THINGS THAT MAKE US SMART

DEFENDING HUMAN ATTRIBUTES IN THE AGE OF THE MACHINE

Donald A. Norman

A William Patrick Book

PERSEUS BOOKS
Cambridge, Massachusetts

THREE

THE POWER OF REPRESENTATION

THE POWER OF THE UNAIDED MIND IS HIGHLY OVERRATED. WITHOUT external aids, memory, thought, and reasoning are all constrained. But human intelligence is highly flexible and adaptive, superb at inventing procedures and objects that overcome its own limits. The real powers come from devising external aids that enhance cognitive abilities. How have we increased memory, thought, and reasoning? By the invention of external aids: It is things that make us smart. Some assistance comes through cooperative social behavior; some arises through exploitation of the information present in the environment; and some comes through the development of tools of thought—cognitive artifacts—that complement abilities and strengthen mental powers.

The limits of the average mind are most easily demonstrated by noting the attention paid to those who have managed to overcome them. We pay homage to those who can remember large quantities of information without any external aid. We pay them money to perform before us on the stage, and we clap delightedly when they learn the names of everyone in the room or tell us the cube root of the serial number of a dollar bill held up by someone in the audience or tell us on what day of the week some arbitrary event a hundred years ago fell. We admire these abilities because they are so unusual and so difficult for the average person to perform. Actually, these skills are difficult even for the expert. They take years to perfect; they require the memorization of numerous tables and word lists, and the learning and continued practice of the computational and mnemonic algorithms. More important, however, is that these are unessential skills. The rest of us live quite productive lives without ever acquiring them. We substitute paper and pencil for mnemonic skills, pocket calculators for computational skills, and printed calendars and tables for extensive memorization and mental calculation.

Probably the most important of our external aids are paper, pencil, and the corresponding skills of reading and writing. But because we tend to notice the unique, not the commonplace, few recognize them for the powerful tools that they are, nor does the average person realize what breakthroughs in reasoning and technology were required to invent writing, numerical representations, portable and reliable pens and pencils, and inexpensive, functional writing paper.

Oral cultures, societies that do not yet have a written language and that also lack the mechanical tools of technological cultures, do not share the benefits. These cultures have not developed advanced mathematics or formal methods of decision making and problem solving. The society that does not yet have writing also has less formal schooling. Instead, most education is conducted through apprenticeships, by watching, copying, and being guided by those who know how to do the task being learned. Their need for formal schooling is limited. They haven't developed mathematics or science, formal history, or extensive commercial records because they can't without the aid of artificially constructed artifacts. It is *things* that make us smart.

Two thousand years ago, Plato wrote the collected dialogues in which he presented the views of Socrates on the important issues of those times. Socrates, Plato tells us, argued that books would destroy thought. How could this be? After all, books, reading, and writing are considered to be the very essence of the educated, intellectual citizen. How could one of the foremost thinkers of civilization deny their importance?

Socrates is famous for his dialogues between teacher and student in which each questions and examines the thoughts of the

other. Questioning and examination are the tools of reflection: Hear an idea, ponder it, question it, modify it, explore its limitations. When the idea is presented by a person, the audience can interrupt, ask questions, probe to get at the underlying assumptions. But the author doesn't come along with a book, so how could the book be questioned if it couldn't answer back? This is what bothered Socrates.

Socrates was concerned with reflective thought: the ability to think deeply about things, to question and examine every statement. He thought that reading was experiential, that it would not lead to reflection.*

SOCRATES: *Then anyone who leaves behind him a written manual, and likewise anyone who takes it over from him, on the supposition that such writing will provide something reliable and permanent, must be exceedingly simple-minded; he must really be ignorant of Ammon's utterance, if he imagines that written words can do anything more than remind one who knows that which the writing is concerned with.*

PHAEDRUS: *Very true.*

SOCRATES: *You know, Phaedrus, that's the strange thing about writing, which makes it truly analogous to painting. The painter's products stand before us as if they were alive, but if you question them, they maintain a most majestic silence. It is the same with written words; they seem to talk to you as if they were intelligent, but if you ask them anything about what they say, from a desire to be instructed, they go on telling you just the same thing forever. And once a thing is put in writing, the composition, whatever it might be, drifts all over the place, getting into the hands not only of those who understand it, but equally of those who have no business with it; it doesn't know how to address the right people, and not to address the wrong. And when it is ill-treated and unfairly abused it always needs its parent to come to its help, being unable to defend or help itself.*

PHAEDRUS: *Once again you are perfectly right.*

Socrates was an intellectual, and to him thinking was reflection or nothing. He didn't go for this experiential stuff. The worst kind of writing for people like Socrates would be novels, storytelling. A story engages the mind in an experiential mode, capturing the reader in the flow of events. All such experiential modes—music, drama, and novels—were considered to be the entertainment of the masses, not worthy of serious respect. Socrates worried that reading would be too passive, an acceptance of the thoughts of the writer without the chance to question them seriously.

In the Middle Ages, just the opposite was true. Reading was generally done aloud, often to an audience. It was an active process, so active that Susan Noakes, in her analysis of medieval reading, points out "that it had been recommended by physicians, since classical times, as a mild form of exercise, like walking."

Moreover, Noakes observes that the characteristics of a good novel today were unheard of in earlier times: "Today, many readers take as the hallmark of the good novel the way it propels them to read it continuously, without putting it down, from beginning to end. Readers of many late medieval books would have been forced, on the other hand, to read discontinuously, stopping to puzzle over the relationship between complement and text." (The term *complement* refers to the dialogue provided through the illustrations and marginal comments—illuminations and glosses—sometimes put in by the author, sometimes by the copyist, sometimes by other readers.)

During the Middle Ages, readers were taught the rules of rhetoric and were implored to employ them with each sentence: *mnemonics*, to memorize and learn the material; *allegory*, to find the multiple levels of meaning hidden beneath the literal text; *typology*, to think in historical parallels. No text was thought to be complete without mental elaboration in the mind of the individual reader or debates within the social group that might be listening to the read-aloud text.

Readers in the latter part of the Middle Ages did with books exactly what Socrates had claimed was impossible: They questioned and debated each idea. True, the author wasn't around, but in many ways that made the job more challenging, more interesting. Read a

sentence, question it. Read a page, criticize it. No authors to object. No authors to refute your arguments with the force of their rhetoric. Readers were free to develop their own objections and opinions without interference from meddling authors. Today we may have regressed to match the fears of Socrates: We read too quickly, without questioning or debating the thoughts of the author. But the fault does not lie with the book, the fault lies with the reader.

