some of these errors may resemble those that would occur without the use of the checklist, and some may not.

When we compare the activities performed with an without the aid of a reminder list, we see that the conclusion one draws depends on the point of view being taken. To the outside observer (who takes the system view), the same actions are intended to be performed with and without the list, but (usually) they are carried out more accurately and reliably with the list. To the individual user (who takes the personal view), the list is not a memory or planning enhancer; it is a set of new tasks to be performed, with the aspects of the list relevant to memory and planning separated from the aspects of the list relevant to performance.

Every artifact has both a system and a personal view, and they are often very different in appearance. From the system view, the artifact appears to expand some functional capacity of the task performer. From the personal view, the artifact has replaced the original task with a different task, one that may have radically different cognitive requirements and use radically different cognitive capacities than the original task.

This analysis points out that from all points of view, artifacts change the way a task gets done. In particular, artifacts can:

Distribute the actions across time (precomputation);

Distribute the actions across people (distributed cognition);

Change the actions required of the individuals doing the activity.

Levels of Directness and Engagement

When we use an artifact to do a task, of necessity we make use of a representation. Artifacts act as mediators between us and the world, both in execution (between actions and the resulting changes to the world state) and in perception (between changes in the world and our detection and interpretation of the state). The nature of the interaction between the person and the object of the task varies from direct engagement to a very indirect, remote form of interaction. Thus, when we write or draw with a pencil on paper, there is a direct relationship between movement of the pencil and the resulting marks on the paper. When we ask someone else to write or draw for us, the relationship is much less direct. Some interactions are so indirect and remote that feedback and information about the world state are difficult to get and possibly delayed in time, and incomplete or of unknown accuracy. These differences can have a major impact upon task performance and to a large extent are controlled by the design of the task and the artifact. (See the important discussion by Laurel, 1986, which introduces the concept of "direct engagement.")

Bodker (1989) distinguishes among several possible relationships among the person, the artifact, and the objects being operated upon. Thus, the artifact can be used to mediate directly between the person and the object (as in using a hammer or chisel to operate upon nails or wood). Or the artifact can present a virtual object or world upon which operations are performed, eventually to be reflected onto the real object.

In some cases, the virtual world exists only within the computer (as in building a spreadsheet or graphic object that will never exist outside the computer). The object might actually exist outside the computer, but be created or operated upon through the virtual world of the artifact (as in controlling an industrial process through the computer display, or developing the content and format of a publication within the computer word processor and publishing system). In these cases, here are several layers of representation: representations the represented world of he real object; representations the representing world within the artifact; represention3, the way the artifact displays the virtual world; and representations the ental representation of the human.

Actions are performed through a feedback mechanism involving both an execuon and evaluation phase (Figure 2.3). Both phases of the action cycle need suport from the representational format used by the artifact. The choice of representaon and interactions permitted by the artifact affect the interaction of the person ith the object, whether real or virtual (Hutchins, Hollan, & Norman, 1986; Nortan, 1986, 1988). Different forms of artifacts have different representational implitions, which in turn dramatically affect the interactions.

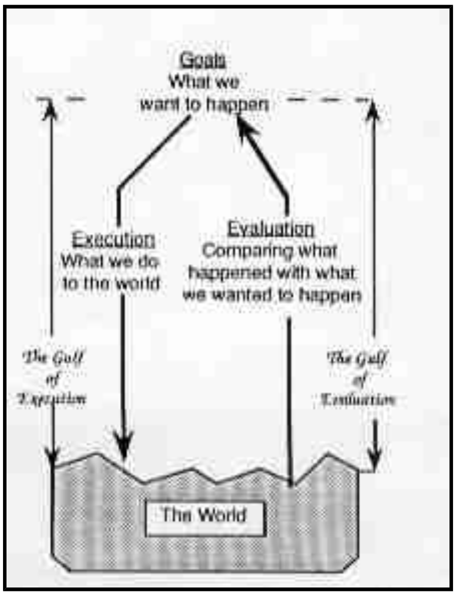

Figure 2.3. The action cycle. Artifacts that support action must support both the execution and evaluation phases of the action cycle, usually through different representations. The gulfs of execution and evaluation refer to she mismatch between our internal Coals and expectations and the availability and representation of informaion about the state of the world and low it might be changed. The gulf of xecution refers to the difficulty of acting upon the environment (and low

well the artifact supports those actions). The gulf of evaluation refers o the difficulty of assessing the state f the environment (and how well the rtifact supports the detection and interpretation of that state).

Activity Flow

The gulf of execution refers to the difficulty of acting upon the environment (and how well the artifact supports those actions). The gulf of evaluation refers to the fficulty of assessing the state of the environment (and how well the artifact suprts the detection and interpretation of that state). There are two ways of bridging e gulfs. One is by appropriate design of the artifact, the other through mental effort and training. Thus, with increasing skill, a person mentally bridges the gulfs, so that the operations upon the artifact are done subconsciously, without awareness, and the operators view themselves as operating directly upon the final object (Bodker, 1989; Hutchins, 1986; Hutchins et al., 1986).

