Response time in man-computer conversational transactions

by ROBERT B. MILLER

International Business Machines Corporation
Poughkeepsie, New York

INTRODUCTION AND MAJOR CONCEPTS

The literature concerning man-computer transactions abounds in controversy about the limits of "system response time" to a user's command or inquiry at a terminal. Two major semantic issues prohibit resolving this controversy. One issue centers around the question of "Response time to what?" The implication is that different human purposes and actions will have different acceptable or useful response times.

This paper attempts a rather exhaustive listing and definition of different classes of human action and purpose at terminals of various kinds. It will be shown that "two-second response" is not a universal requirement.

The second semantic question is "What is a need or requirement?" In the present discussion, the reader is asked to accept the following definition: "A need or requirement is some demonstrably better alternative in a set of competing known alternatives that enable a human purpose or action to be implemented." This definition intentionally ignores the problem of value versus cost. It is not offered as a universally useful definition of "need." It does enable us to get into a systematic exposition of problems, alternatives and implications. A value-based definition, in contrast to the rational one given here, inevitably leads to a vicious regress that dead-ends only with the agreement that all that humans *really* need are food, water, and a place to sleep.

Another point of view, compatible with the present one, is that need is equivalent to what is demanded and what can be made available; need, therefore, is a cultural and technical outcome. It is the outcome of many vectors, at least one of which is what the marketplace has to offer and the number of Joneses who have one, too.

Operating needs and psychological needs

An example of an operating need is that unless a given airplane's velocity exceeds its stall speed, the airplane will fall to earth. Velocity above stall speed is an undebatable operating need. In a superficially different context, it is a "fact" (let's assume we know the numbers) that when airline customers make reservations over a telephone, any delays in completing transactions above five minutes will reduce their making future reservations with this airline by 20%. A related form of need in this context is that the longer it takes to process one reservation, the larger the number of reservation clerks and reservation terminals that will be required. These are just two examples of the context of operating needs. This report will not look into the problems of operating needs except to mention when they may be more significant than a psychological need. The following topics address psychological needs.

Response to expectancies

Psychological "needs" (in the information processing context) have two major forms, with overlap. One is in the nature of response to an expectation. If you address another human being, you expect some communicative response within x seconds—perhaps two to four seconds. Even though his response may not have the message context you want, you expect him to respond within that time in some fashion, if by no more than a clearing of the throat or a grunt. In conversation of any kind between humans, silences of more than four seconds become embarrassing because they imply a breaking of the thread of communication. This is similar to a phone line going dead. Conditioning experiments (which, of course, should be intro-

duced only with great caution in the context of cognitive activities) suggest an almost magical boundary of two-second limits in the effectiveness of feedback of "knowledge of results," with a peak of effectiveness for very simple responses at about half a second. There is much evidence to suggest that two seconds in human behavior is a relatively long time. Of course even the lower animals can be conditioned (acquire expectancies) to delays, although as the delay is extended the reliability of the performance rapidly deteriorates. The parameters differ for different species.

These points are made only to suggest that the behavior of organisms is time-dependent, and that time spans in the order of one to ten seconds have significance for some forms of behavior involving information transactions with an environment.

Activity clumping and psychological closure

There is a second class of psychological need in communications. This need recognizes that humans spontaneously organize their activities into clumps that are terminated by the completion of a subjective purpose or subpurpose. When I search in a phone book for a telephone number with which to dial a person I want to talk with, I have a sense of temporary completion when I find the telephone number. I have another when I have completed dialing the number. I will more readily tolerate an interruption or delay after such a completion than during the activities preceding this completion. Psychologists call this subjective sense of completion a "closure" and that is the term used henceforth in this report. The rule is that more extended delays may be made in a conversation or transaction after a closure than in the process of obtaining a closure.

Human short-term memory

Here is a rationale for this phenomenon. Performing any task calls for holding a body of information in mind—I call this short-term memory. When I am looking up the telephone number, I am holding in mind the image of the name I am searching for as well as the goal—which is locating this name in the list. When I shift from temporarily memorizing the telephone number to dialing it, short-term memory is holding this set of digits and the goal action of completing the dialing. An interruption or delay in achieving a goal usually results in a degree of frustration. The longer a content must be held in short-term memory, the greater the chances of forgetting or error. Thus, on both counts (short-term memory and goal aspiration), waiting times within a clump of activities have deleterious effects. A psychological closure results in at least a partial purging of short-term memory or the internal activities that support it.

In very complex problem solving, short-term memory is heavily filled. It is becoming clear in the psychological literature that the degree of complexity of problems that can be solved by a human is dependent on how much information (and in what form) he can hold in short-term memory. Human memory is never passive. Spontaneous noise from within the thinking system, as well as distractions from outside, can interfere with short-term memory contents, and of course these effects rapidly increase when the individual has an awareness of waiting. This awareness comes as soon as several seconds—two seconds still seem to be a good number here.

That is why the tasks which humans can and will perform with machine communications will seriously change their character if response delays are greater than two seconds, with some possible extension of another second or so. Thus, a system with response delays of a standard ten seconds will not permit the kind of thinking continuity essential to sustained problem solving, and especially where the kind of problem or stage of its solution contains a high degree of ambiguity. Such a system will have uses, but they will be different from those of a two-second system.