Cognitive artifacts are tools, cognitive tools. But how they interact with the mind and what results they deliver depend upon how they are used. A book is a cognitive tool only for those who know how to read, but even then, what kind of tool it is depends upon how the reader employs it. A book cannot serve reflective thought unless the reader knows how to reason, to reflect upon the material.

COGNITIVE ARTIFACTS

The cognitive age of humans started when we used sounds, gestures, and symbols to refer to objects, things, and concepts. The sound, gesture, or symbol is not the thing itself; rather, it stands for or refers to the thing: It represents it.

The powers of cognition come from abstraction and representation: the ability to represent perceptions, experiences, and thoughts in some medium other than that in which they have occurred, abstracted away from irrelevant details. This is the essence of intelligence, for if the representation and the processes are just right, then new experiences, insights, and creations can emerge.

The important point is that we can make marks or symbols that represent something else and then do our reasoning by using those marks. People usually do this naturally: This is not some abstract, academic exercise. Suppose Henri had an auto accident and is describing it to his friends. The description might go something like this:

"Here," Henri might say, putting a pencil on the tabletop, "this is my car coming up to a traffic light. The light is green, so I go through the intersection. Suddenly, out of nowhere, this dog comes running across the street." With this statement, Henri places a paper clip on the table in front of the car to represent the dog. "I jam on my brakes, which makes me skid into this other car coming from the other direction. We don't hit hard, but we both sit there stunned."

Henri takes another pencil and lets it represent the second car. He manipulates the pencil representing his car to show it skidding, then turning and hitting the other pencil. Now the tabletop has two pencils touching each other and a paper clip.

"The dog disappears," Henri says, moving the paper clip off the table. "Then the light turns red, but I can't move. Suddenly, this car comes rushing down the side street. It has a green light, but here we are, stuck right in the middle of the intersection. Boom, it hits us, like this"—and Henri uses his finger to show the third car coming from the side and hitting the two pencils, scattering them.

In this scenario, the tabletop, pencils, paper clip, and finger are all used symbolically. They stand for the real objects—the street, the three cars, and the dog. In the listener's head are other symbols to represent the streets and the traffic light. Notice how difficult it would have been to tell this story without the artifacts, without the tabletop, pencils, and paper clip. In fact, you, the reader, may have had some problems following it in this text unless you tried to visualize the scene in your head: What was the path of the dog? Exactly where were the two cars stopped in the intersection? How did the third car travel?

The story on the tabletop helps show some of the simpler properties of artifacts. You can see how they help the mind keep track of complex events. The same representational structure is also a tool for social communication: Several different people can share the tabletop and the story at the same time, perhaps suggesting alternative courses of action. "Look," Marie might say, picking up one of the pencils, "when you saw the dog, you should have gone like this." "Ah, but I couldn't," Henri might respond, "because there was another car there," and he puts yet another pencil on the tabletop. The tabletop becomes a shared workspace with shared representations of the event.

Note what is now happening: People are using the artifacts themselves to reason about alternative courses of action. The representation substitutes for the real event. A problem, of course, is that the representations are abstractions. The pencil may represent the car, but it doesn't have the correct size or mass. It isn't possible to show how fast the real car was going or how much it would skid if the brakes were applied. All this would require more powerful representations. Nonetheless, the representation adds dramatically to the person's power to describe the event. It enables other people to understand better. It makes it easier to analyze alternative actions. It adds power and precision to the memory of the unaided mind.

A good representation captures the essential elements of the event, deliberately leaving out the rest. Pencils don't look anything at all like cars, yet for the purposes of understanding the incident, that difference doesn't matter. A representation is never the same as the thing being represented, else there would be no advantage to using one. The critical trick is to get the abstractions right, to represent the important aspects and not the unimportant. This allows everyone to concentrate upon the essentials without distraction from irrelevancies. Herein lie both the power and the weakness of representations: Get the relevant aspects right, and the representation provides substantive power to enhance people's ability to reason and think; get them wrong, and the representation is misleading, causing people to ignore critical aspects of the event or perhaps form misguided conclusions.

To understand cognitive artifacts, we must begin with an understanding of representation. A representational system has two essential ingredients, shown in Figure 3.1:*

1. *The represented world:* that which is to be represented;
2. *The representing world:* a set of symbols, each standing for something—representing something—in the represented world.

Representations are important because they allow us to work with events and things absent in space and time, or for that matter, events and things that never existed—imaginary objects and

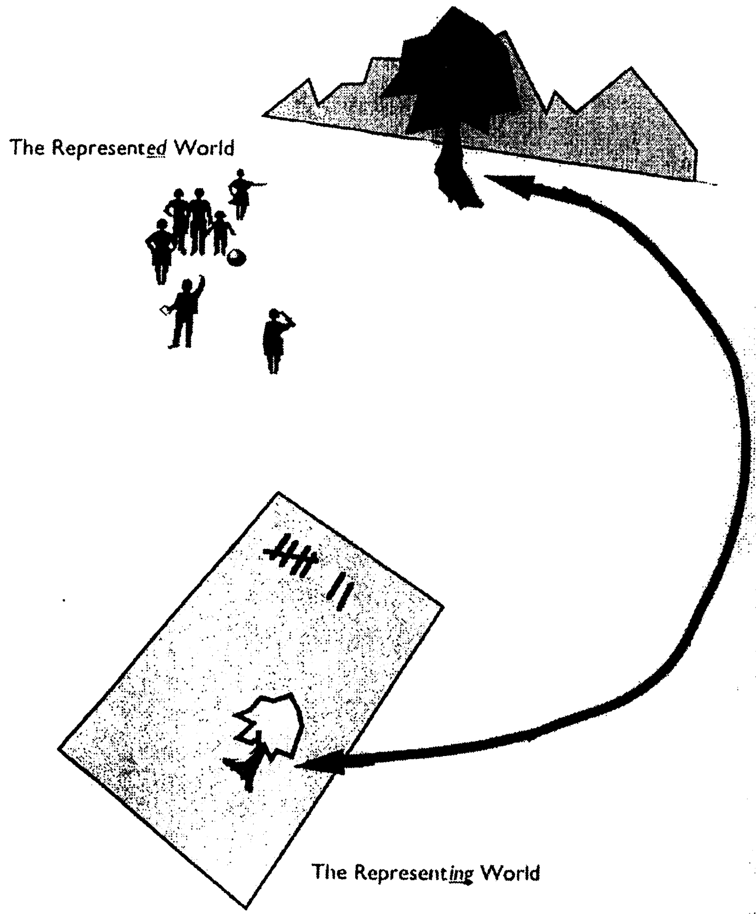


Figure 3.1 The represented and representing worlds. The world to be represented is shown on top—the "represented" world consisting of people, a tree, mountains, and a ball. The "representing" world is shown as marks—symbols—on a sheet of paper. The representing world is an abstraction and a simplification of the represented world. In this example of a representing world, the tally marks each represent one person, and the drawing represents the tree. The other aspects of the real (represented) world are absent from the representing world.

concepts. External representations, especially ones that can be part of a workspace shared with others, require some sort of constructed

device to support them: an artifact. Even if the representation is as simple as stones placed in a special arrangement on the ground or a diagram drawn in the sand, its use as a representation is artificial, with a designated space and often with a verbal explanation to interpret for each object in the representing world just what aspect of the represented world it stands for. We have invented more powerful artifacts than sticks, stones, and sand, of course—artifacts that support a variety of representations, that are long-lasting, portable, easily reproduced and communicable over distances, and capable of powerful computational abilities in their own right.