Bodker introduces the notion of "activity flow" to describe the activity cycle in accomplishing a task. Automatization of effort and the resulting feeling of direct engagement can occur where a consistent, cohesive activity flow is supported by the task, artifact, and environment. Interruptions and unexpected results break the activity flow, forcing conscious attention upon the task. For many activities, this "bringing to consciousness" is disruptive of efficient performance.

The problem with disrupting activity flow is that the disruption brings to conscious awareness the disrupting activity, even when this is not the main focus of attention. This is usually undesirable, for it can have negative impact upon the task being performed. In fact, disruptions of this sort can lead to errors when the interrupting activity interferes with the maintenance of working memory for the task. The resulting memory difficulties may mean that the interrupted task is not resumed properly, either by being delayed beyond its proper execution time, by returning to the wrong point in the task, or by being forgotten altogether and never resumed: three classic forms of action errors. But deliberate disruption of the activity flow might be a useful safety device if it forces conscious attention upon critical, safetyrelated aspects of the task.

Automatic behavior is valuable in many skilled operations, for it permits the attention to be directed to one area of concern even while performing smoothly the operations required for another area for example, the way in which a skilled typist can enter text automatically while concentrating upon the construction of future sentences. But at times, it might be valuable to force conscious attention to some aspect of performance by deliberately breaking the activity flow.

Thus, "forcing functions" physical constraints that prevent critical or dangerous actions without conscious attention could be viewed as serving their function by a deliberate disruption of normal activities. A good example of a deliberate disruption of activity for safety purposes is the use of checklists in industry and, especially, in commercial aviation. In aviation, the checklist is often reviewed by both pilots, one reading aloud the items, the other confirming and saying aloud the setting of each item as it is read. These actions are intended to force a deliberate, conscious disruption of skilled behavior, deliberately breaking the normal activity flow. Safetyrelated checks and cautions should be disruptive in order to receive conscious attention. Automatic actions are the most susceptible to errors by action slips and to disruption by external events and interruptions. In fact, the checklist can fail in its function: After thousands of usages and years of experience, checklist use can be so routine that it does become automatic, sometimes with serious consequences (Degani & Wiener, 1990; Norman & Hutchins, 1990; NTSB, 1989).

The point is not that one class of interaction or representation is superior to another but that the different forms and modes each have different properties.

Representation and Artifacts

The power of a cognitive artifact comes from its function as a representational device. Indeed, I define a cognitive artifact as an artificial device designed to maintain, display, or operate upon information in order to serve a representational function. It is now time to take a look at some of the representational features of artifacts. This will be brief and incomplete: This work is just beginning and although the work so far is suggestive, a more complete analysis will have to come later.

Representational Systems

A representational system has three essential ingredients (Newell, 1981; Rumelhart & Norman, 1988):

The represented world that which is to be represented;

The representing world a set of symbols;

An interpreter (which includes procedures for operating upon the representation).

Surface Representations

Some artifacts are capable only of a surface level representation. Thus, memory aids such as paper, books, and blackboards are useful because they allow for the display and (relatively) permanent maintenance of representations. The slide rule and abacus are examples of computational devices which only contain surface representations of their information. These devices are primarily systems for making possible the display and maintenance of symbols: They implement the "physical" part of the physical symbol system. These are called surface representations because the symbols are maintained at the visible "surface" of the device for example, marks on the surface, as pencil or ink marks on paper, chalk on a board, indentations in sand, clay or wood.

Internal Representations

Artifacts that have internal representations are those in which the symbols are maintained internally within the device (unlike paper and pencil where the symbols are always visible on the "surface"). This poses an immediate requirement on the artifact: There must be an interface that transforms the internal representation into some surface representation that can be interpreted and used by the person. Artifacts that have only surface representations do not have such a requirement, for the surface representation itself serves as the interface.

The Interface between Artifact and Person

Cognitive artifacts need interfaces for several reasons. In the case of artifacts with internal representations, the internal representation is inaccessible to the user, so the interface is essential for any use of the artifact. Moreover, even for artifacts that have only surface representations, the style and format of the interface determine the usability of the device. Here, the standard issues in the field of interface design apply.

We can conceptualise the artifact and its interface in this way. A person is a system with an active, internal representation. For an artifact to be usable, the surface representation must correspond to something that is interpretable by the person, and the operations required to modify the information within the artifact must be performable by the user. The interface serves to transform the properties of the artifact's representational system to those that match the properties of the person.

To the user of an artifact, the representing world is the surface of the artifact the information structures accessible to the person employing the artifact. One of the basic issues in developing an artifact is the choice of mapping between the representing world and the represented world (or between the surface representation and the task domain being supported by the artifact). In the mapping between the represented world and the representing world of the artifact. the choice of representation determines how faithfully the match is met.

The Object Symbol

One major concern in interfaces is the relationship between control operation and system state. Usually, these two aspects of the interface are separated and handled by different components. The two different aspects are not always present, and even when they are, they may differ considerably from one another in physical location, conception, and form of representation. This independence of control and display was not always true, and it seems to have arisen more by historical accident than by design.

Some controls have the interesting representational property that they serve both as the objects to be operated upon and also as representations of their states (see Figure 2.4). Simple examples occur for any controls operated by physical levers, where the act of moving the lever changes both the system state and also the physical appearance of the device: The position of the lever is both the actual state of the device and also its representation. Norman and Hutchins named the situation where the physical object is both the object operated upon and the symbol of its state the "object symbol" (Norman & Hutchins, 1988). The special case in which the same object serves as both a control of its value and a representation of its value was first described by Draper (1986). who argued for the importance of treating input and output to a computer system as a unified activity.

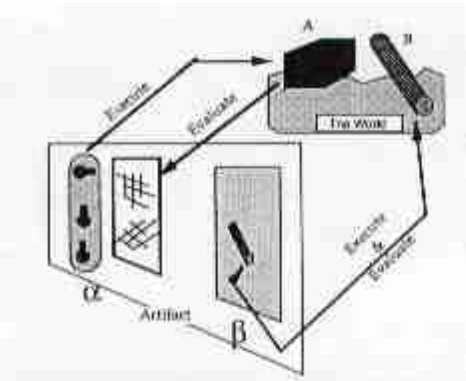

Figure 2.4. The object symbol. When a person manipulates a real or virtual world through an artifact, when the object in the artifact is both the means of control (for execution of actions) and also the representation of the object state (for evaluation), then we have the case of an object symbol. In condition a, the execution and evaluation are done separately. In condition alfa, the same representation is used for execution and evaluation; beta represents the case of the object symbol.

Object symbols used to be the prevailing mode of operation, for they represent the natural and frequently occurring mode of operation with mechanical systems, especially simpler systems. Many mechanical systems have the property that one directly manipulates the parts of interest and that one assesses the state of the device from the position of those same parts. The object symbol situation disappears when controls are physically removed from the site of action.

In the modern world of computer controls, the object symbol is rare. With modern electronic systems, the controls and indicators have almost no physical or spatial relationships to the device itself, which introduces an arbitrary or abstract relationship between the controls, the indicators, and the state of the system. But this state of affairs has come about by accident, not by design. The advantages of separating controls from physical equipment led to a natural separation of object and symbol. Once there was a separation, then the control no longer signaled system state. The result has been separation of the control of state from the indicator of state and, in some systems, a complete neglect of the development of appropriate representational forms for either control or display.

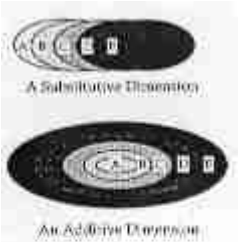

Figure 2.5. Substitutive and additive dimensions. Each of the ovals represents the representational aspects of values along the dimension from A to E. In the substitutive case, the representations replace one another. In the additive case, each successive representation includes the previous. Examples of additive dimensions are loudness and brightness. Examples of substitutive dimensions are pitch and hue.

Additive and Substitutive Dimensions

Many years ago, Stevens identified two forms of psychological representational dimensions or scales: additive and substitutive (Stevens, 1957). In an additive scale, the representations could be ordered, with each succeeding one containing the one before it, plus perhaps new aspects. The psychological percepts of loudness and brightness (which are the psychological mappings of physical sound and light intensities) form additive scales. In a substitutive scale, each new item replaces the one before it, with perhaps some overlap of attributes. The psychological percepts of pitch and hue (which are the psychological mappings of physical sound frequency and light wavelength) form substitutive scales.

Restle (1961) showed that these two scale types could be represented in settheoretic terms (as shown in Figure 2.5). In an additive scale, "as one moves along the sequence of sets one picks up new aspects, and one never loses any of the old ones. Any such sequence of sets is ordered in a strict way, and distances are additive" (Restle, 1961, p. 49). In a substitutive scale where, for example, one is moving from state A to state B. "each step of the process involves discarding some elements from A and adding some new elements from B. Elements from A which have earlier been discarded are never reused and elements from B which have been added are never discarded.... each move along the scale involves substituting some elements from B for some of the elements of A" (Restle, 1961, p. 50).

Representational Naturalness

I propose the following hypotheses about the form of representation used in a cognitive artifact.

Hypothesis 1: The "naturalness" of a mapping is related to the directness of the mapping, where directness can be measured by the complexity of the relationship between representation and value, measured by the length of the description of that mapping.

The use of "length of description" as the measure of naturalness is taken from the analogous use in specifying the complexity of a statement in complexity theory. The length of the description is, of course, a function of the terms used for the description. I propose that the terms be psychological, perceptual primitives.

It is important not to confuse the idea of the mapping terms with natural language or conscious awareness. The mapping terms are purely formal and do not imply that the person is aware of them. They are not terms in natural language.

Hypothesis 2: Experts derive more efficient mapping terms, thus reducing the complexity of a mapping and increasing its feeling of "naturalness." However, although these derived terms may simplify the mapping relationship, they always extract some penalty in time or computation (and, thereby, in mental workload) for their interpretation .

Hypothesis 2 accounts for the phenomenon that experts can apparently get used to any representation, without obvious decrease in performance (except for learning time). This hypothesis allows the apparent complexity and naturalness of a representation to change with the development of expert skill. However, because the derived mapping terms are built upon some set of perceptual primitives, these derived terms will need to be interpreted, thereby extracting some informationprocessing workload. In normal behavior, this will probably not be noticeable, but in times of heavy workload or stress, the extra workload required to use the derived terms should degrade performance.