The critical property of the representations supported by cognitive artifacts is that they are themselves artificial objects that can be perceived and studied. Because they are artificial, created by people, they can take on whatever form and structure best serves the task of the moment. Instead of working with the original idea, concept, or event, we perceive and think about representations that are better suited to match our thought processes. Figure 3.1 serves as an example of this ability to represent knowledge. The figure is itself a representation, one that represents the concept of representation. It contains a representation of yet another artifact (labeled "The Representing World") and the symbols on it, as well as the relationship between that artifact and the world that it represents. Hence, the figure is a metarepresentation: a representation of a representation.

This ability to represent the representations of thoughts and concepts is the essence of reflection and of higher-order thought. It is through metarepresentations that we generate new knowledge, finding consistencies and patterns in the representations that could not readily be noticed in the world. These higher-order representations are very difficult for the unaided mind to discover. In principle, it can be done without artifacts, with just the unaided mind, but in practice, the limited ability to keep track of things in active consciousness severely reduces that possibility.

Once we have ideas represented by representations, the physical world is no longer relevant. Instead, we do our thinking on the representations, sometimes on representations of representations. This is how we discover higher-order relationships, structures, and consistencies in the world or, if you will, in representations of the world. The ability to find these structures is at the heart of reasoning, and critical to serious literature, art, mathematics, and science. The ideal, of course, is to develop representations that

- Capture the important, critical features of the represented world while ignoring the irrelevant
- Are appropriate for the person, enhancing the process of interpretation
- Are appropriate for the task, enhancing the ability to make judgments, to discover relevant regularities and structures

There are many kinds of artifacts. Experiential artifacts have different functions from reflective ones. Experiential artifacts provide ways to experience and act upon the world, whereas reflective artifacts provide ways to modify and act upon representations. Experiential artifacts allow us to experience events as if we were there, even when we are not, and to get information about things that would be inaccessible, even if we were present. A telescope gives us information about something distant in space. A movie or recording lets us experience events distant in time and space. Instruments, such as the gas gauge of an automobile, give us information about states of equipment that would otherwise be inaccessible. Experiential artifacts thus mediate between the mind and the world.

Reflective artifacts allow us to ignore the real world and concentrate only upon artificial, representing worlds. In reflection, one wants to contemplate the experience and go beyond, finding new interpretations or testing alternative courses of action. The process can be both powerful and dangerous. The power comes from the ability to make new discoveries. The danger occurs whenever we fool ourselves into believing that the representation is the reality.

When we concentrate only upon the information represented within our artifacts, anything not present in the representation can conveniently be ignored. In actuality, things left out are mostly things we do not know how to represent, which is not the same as things of little importance. Nonetheless, things not represented fall

in importance: They tend to be forgotten or, even if remembered, given little weight. This is the lesson of Chapter 1: We value what we can measure (or represent).

MATCHING THE REPRESENTATION TO THE TASK

> Solving a problem simply means representing it so as to make the solution transparent. *(Simon, 1981)*

Let's play a game: the game of "15." The "pieces" for the game are the nine digits—1, 2, 3, 4, 5, 6, 7, 8, 9. Each player takes a digit in turn. Once a digit is taken, it cannot be used by the other player. The first player to get three digits that sum to 15 wins.

Here is a sample game: Player A takes 8. Player B takes 2. Then A takes 4, and B takes 3. A takes 5.

> *Question 1:* Suppose you are now to step in and play for B. What move would you make?

This is a difficult problem for several reasons, all traceable to the way I described the problem—to the representation. The task is described as a problem in arithmetic. To figure out what move to make, you have to consider what possibilities both you and A have for winning. This requires a lot of calculation to determine which triples of digits sum to 15. There are few aids to memory, so it is difficult to keep track of which player has chosen which digits, which ones remain. I have deliberately presented the game information to you in a representational form that is awkward to use: The moves are listed sequentially, making it difficult to see just which digits A and B each have. Although the arithmetic is simple, keeping track of all the possibilities while doing the arithmetic makes the game difficult.

Now let's play a different game, this one the children's game of ticktacktoe (also called "naughts and crosses" and "three in a row"). Players alternately place a naught (the symbol O) or a cross (the symbol X) in one of nine spaces arranged in a rectangular array (as shown in the following illustration). Once a space has been taken, it cannot be changed by either player. The first player to get three symbols in a straight line wins. Suppose player A is X and B is O, and the game has reached the following state:

X	O	X
	X	
O		

Question 2: Suppose you are now to step in and play an O for B. What move would you make?

Unlike the game of 15, this time the task is easy. This is a spatial game, not one of arithmetic. To see what is happening, just look at the board: A quick glance shows that A is all poised to win (by completing a diagonal line of Xs) unless blocked by an O in the lower right-hand corner.

Question 1 was hard because the game of 15 requires reflection, with few external aids. Question 2 was easy because it could be answered experientially, perceptually: No computation required—just look at the board and see the proper move.

But note, the two games are really the same. If you think of the nine digits of the game of 15 arranged in a rectangular pattern, you see that it is identical to the game of ticktacktoe:

4	3	8
9	5	1
2	7	6

Remember the moves in the game of 15? A had selected 8, 4, and 5; B had selected 2 and 3:

X	O	X
	X	
O		

4	3	8
	5	
2		

Player B, you, had better select the digit 6, in the lower right corner.

The two games are what we call "problem isomorphs" (from the Greek *iso*, for "the same," and *morph*, for "form").* Technically, questions 1 and 2 are identical, but as the example shows, the choice of representation changes the task and the difficulty dramatically.

Although the spatial representation of ticktacktoe is much easier for people to play than the arithmetic one of 15, for computers the arithmetic representation is much easier. A computer program to solve ticktacktoe spatially would have to figure out whether the Xs and Os were on a straight line: It would have to solve the trigonometric relationships among the points. How much easier for us, since we can simply look and see: The human perceptual system is designed for this task. We find the method used by the computer difficult and cumbersome, although we are quite capable of programming the computer to follow the method. In return, it is very difficult for the computer to do the perceptual processing.