In other words, although experts can get used to anything and even claim it to be natural and easy to use, less natural representations will suffer first under periods of heavy workload and stress.

Finally, I suggest that the choice of representation for the mapping between the representing world (the surface representation) and the represented world (the task domain being supported by the artifact) follow a guiding principle for appropriateness taken from the work of Mackinlay, Card, and Robertson (1989):

Appropriateness principle: The surface representation used by the artifact should allow the person to work with exactly the information acceptable to the task: neither more nor less.2

This principle is a direct paraphrase of the expressiveness principle for input devices developed by Mackinlay, Card, & Robertson (1989), namely: "An input device should allow the user to express exactly the information acceptable to the application: neither more nor less" (emphasis added). Mackinlay et al. were developing a language for describing the mapping between input device and function, which meant they were on a parallel undertaking to the one described here. In principle, their analyses can be translated into the ones needed for the study of the representational properties of artifact.

EST %

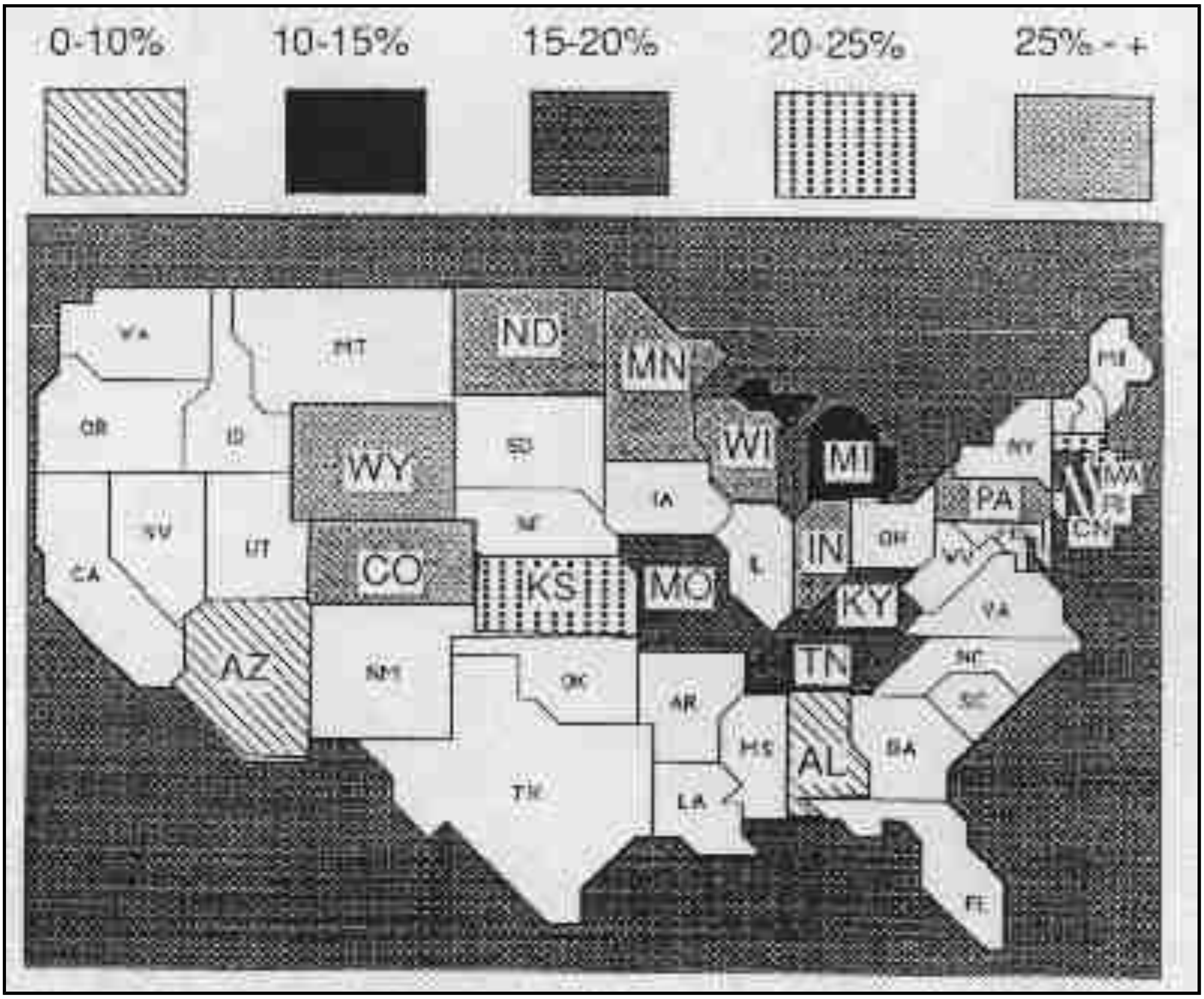

Figure 2.6. An unnatural mapping. Here, percentage (which is an additive dimension) is represented by a substitutive scale different shadings. And where the shadings can be ordered along an additive scale, the ordering conflicts with the ordering of percentages. (Redrawn from a figure in the Los Angeles Times, September 13, 1988, p. 21.)