This example illustrates two points. First, the form of representation makes a dramatic difference in the ease of the task, even though, technically, the choice does not change the problem. Second, the proper choice of representation depends upon the knowledge, system, and method being applied to the problem. In this case, the method hardest for the human is easiest for the computer, and the method easiest for the human is hardest for the computer. The example therefore also illustrates the differing yet complementary powers of human and computer information processing.

The power of a representation that fits the task shows up over and over again. Bad representations turn problems into reflective challenges. Good representations can often transform the same problems into easy experiential tasks. The answer so difficult to find using one mode can jump right out in the other.

Consider the task of planning an airline trip between two cities. Suppose I want to travel from my hometown of San Diego (California, U.S.A.) to London (England, U.K.). The way in which airline information is typically presented is shown in the accompanying table: the format employed by the *Official Airline Guide* (the OAG), perhaps the most widely used source of airline information for professional travelers within the United States.

	1131	SAN	0820+1	LGW	AA	2734	FCYBM	D10	1
		AA	2734	CHG PLANE AT DFW					
X12	1805	SAN	1425+1	LGW	BA	284	FJMSB	D10	1
	2100	SAN	2030+1	LHR	TW	702	FCYBQ	*	2
		TW	702	EQUIPMENT 767 LAX-L10					

This excerpt from the *Official Airline Guide Worldwide Edition* (November 1990) shows three flights between San Diego and London. Reading left to right, the top line shows a flight leaving at 11:31 AM from San Diego (SAN) and arriving at 8:20 AM the next day (the +1) at London's Gatwick airport (LGW). This is American Airlines flight 2734, with five classes of service (FCYBM), using a DC-10 and making one stop. The second line states that the flight has a plane change at the Dallas/Fort Worth (DFW) airport. The third line shows a flight that goes every day except Monday and Tuesday (X12): British Air flight 284, with one stop. (The arrival time, 1425, is given in European, twenty-four-hour time: 1425 is 2:25 PM.) The fourth line is a TWA flight that makes two stops and lands at London's Heathrow airport (LHR) at 8:30 PM, and the last line indicates that between San Diego and Los Angeles (LAX), the flight is on a Boeing 767, but from Los Angeles, it will be a Lockheed L-1011.

The OAG's presentation is designed to pack as much information as possible into the smallest amount of space. The monthly worldwide edition is printed in tiny type on over fifteen hundred large pages. Although the publishers have done a creditable job of making the entries usable, the user still has to do considerable mental processing and copying of information. The publishers have unwittingly transformed the selection of a flight into a reflective task.

Suppose my desire is a flight that arrives in London late in the afternoon. At first glance, the OAG format would appear to be perfect because column four shows arrival time directly: I need only scan the arrival times for the one most convenient. This would suggest the TWA flight that leaves San Diego at 9:00 in the evening and arrives in London at 8:30 in the evening the next day. Wonderful: I get on the plane, read a book, have a brief sleep, and when I get to London, clear customs, and get to my hotel, it is time for bed.

But is this true? Closer reading indicates that I had better not go to sleep right away: There is a plane change at Los Angeles. And there are two stops: Los Angeles and where? Is this flight longer than the others?

While I want to arrive late in the afternoon, I do not want to spend several extra hours in traveling. So let me see which flight has the shortest duration. Now we see how the display affects the task: It was easy to search for a flight by arrival time, but it is not so easy to find a flight by duration. I have to do some arithmetic, subtracting departure times from arrival times, which is not easy given that they are on different days and that there is a seven-hour time difference between the two cities.*

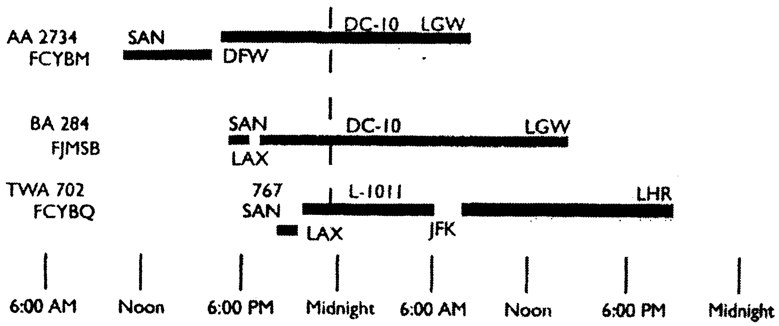
To do the arithmetic, I must invent an intermediate notation. Whenever an arrival occurs on the day following departure, add twenty-four hours to the arrival time: 1:00 AM the next day is 25:00; 8:30 PM the next day is 44:30 (20:30 + 24:00). This puts arrival times in a format that makes it easy to find the differences in time between departure and arrival, and when I subtract the seven-hour time difference, I get the following durations:

Flight	**Depart**	**Arrive**	**Difference**	**Duration (Diff. −7)**
AA 2734	11:31	32:20	20:49	13:49
BA 284	18:05	38:25	20:20	13:20
TW 702	21:00	44:30	23:30	16:30

The TWA flight takes almost three hours longer than the others, so even though it arrives at a good time, the tradeoff of an extra three hours travel time is not acceptable. Note that this new arrival notation makes it easier to do the arithmetic but harder to figure out what time the flights arrive. One table makes it easier to choose the shortest flights; the other table makes it easier to check what time the flights arrive. Of course, I could simply add an extra column to the first table giving duration, but in the crowded pages of the OAG, there is simply no room for any information that can be derived.

All of this comparing and planning is reflective. I look at the information given in the OAG and ask questions of it, restructuring the information and performing new computations. This is an excellent example of the power of reflection, except it shouldn't be needed. A different form for presenting this information would change the task to an experiential one, where the answers would appear through inspection.

The OAG uses a table to present its information, and this made some of the comparisons difficult. Stephen Casner has shown how the graphic presentation of scheduling information can simplify some of the decision making in flight planning.* So, borrowing from his work, let us examine these three flights.



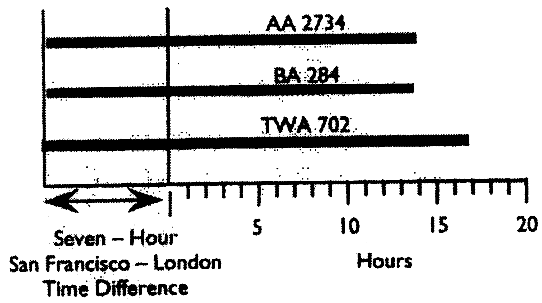
All times are local. Note seven hour time difference between San Diego and London.