Using Density to Represent Numerical Value

Example: Contrast the case where an additive scale is used to represent an additive domain with one in which a substitutive scale is used to represent an additive domain. Figure 2.6 illustrates the representation of percentages (an additive scale) by arbitrary shadings. According to Hypothesis 1, the superior representation would be to use an ordered sequence of density (an additive scale) to represent percentages (an additive scale), as shown in Figure 2.7.

Note that there is still a problem with the representation in Figure 2.7, but the problem helps emphasize the point about the importance of matching representational format. The white areas, perceptually, appear to represent the states with the least concentration of radon. This is because white fits on the ordered density scale to the left of (less than) the 010% density. In fact, white represents those states for which there are no data. One way to represent this situation to avoid the conflict in representational interpretation would be to delete the states for which there is no information from the map. I chose the method shown because the natural misinterpretation helps make the point about the impact of representational scale.

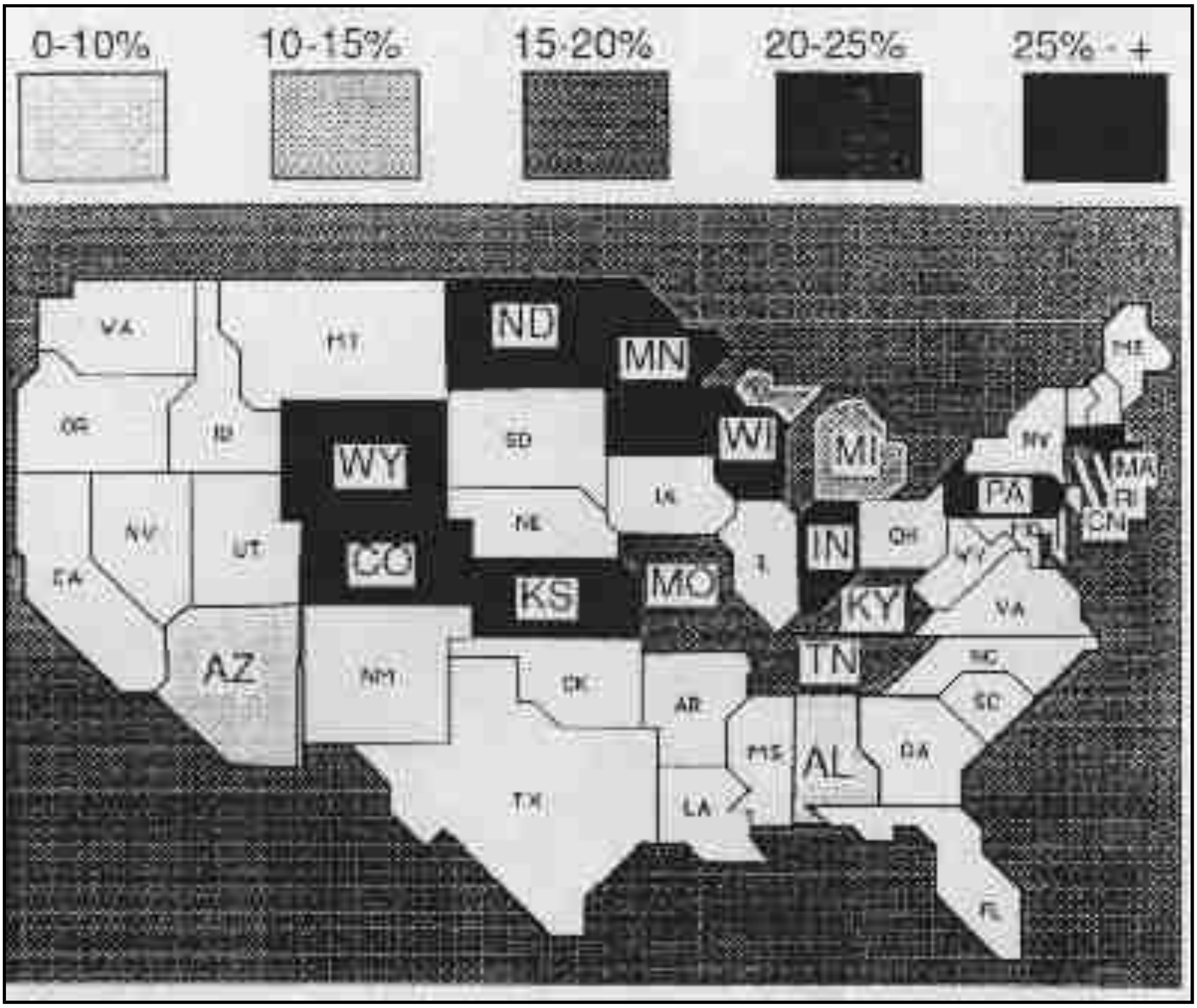

Figure 2.7. A natural mapping. Here, the map of Figure 2.6 has been redrawn so that percentage (which is an additive dimension) is represented by an additive scale ordered densities of shading. Now, the density ordering matches the percentage ordering. (Redrawn from a figure in the Los Angeles Times, September 13, 1988, p. 21.)