This graphic display does appear to make some of the information about the flights much easier to comprehend. It shows all the information in the OAG display, plus more information about the stops. Durations of the three flights are indicated by the lengths of the lines connecting departure and arrival times. The notation also provides a simple way to represent plane changes (the "steps" in the lines) as well as the amount of time spent at stops (the gaps in the lines). The AA flight has a stop, plane change, and delay at Dallas/Fort Worth (DFW). The BA flight stops with a delay but no plane change at Los Angeles (LAX). The TWA flight has a plane change at Los Angeles and a stop with no plane change but a long layover in New York at Kennedy airport (JFK).

Which is the shortest-duration flight? We have already discussed the difficulties of answering this question from the OAG table. In theory, the answer should be easy to discover in the graphic display because all that needs to be done is to compare the lengths of the three lines. In practice, as you can readily see for yourself, the comparisons are not so easy to make. To compare flight durations, you must mentally line up the lines to determine which is the shortest. This example shows that perceptual pro-

cessing alone does not guarantee success. Whenever mental transformations are necessary in order to make comparisons of the configurations, graphic representation presents the viewer with a difficult task.

The comparison is finally transformed into an experiential task by lining up the starting points: Now you can just look and immediately see the answer.



The TWA flight is the longest, and the other two have approximately the same duration. Line up the starting points, remove some distracting clutter, and we have an easy task: The task that used to require arithmetic in the table or mental superposition of lines in the other graphic display can now be done by simply scanning the diagram to find the line that sticks out most (for the longest flight) or least (for the shortest flight).

This new representation also has another advantage. Because the flight times are given in local time, the flight duration is seven hours less than the lengths of the lines would suggest. To determine the actual amount of time on the airplane, you have to subtract the seven-hour time difference. But with this new graph, even the subtraction task is easy. We simply need to move the starting point for the comparison of the lines seven hours to the right, as illustrated on the diagram.

What do we conclude about the appropriate representation for a task? The answer depends upon the task. To know the class of service or the type of airplane, text is superior. To know the exact minute of departure (11:31, say), the printed number is needed. To make a rapid comparison of flight duration, the graphic display is best.

Now that we have seen how graphic displays can simplify the task, what should the OAG do? I recommend that it continue as it is. The publishers of the OAG have a different task from the users. They need to make available as much relevant information as possible. Space is clearly of great importance, and the textual presentation the OAG provides is both efficient and relatively usable. The OAG has changed its format over the years to improve the usability. The graphic display takes up much more space than the tabular one. The most appropriate format depends upon the task, which means that no single format can ever be correct for all purposes.

Someday, not too far in the future, all the information will be available on electronic devices whose displays will allow the same information to be presented in a variety of ways: different layouts for different needs. Wouldn't it be nice to be able to see a listing of all flights organized by time of arrival or by duration of flight or by price of the ticket? Displays that let us switch instantly from numerical format to graphic, depending upon the task? And that let us move among all the formats until all the information needed was available, neither too much nor too little?*

HOW REPRESENTATIONS AID INFORMATION ACCESS AND COMPUTATION

There are two major tasks for the user of an information display:

1. Finding the relevant information;
2. Computing the desired conclusion;

In our examination of information displays, we can note what kinds of assistance the displays provide for these two aspects: What aids are given to help the person's access to the appropriate information? What aids are given to help with the computations?

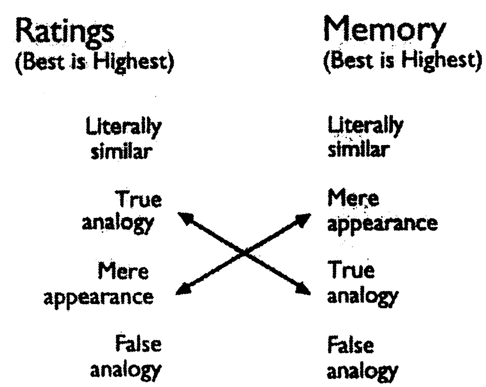
Consider this example, taken from the draft of a Ph.D. dissertation:

> They found that while subjects would rate the analogies, from best to worst, as literally similar, true analogy, mere appearance, and false analogy, their recall for stories, from best to worst, was literally similar, mere appearance, true analogy, and false analogy.

Why is the sentence so unintelligible? Just consider what you have to do to figure out what it means:

> Best to worst, um, best for analogies is literally similar. And stories, best is literally similar. Gee, those are the same. Let's see, next best for analogies is, um, true analogy. Next best for, um, stories, is, um, mere appearance. Hmm, that's different.

The task of understanding the sentence is an example of reflective thought, unnecessarily reflective, for the information in the sample sentence can also be displayed in a chart like this:



This diagrammatic format uses several techniques to aid the reader:

Needs of the Reader	Provided By
Finding critical comparisons	The lines with arrowheads make the significant comparisons easy to find.
Finding the relevant variables to be compared	Lining up the items.
Remembering the ranking of conditions	Ordering them vertically—the higher, the better.
Comparing the different conditions	Putting the four conditions into two vertical columns, lined up horizontally.
Search and computation	Lining up the right borders of the left list and the left borders of the right list.

The diagram contains exactly the same information as the original, written sentence, but in a form much easier to understand. The tabular arrangement has made both the search and the computations—which, in this case, are comparisons—simpler. Is this a graphic, a chart, or a table? It doesn't matter: It uses an appropriate display format for the task.

Example: Medical Prescriptions

Medical prescriptions are growing ever more complex, with many people being required to take numerous medications daily. How well do people cope with following their prescriptions? Not very well. Several surveys have shown that between 10 percent and 30 percent of the people studied were unable to determine how much medication they should take at any time. In one study, arthritic patients were asked to bring their medication to the experimenters and then, with the bottles and containers in front of them, to write down their daily medication. They were allowed as much time as needed. The results showed great difficulty in doing the task, with an average error rate of about 14 percent. It should hardly be a surprise that the more medications prescribed, the greater the percentage of error. Those

people who were prescribed the largest amounts (seven or more drug dosages a day) made both the highest absolute number of errors and the highest percentage of errors: slightly over 30 percent.

The problems of keeping track of medication are well known. In my local drugstore, several different memory aids are available, all aimed at making it easier to keep track of pill taking. All of them are "pill organizers," boxes divided into compartments labeled by time of day, day of week, or both day and time. In principle, these should be beneficial to patients, once the pills are loaded into the proper compartments. Alas, loading the boxes is not very easy. The boxes do not overcome the fundamental problems of interpreting the prescriptions.