Color hue is frequently used to represent density or quantity, especially in geographic maps, satellite photographs, and medical imagery. But hue is a substitutive scale, and the values of interest are almost always additive scales. Hence, according to Hypothesis 1, hue is inappropriate for this purpose. The use of hue should lead to interpretive difficulties. In fact, people who use these color representations do demonstrate difficulties by their continual need to refer to the legend that gives the mapping between the additive scale of interest and the hues. According to the hypothesis, density or brightness would provide a superior representation. It would probably be even better to use a spatial third dimension for representing this information.

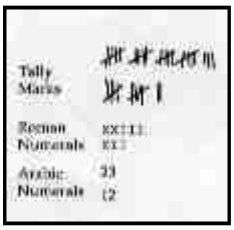

Figure 2.8. Differing representations for numerical quantity. If one simply wishes to compare numerical values, tally marks are superior to Arabic numerals, for the length of the representation is analogous to the numerical value. If one wants to do arithmetic operations, the symbolic (Arabic) representation is better, even though length is not a good indication of value. The Roman numeral representation is a compromise, being somewhat symbolic, but also approximately proportional to the value being represented.

Legends of maps and graphs are usually used to present the mapping rule for the representational code being used. According to my hypotheses, frequent use of legends is a sign of inappropriate representational mapping. With appropriate representations, the mapping code is easily learned and applied: Legends should not be essential to understanding.

Representations for Comparing Numerical Counts

Even such a simple example as counting items in order to compare quantity provides another instance of the use of mapping rules. When one is interested in comparing the values of counts to determine which is greater, according to these hypotheses, the superior form of representation will have the size of the representation itself map onto the size of the number. Size comparisons require additive comparisons.

Line length provides an additive representation. The Arabic numeral method for representing number does not. Counting methods that use tally marks to represent the number of objects translate number into length in this case, the length of the space required to show the tally marks (Figure 2.8). Tally marks, therefore, provide an additive representation in which the size of the representation is related to the value of the number.

Thus, according to Hypothesis 1, Arabic notation is inferior for simple Boolean comparisons because its perceptual representation bears little relationship to its numerical value: There is only a weak perceptual relationship between the physical dimensions of a numerical representation and its numerical value (the physical length of the number how many digits it contains is proportional to the logarithm of its value but with a discreteness of resolution good only to within a factor of 10). But Arabic notation is superior to all other common notations when numerical operations need to be performed.

Most people feel uncomfortable with this result because the comparison of Arabic numerals seems natural and straightforward. Here is where Hypothesis 2 comes into play. Most people forget the years of training it has taken to reach this state of naturalness. Moreover, there is psychological evidence that the time to compare two different (Arabic) numbers varies with the size of the difference between the numbers, strongly suggesting that an internal translation has to be made into the more primitive, additive representation, as suggested by Hypothesis 2. Moreover, I would predict that under heavy workload, comparisons of Arabic numbers would suffer.

However, in cases where an exact numerical value is required or where numerical operations need to be performed, Arabic notation is clearly superior which is why it is the standard notation used today. The form of representation most appropriate for an artifact depends upon the task to be performed, which is one reason that so many different numerical representations do exist (Ifrah, 1987; Nickerson, 1988).

Intrinsic Properties of Representation

Some years ago, Palmer described several properties of representations, including two that he called "intrinsic" and "extrinsic" (Palmer, 1978). The important point of these attributes is that they constrain what one can do with representations. A simple example will suffice.

Consider three objects: A, B. and C. Suppose that we know that object A is taller than both object B and object C, but we don't know which is taller, B or C. We can represent this state of affairs very nicely by symbolic expressions. Let H(i) be the height of object i. Then we know that:

$H(A) > H(B);$

$H(A) > H(C)$

We do not know the relationship between H(B) and H(C), and this symbolic form of representation does not force us to represent the relationship. That is an important, positive aspect of this form of representation. However, on the negative side, there is nothing to stop us from writing a contradictory statement:

H(B) > H(A),

or even

H(A) > H(A)

Suppose we represented the objects by a visual image: In the image, height of the object would be represented by height of the image. A possible representation for the three objects is shown in Figure 2.9.

Note that with an image, it is simply not possible to represent an object without also representing its form and size: In this case, the representation of height is an intrinsic property of a visual image. Moreover, it is simply not possible to enter a contradictory statement in the same way that we could with the other representational format.

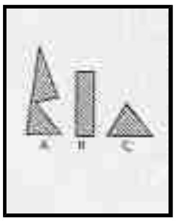

Figure 2.9. Intrinsic properties of a representation. Using images to represent the objects A, B. and C, we cannot also avoid representing their form and dimensions. Even if we did not know the height of C, we would be forced to select some value under this form of representation.

The form of representation used by an artifact carries great weight in determining its functionality and utility. The choice of representation is not arbitrary: Each particular representation provides a set of constraints and intrinsic and extrinsic properties. Each representation emphasises some mappings at the expense of others, makes some explicit and visible, whereas others are neglected, and the physical form suggests and reminds the person of the set of possible operations. Appropriate use of intrinsic properties can constrain behavior in desirable or undesirable ways.