The same study that revealed the 30 percent error rate in taking pills also examined how well patients could use these organizers. Again, the answer is not very well. One patient put twice the recommended medication into one of the boxes. Another box tended to be loaded properly, but the average loading time was over nine minutes! These organizers do not appear to work, not when they still lead to errors or when they require so much time to be loaded with pills.

This is an area crying out for help. Solutions, to be effective, must include and support the needs of all the people involved with the prescription: the patient, the physician and physician's aides, and the pharmacist. This issue can truly be a matter of life or death.

One of the problems is that the prescriptions themselves are not written from the patient's point of view. Consider the following medical prescription from the work of psychologist Ruth Day, a prescription that was given to a patient following hospitalization for a mild stroke.

Inderal	—1 tablet 3 times a day
Lanoxin	—1 tablet every a.m.
Carafate	—1 tablet before meals and at bedtime
Zantac	—1 tablet every 12 hours (twice a day)
Quinaglute	—1 tablet 4 times a day
Coumadin	—1 tablet a day

This set of instructions is very difficult to follow. Speaking of the patient, Day reports:

> Over the next few days, he had difficulty remembering what pills to take, as well as what pills he had already taken. It would be easy to blame the patient: after all, he was 81 years old and had just had a stroke. However, he was highly intelligent, was still working full time (and had even begun a new and demanding career a few years earlier), was not otherwise disoriented, and was highly motivated to return to work and an active life style. (Day, 1988, p. 276)

The physician's list, as presented here, is neatly organized, precise, and easy to read. It is very similar to the format used for most prescriptions in the United States. The problem is that it is set up for the wrong task. The representation is appropriate from the point of view of the prescribing physician: Figure out what the patient needs and write it down. But it simply does not lend itself to usage. The list is organized by medicine, which makes it easy for the physician and the pharmacist to look for any medication and see how it was prescribed. But the patient needs it organized by time: Given the time of day, what actions should be performed? Day tested the usability of the prescription by having people try to answer the following two questions:

1. It is lunchtime (noon). Which pills should you take?
2. If you leave home in the afternoon and will not be back until breakfast time the next day, how many pills of each type should you take along?

As you can determine for yourself, it is not easy to answer these questions. The problem is that following this prescription is a reflective task, when it should be an experiential one. Reflection requires mental effort, something a sleepy, ill patient is apt to have trouble with. To fit the needs of the patient, the prescription should be organized by time of day. Note this organization is still appropriate for the physician or pharmacist. Here is Day's suggested presentation of the information:

	Breakfast	Lunch	Dinner	Bedtime
Lanoxin	✔			
Inderal	✔	✔	✔	
Quinaglute	✔	✔	✔	✔
Carafate	✔	✔	✔	✔
Zantac		✔		✔
Coumadin				✔

Notice that with Day's solution, the items can be organized by time of day (the columns) or by medication (the rows). The users simply scan the list by whichever starting point they prefer. A simple change in representation transforms the earlier, difficult reflective task into a much simpler experiential one. Day's experiments showed that the matrix form was not only easier but also conducive to more accurate interpretation than the original (and more common) format.

As Day points out, the matrix has major advantages over lists. Lists are organized by one factor (medication name, in this example). Matrices allow several different dimensions to serve as organizational keys: in this case, medication name or time of day. Whenever several different needs have to be met, a matrix is apt to be superior.

The matrix organization aids both search and computation. In the original prescription, in order to answer the question "How many pills are taken at lunchtime?" the entire list had to be read and then interpreted. The computations were reasonably extensive, even if simple in nature. With the matrix, the computation merely involves scanning down the "Lunch" column and counting. Once again, the proper choice of cognitive artifact aids the task by transforming it from reflection to experiencing, simplifying the operations that must be performed to reach the desired answer.

REPRESENTING NUMBER

Imagine trying to multiply using Roman numerals—say, CCCVI times CCXXXVIII. It's possible, but very difficult. The same numbers written in modern notation—306 times 238—present an easier challenge. The modern Arabic notation lends itself to efficient algorithms for arithmetic, although to multiply these three-digit numbers will require writing something down. In Roman numerals, each symbol stands for a quantity, and in their original form (where *4* was written as "IIII" and *9* as "VIIII"), it doesn't even matter in what order you write the symbols: CCXXXVIII is the same quantity as ICVXIICXX.* With our modern Arabic numbers, we also use the same symbols repeatedly, but the meaning of each symbol depends upon its location. That's why we need the *0* in *306*: The *3* means "300" only in the third position from the right. Roman numerals had no need for a zero.

The choice of representation for numbers makes a big difference in how easy or hard it is to do certain operations. Arabic numbers are not always the best choice for representation.

One of the oldest forms of representing numerical quantities—tally marks—is still the best form when we need a way of counting something rapidly. To count an item, I make a short vertical line, I; adding a second one, II; a third, III; and a fourth, IIII: one new mark for each new item.

Tally marks are easy to make and easy to compare, which is why they are still in use today. Roman or Arabic numerals are much more difficult. Why? Because tally marks are *additive:* With an additive representation, if I wish to increase the value of a previous symbol, I simply add extra marks to the symbol already there. Thus the symbol for 3 (III) readily becomes the symbol for 4 (IIII). Nothing already present has to be changed.

Contrast this with Arabic numerals, which are *substitutive:* With a substitutive representation, if I wish to increase the value of a previous symbol, I must substitute a new symbol for the previous one. To increase the value by 1, I have to cross out the previous value and write the new one. The symbol 1 becomes ~~1~~ 2, and then ~~1~~ 2 becomes ~~1~~ ~~2~~ 3.

Of course, there are other differences between Arabic and tally representations besides the ease of making the marks. Arabic numbers are harder to make than tally marks, but easier to read and to use for computations. To make it easier to read tally marks, we usually modify them somewhat, so we group them into fives, generally like this: ~~IIII~~I.

Additive notations have another important property: The size of the representation is proportional to the value of the number. So tally marks also serve as a graph. (See Figure 3.2)

These examples show that changes in representation often provide us with tradeoffs: One aspect of the task gets easier while another gets harder. Thus, while counting, tally marks are easier to make than Arabic numbers and easier to compare, especially if the number of objects is relatively small. But for doing calculations, tally marks are much harder to use than Arabic numbers.

Addition Is Easier with Roman than Arabic Numerals

Strange as it may seem, it is easier to add two numbers using Roman numerals than using our everyday Arabic numerals. Today students have to learn the arithmetic table: They start by learning the ten arbitrary symbols for the ten digits, then learn place notation to know that 46 is the same as four 10s plus six 1s. Next, they must memorize the sums for the forty-five possible pairs of numbers. (The ten digits, 0 through 9, have 100 possible combinations. But because of the property called reflexivity—e.g., 4+5 = 5+4—and the ease of adding zero, only forty-five combinations need to be learned.) Finally, students have to learn what to do if there is a carry from one column to the other. All this takes a surprisingly long time to learn.