Forcing functions are design properties that use the intrinsic properties of a representation to force a specific behavior upon the person (Norman, 1988). Thus, in normal operation, it is not possible to start a modern automobile without the proper key, for the ignition switch is operated by turning the key: The switch has a builtin forcing function that requires insertion of the key. One of the intrinsic properties of the lock is the lack of affordances for turning. One of the intrinsic properties of a key is the affordance it offers for rotation of the lock (assuming it is the proper key for the particular lock). However, it is possible to leave the automobile without removing the key from the ignition there is no forcing function. Bells and alarms that accompany the opening of the door without removing the key are not forcing functions. These are reminders extrinsic or addedon properties of the system. They can remind the user but they allow the behavior. A forcing function would require the key to open the door, or perhaps make it so that the door would not open with the key still in the ignition. These forcing functions, of course, have undesirable consequences.

Any design can be thought of as a representation. The designer has to decide how to represent the features of the device, how to implement the operation, and how to represent the current state. In the choice of design, many factors come into play, including aesthetics, cost, manufacturing efficiency, and usability. The face that the device puts forward to the person is often a compromise among the competing requirements of these different factors, but this face the interface is a representation. Forcing functions are simply the manifestations of the intrinsic properties of the design represention.

Representations carry with them many subtle intrinsic properties, often ones not intended by the designer. Line lengths represent quantity, and two lines of different lengths thereby intrinsically present a comparison of the lengths, even if that is not intended by the designer. Many inappropriate uses of graphs can be traced to conflicts with the unintended intrinsic properties of the graphs.

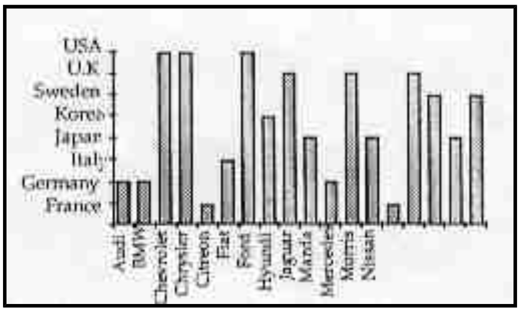

Figure 2.10. Inappropriate use of an additive scale. This example, inspired by Mackinlay (1986), shows that additive scales have the intrinsic property of numerical value andn therefore, they imply numerical comparison. This, of course is an inappropriate operation for these data. Note that there is no formal problem with the representation save for the erroneous implication.

Additive Scales for Qualitative Information

A marvelous demonstration of how representational format can be misused in graphs is presented by Mackinlay (1986). Suppose we wish to represent the country of origin of various automobiles. Mackinlay points out that the example shown in Figure 2.10 is clearly inappropriate.

Clearly, the choice of a bar graph is inappropriate for this purpose. But why? What is the problem with Figure 2.10? The bar graph does uniquely specify the desired relationship between manufacture and country: There is no formal problem with the presentation. The problem arises from the intrinsic, additive properties of the lengths of bars. Additive scales have the intrinsic property of numerical value and, therefore, they imply numerical comparison. This, of course, is an inappropriate operation for these data. Finally, the bar graph violates the appropriateness principle that the surface representation used by the artifact should allow the person to work with exactly the information acceptable to the task: neither more nor less. In this case, the bars are capable of carrying more informational structure than the task permits. The excess informational value permitted by the graph is clearly inappropriate: The graph and any artifact should use a representation that is nether too rich nor too poor.

Summary

Cognitive artifacts play an important role in human performance. In this chapter I provide the beginning of an analysis of their critical components by focusing upon three aspects of artifacts:

their role in enhancing cognition (the difference between the system and the personal point of view);

the degrees of engagement that one can experience;

the role of representational format.

The study of artifacts can lead to several advances. First, because so many human activities depend upon artifacts, a full understanding of those activities requires an understanding of the human informationprocessing mechanisms, the internal knowledge of the human, and also the structure, capabilities, and representational status of the artifacts. Second, by understanding the ways in which cognitive artifacts serve human cognition, we may be better able to design new ones and improve old ones.

A major theme of the chapters in this book is the role of artifact, both in support of human activities and also as a tool for the understanding of human cognition. Artifacts play a critical role in almost all human activity. Indeed, as the quotation from Cole with which I opened this chapter suggests, the development of artifacts, their use, and then the propagation of knowledge and skills of the artifacts to subsequent generations of humans are among the distinctive characteristics of human beings as a species. The evolution of artifacts over tens of thousands of years of usage and mutual dependence between human and artifact provides a fertile source of information about both. The study of the artifact informs us about the characteristics of the human. The study of the human informs us of the appropriate characteristics of artifacts. And the study of both the artifact and the human must emphasise the interactions between and the complementarity of the two. The study of the relationship between humans and the artifacts of cognition provides a fertile ground for the development of both theory and application.

Acknowledgments

Much of this chapter reflects joint work and thinking with my research collaborator, Ed Hutchins. I wish to acknowledge my strong debt and gratitude to Ed for his contributions and my thanks for his permission to use them in this way throughout this chapter.