Roman students simply had to learn the Roman characters for digits—seven different characters go from 1 to 1,000 (I, V, X, L, C, D, M). After that, to add two numbers, they simply combined the symbols together and reordered them, all similar symbols together, the symbols with the greatest value on the left. Then they applied some simplification rules (one rule for

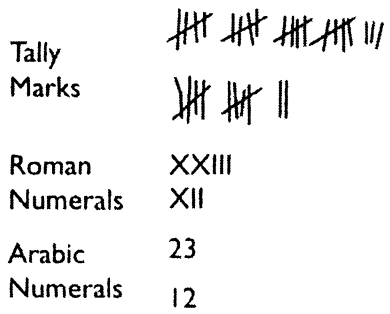


Figure 3.2 Comparing *23* and *12* with tally marks, Roman numerals, and Arabic numerals. Tally marks are an additive representational system in which the length of the representation is proportional to the value being represented. The values of additive representations can be compared experientially. A glance at the figure shows that one value is roughly twice as much as the other. Roman numerals are a modification of tally marks, and so they too can have an additive character, with their length related to the value. A glance at the figure shows that the top value is greater than the bottom one, but the ratios of the lengths of the numerals do not accurately reflect the ratios of the numerical values. Within each place position, Arabic numbers are a substitutive representation, and as this example shows, for small numerical differences, the length of the representation does not provide any information about its value.* The values of substitutive representations have to be compared reflectively, through mental computation.

each symbol) that tell how small symbols combine to make bigger ones (e.g., IIIII = V, VV = X). This is a lot less to learn than the ten symbols of Arabic numerals, the forty-five arithmetic combinations, and the rules for place notation and carry. It's a lot easier too.

Example: 306 + 238

The problem: CCCVI + CCXXXVIII

Combining the symbols: CCCVICCXXXVIII

Reordering the symbols: CCCCCXXXVVIIII

Simplifying gives the answer: DXXXXIIII

The answer, in Arabic numerals: 544

No arithmetic sums have to be known, just how to combine, reorder, simplify, and read the symbols. Roman children had it easier than today's—at least, until they tried to multiply or divide.

Additive and Substitutive Representations

The distinction between additive and substitutive dimensions is important, one that makes a big difference in the ease of understanding graphic representations. The distinction is not well respected by many graphic designers.

Look at Figure 3.3, my redrawing of a chart that was published in a newspaper. The chart uses different kinds of shading superimposed on a map of the United States to indicate what percentages of homes exceed the recommended level of radon, a radioactive gas that we all wish to avoid. Alas, the chart uses the wrong representation: A substitutive representation (different types of shading) is used to represent additive information (percentage of homes that exceed the recommended level of radon). Look at that graph and try to figure out where in the United States radon is most prevalent, least prevalent, and at an average value. The task is hard because the shadings are arbitrary: You have to keep going back to the legend to remember whether a particular shading represents a greater or lesser value than another. The choice of shading transforms this into a reflective task when it should be experiential.

The proper way to draw the figure is to use an ordered sequence of density (an additive scale) to represent percentages (an additive dimension). Try the same task (to determine where radon is most prevalent, least prevalent, and at an average value) with the map shown in Figure 3.4.

I have deliberately introduced a problem with the representation in Figure 3.4 to emphasize the point about the importance of representational format. If you look at the map, it appears that the northwest part of the United States has a very low concentration of radon. That's because that portion of the map is white,

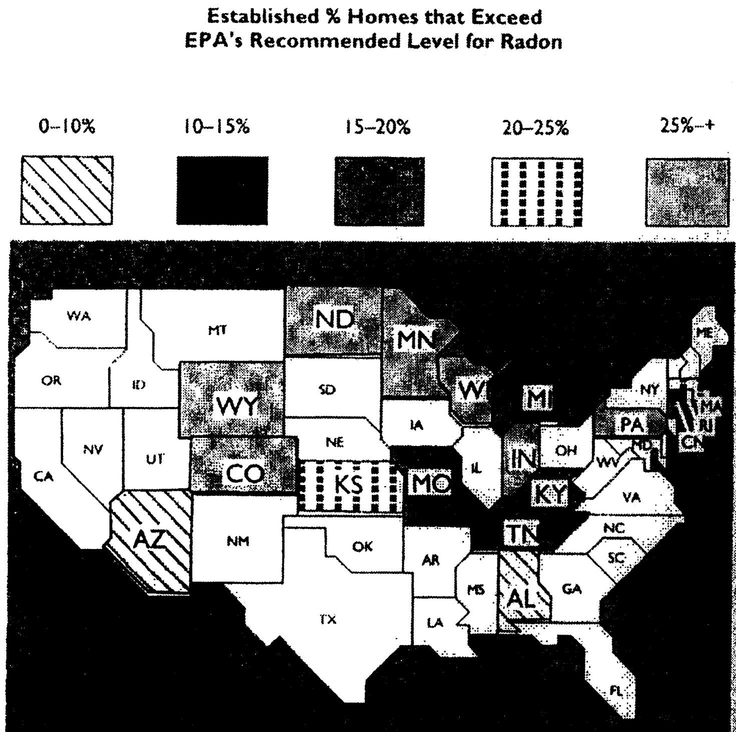


Figure 3.3 An unnatural mapping. Here percentage (which is an additive dimension) is represented by a substitutive scale—different shadings. And where the shadings can be ordered along an additive scale, the ordering conflicts with the ordering of percentages. (Redrawn from a figure in the *Los Angeles Times* [September 13, 1988], p. 21.)

and on the scale of density, white falls to the left of (less than) the 0–10% density. In this case, however, white actually represents those states for which there are no data. A better way to make this graph would be to delete the names of states for which there is no information. I left them in because the natural misinterpretation helps make the point about the impact of representational format.