I am grateful to the many people who have commented and aided in this work. In particular, l thank Jack Carroll, Mike Cole, Emmy Goldknopf, and Hank Strub for their comments, intensive critiques, and suggestions. The working group for this conference, of course, was a valuable source of feedback and considerable thanks must go to them and to Jack Carroll for providing the framework for the interaction.

Research support was provided by grant NCC 2591 to Donald Norman and Edwin Hutchins from the Ames Research Center of the National Aeronautics and Space Agency in the Aviation Safety-Automation I'rogram. Everett Palmer served as technical monitor. Additional support was provided by funds from the Apple Computer Company and the Digital Equipment Corporation to the Affiliates of Cognitive Science at UCSD.

References

Bodker, S. (1989). A human activity approach to user interfaces. HumanComputer Interaction, 4, 171195.

Cole, M. (1990). Cultural psychology: A once and future discipline? Paper presented at the Nebraska Symposium, 1989.

Cole, M., & Griffin, P. (1980). Cultural amplifiers reconsidered. In D. R. Olson (Ed.), The social foundations of language and thought. New York: Norton.

Degani, A., & Wiener, E. L. (1990, February). Human factors of JQightdeck checklists: The normal checklist (Contractor's Report). Moffett Field, CA: National Aeronautics and Space Administration, Ames Research Center.

Draper, S. W. (1986). Display managers as a basis for usermachine interaction. In D. A. Norman & S. W. Draper (Eds.), User centered system design: New perspectives in humancomputer interaction (pp. 339352). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hutchins, E. (1986). Mediation and automatization. Quarterly Newsletter of the Laboratory of

Comparative Human Cognition: University of California, San Diego, 8(2), 4758.

Hutchins, E., Hollan, J., & Norman, D. A. (1986). Direct manipulation interfaces. In D. A. Norman & S. Draper (Eds.), User centered system design: New perspectives in humancomputer interaction. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ifrah, G. (1987). From one to zero: A universal history of numbers (L. Bair, Trans.). New York: Penguin Books. (Original work published 1981)

Laurel, B. K. (1986). Interface as mimesis. In D. A. Norman & S. W. Draper (Eds.), User centered system design: New perspectives in humancomputer interaction (pp. 6785). Hillsdale, NJ: Lawrence Erlbaum Associates.

Leont'ev, A. N. (1981). Problems of the development of mind. Moscow: Progress Publishers.

Luria, A. R. (1979). The making of mind: A personal account of Soviet psychology. (M. Cole & S. Cole, Eds.). Cambridge, MA: Harvard University Press.

Mackinlay, J. D. (1986). Automating the design of graphical presentations of relational information. ACM Transactions on Graphics, 5 (2), 110141.

Mackinlay, J. D., Card, S. K., & Robertson, G. G. (1989). A semantic analysis and taxonomy of input devices. Unpublished manuscript, Xerox Palo Alto Research Center.

Newell, A. (1981). The knowledge level. Al Magazine, 2, 120. Also published in Artificial Intelligence, 1982, 18, 87127.

Nickerson, R. (1988). Counting, computing, and the representation of numbers. Human Factors, 30, 181199.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper (Eds.), User centered system design: New perspectives in humancomputer interaction. Hillsdale, NJ: Lawrence Erlbaum Associates.

Norman, D. A. (1988). The psychology of everyday things. New York: Basic Books. Also published in paperback as D. A. Norman. (1990). The design of everyday things. New York: Doubleday.

Norman, D. A., & Hutchins, E. (1990). Checklists. Unpublished manuscript, Department of Cognitive Science, University of California, San Diego.

Norman, D. A., & Hutchins, E. L. (1988). Computation via direct manipulation (Final Report: ONR Contract N0001485C0133). La Jolla, CA: University of California, San Diego, Institute for Cognitive Science.

NTSB. (1989). Aircraft accident report Delta Air Lines, Inc., Boeing 727232, N473DA, DallasFort Worth International Airport, Texas, August 31, 1988 (Report No: NTSB/AAR88/05, Govt. Accession No: PB 89/04. September 26, 1989). Washington, DC: National Transportation Safety Board.

Palmer, S. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. B. Lloyd (Eds.), Cognition and categorisation. Hillsdale, NJ: Lawrence Erlbaum Associates.

Restle, F. (1961). Psychology of judgment and choice. New York: Wiley.

Rumelhart, D. E., & Norman, D. A. (1988). Representation in memory. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), Stevens' handbook of experimental psychology. New York: Wiley.

Stevens, S. S. (1957). On the psychophysical law. Psychological Review, 64, 153181.

Vygotsky, L. S. (1978). Mind in society: The development of higher mental processes. (M. Cole, V. JohnSteiner, S. Scribner, & E. Souberman, Eds.). Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1985). Vygotsky and the social formation of mind. Cambridge, MA: Harvard University Press.

Wundt, W. M. (1916). Elements offolk psychology: Outlines of a psychological history of the development of mankind (Edward Leroy Schaub, Trans.). London: Allen & Unwin.