Figure 3.3, with an inappropriate use of substitutive shading to represent additive percentages, makes the comparison task one of reflection. Figure 3.4, which uses an additive representation of

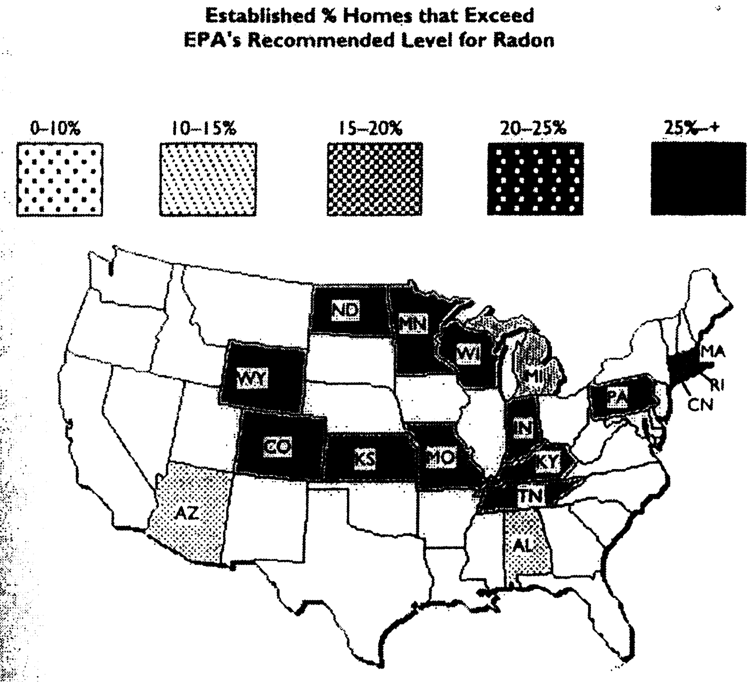


Figure 3.4 A natural mapping. Here the map in Figure 3.3 has been redrawn so that percentage (which is an additive dimension) is represented by an additive scale—ordered densities of shading. Now the density ordering matches the percentage ordering. (Redrawn from a figure in the *Los Angeles Times* [September 13, 1988], p. 21.)

shading to represent the additive percentages, allows the task to be performed experientially.

Color (hue) is frequently used to represent density or quantity, especially in geographic maps, satellite photographs, and medical imagery. But hue is a substitutive representation, and the values of interest are usually additive scales. Hence hue is inappropriate for this purpose. The use of hue should lead to interpretive difficulties. Many colorful scientific graphics, usually generated by a computer, use different hues to represent numerical value. These graphics force the viewer to keep referring to the legend that gives the mapping between the additive scale of interest and the hues. Density, saturation, or brightness would provide a superior representation.

NATURALNESS AND EXPERIENTIAL COGNITION

The several examples throughout this chapter illustrate an important design principle—naturalness:*

> **Naturalness principle:** Experiential cognition is aided when the properties of the representation match the properties of the thing being represented.

I return to these and other design principles in Chapter 4, but let us explore some of the implications of the principle. We humans are spatial animals, very dependent upon perceptual information. Representations that make use of spatial and perceptual relationships allow us to make efficient use of our perceptual systems, to think experientially. Representations that use arbitrary symbols require mental transformations, mental comparisons, and other mental processes. These cause us to think reflectively, and although in many cases this is appropriate and necessary, it is more difficult than experiential cognition. It is also subject to error, especially when people are under high stress.

Mappings are the relationship between the format of the representation and the actual things being represented. They are easier, more reliable, and more natural with well-designed perceptual or spatial representations than with abstract representations. This leads to the second principle:

> **Perceptual principle:** Perceptual and spatial representations are more natural and therefore to be preferred over nonperceptual, nonspatial representations, but only if the mapping between the representation and what it stands for is natural—analogous to the real perceptual and spatial environment.

Graphs are often superior to tables of numbers because in a graph, the height of the line is proportional to the value, so you can compare the different values perceptually. If all you have to work

with is numbers, then you have to do some mental arithmetic to see the relationships. Graphs are not always superior to tables, mind you: only when the task is appropriate for perceptual judgments.

We have already seen that to decide whether one number is larger than another, tally marks are superior to Arabic notation because the length of the line of tally marks is directly proportional to the value represented. You might very well wonder what the fuss is about here. The comparison of *23* and *12* seems natural and straightforward: *23* is larger than *12*, what's the big deal? The big deal is that numbers are really not natural. They are reflective tools, not experiential ones. Powerful, essential tools for thought, but nonetheless, reflective. When Arabic numerals were first invented, it was only the most highly educated people who could master them, and their use was debated and, in some cases, prohibited. Even today, it takes years of study in childhood to become proficient at arithmetic, years of practice that later on allow adults to regard the comparison as simple and natural. Anything that requires that much study is not natural.

Try these two numerical comparisons:

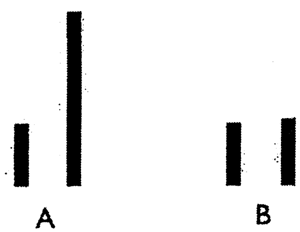
A: Which number is larger?
284 912

B: Which number is larger?
284 312

Much to many people's surprise, experimental psychologists discovered that people can answer problem A faster than they can B. The time differences are small, small enough that you can't notice it yourself, but large enough to be easily measured through the appropriate experiments. Even though we experience both comparisons as immediate and effortless, B takes more time and effort than A. Why is this? So far, the only answer that accounts for all the findings is that the Arabic numbers are translated into a perceptual image—an additive representation—before the comparison is performed. The greater the perceptual difference, the easier the task.

From a logical point of view, the two problems A and B seem equally easy. The point to learn from this is that real psychology is not the same as folk psychology or logic. People have their own commonsense views of how their minds work—a folk psychology. Alas, people are only aware of their conscious experiences, which is a mere fraction of what really goes on. Commonsense views of psychological behavior are reasonable, sensible, and in agreement with everyday experience. Logical views are also reasonable and sensible. Both common sense and logic are often wrong.

Here is the perceptual analog of the numerical comparison. In each problem, the lines are drawn to scale so that they match the earlier questions A and B. Try these two graphic comparisons: In A and B, which line is longer?



The perceptual comparisons are simple and direct, but here, just as with the earlier questions A and B, comparison A can be done more rapidly than comparison B. But the graphic form of the comparison is easier and faster than the numerical one: The first is experiential, the other reflective. To compare the lengths of two lines, you don't even have to know anything about numbers: The perceptual system handles the chore, simply and efficiently.

Representations that match our perceptual capabilities are simpler and easier to use than those that require reflection. Moreover, under a heavy work load (perhaps under severe stress, danger, and time pressure), representations that require reflection—such as the use of Arabic numbers—are not used as rapidly and efficiently as those that can be used experientially, through simple perceptual comparisons. Where simple comparisons are required, graphic notation is superior. But where exact numerical values are required or where numerical operations must be performed, Arabic notation

is clearly superior—that is why it is the standard notation used today.

The power of cognitive artifacts derives from the power of representation. The form of representation most appropriate for an artifact depends upon the task to be performed. The same information may need to be represented differently for different tasks. With the appropriate choice of representation, hard tasks become easy.