Psychological step-down discontinuities with increasing response delays

The point here is that response delays are not merely matters of "convenience" to the user, unless the word "convenience" is made to mean more than it usually does. There is not a straight-line decrease in efficiency as the response delay increases; rather, sudden drops in mental efficiency occur when delays exceed a given point. These sudden drops at given delay points can be thought of as psychological step-down discontinuities. Thus, a ten-second response system (aside from operating inefficiencies) may be no better for the human—in some tasks at least—than a one-minute response or a five-minute response. If the human diverts his attention from the thought matrix (e. g., waiting to be filled or completed by

system response to some other train of thought), the significance of response delay changes dramatically.

The statement that "In the past it took two days to get an answer to a question that now is given in fifteen minutes" means, perhaps, an increase in *operating* efficiency for the system, but does not in itself materially change the cognitive (psychological) behavior of the person getting the information.

Psychological closure comes in different degrees. In the telephone example, I get a partial closure when I find the name in a telephone book, another when I complete dialing the number, and another when I am talking to the right person. Talking to the person I have in mind completes the closure of the series of transactions that led to hearing his voice and name. Just as there is a hierarchy of closures in a given task or goal-directed behavior sequence, so there are probably varying amounts of acceptable delays. The greater the closure, the longer the acceptable delay in preparing for and receiving the next response following that closure.

A general rule for guidance would be: For good communication with humans, response delays of more than two seconds should follow only a condition of task closure as perceived by the human, or as structured for the human.

Response time, or system response time, has not yet been defined in this report, so that the two-second rule applies to "meaningful replies" to human requests or commands, and these are defined, along with others in the pages to follow. In addition to definitions and examples of inquiry and response modes, estimates are made of acceptable response times.

Some qualifications about the analysis

The analysis of qualitative behavior and conversion of the analysis into quantitative limits is prone to misinterpretation. This is especially true when the subject is human behavior. Therefore, the following provisos are made explicit.

1. **The classes of response categories are not exhaustive.**

   The seventeen types of response category and response time cited in the next section of this report are certainly not exhaustive of all the possibilities. Without too much strain, however, they seem to cover a large proportion of interactive behavior between humans and information-processing systems.

2. **A response signal can communicate several messages at the same time.**

   A signal can communicate several messages to the user concurrently. Thus, if the system replies to a user query or command with the statement that is the equivalent of, "I've started doing your work," the user knows (a) his request has been listened to, (b) his request has been accepted, (c) that an interpretation of his request has been made, and (d) that the system is now busy trying to provide him with an answer.

   In the next section, the elements listed above are differentiated into four different kinds of response, but in system operation they may (or may not) be combined into a single communication. If so, the response time that should be met is that demanded by the component in the group which demands the fastest response time.

3. **The language in the text does not indicate the form of inquiry or response.**

   In most cases, a topic will be introduced by a title such as "Response to 'Here I am, what work should I do next?'" This expression is intended to simplify communication to the reader of the report. It does not imply that these words would be entered as such into the system. In many cases, the expression of the inquiry, or of the system's response, may be implicit in some other behavior. Thus, lifting the telephone receiver and putting it to my ear has the implicit question, "Are you listening to me and can you give me service?" The dial tone says that it can.

   The reader, therefore, is urged to look at the context under a topic title for proper orientation.

4. **Tasks can be done in other than the conversational mode.**

   Whereas in traditional batch activities by computer or by humans, responses to queries may have taken days, a response time of

two seconds may be stipulated in the following pages. Therefore, the critic will ask: "Isn't a response time of 30 seconds or even an hour, better than 24 hours? If so, why isn't it good enough?" The answer must be, "Yes, 30 seconds is better than 24 hours for some purposes, but it is not good enough to maintain the continuity of human thought processes." Where discontinuities in human thought processes are irrelevant or unimportant, both to effective problem solving and to effective use of the professional's time, then the conversational mode is beside the point. But it will be easily demonstrated that many inquiries will not be made, and many potentially promising alternatives will not be examined by the human if he does not have conversational speeds—as defined in this report—available to him. But tasks will still be completed as indeed they have been in the past, without conversational interaction, and at least *some* of them will be completed more poorly by any criterion. This assertion is certainly a testable hypothesis.

5. Permissible ranges of variation are not cited.

Any specification intended for implementation should include not only a nominal value but acceptable tolerance values within which the nominal value may randomly fluctuate. These tolerance limits are not generally specified for most of the response time values cited in the following pages. In principle, the range of acceptable variation of delay in a given category of response time is that range within which the human user cannot detect differences under actual conditions of use. By "use" is meant the context of the human performing a task in which the delay of the response element occcurs.

Some laboratory data* of indirect reference are available for making preliminary estimates of response time tolerances. Subjects judged intervals between clicks as "same" or "shorter" or "longer" than a comparison or reference interval between clicks. They gave their full attention to making these judgments. When the duration of the interval was between 2.0 to 4.0 seconds, the subjects made 75% correct judgment of "same" or "different" at the limits of an interval between minus 8% of the stimulus and plus 8% of the stimulus. For example, 75% of the time an interval of 1.84 seconds was judged shorter than 2.0 seconds, and an interval of 2.16 was judged longer than 2.0 seconds. This is about the same as giving a "tolerance" range of 16% of the value of the stimulus. This value of 16% applies in the range of 2.0 to 4.0 seconds.

The most accurate judgments of time (under these experimental conditions) were between 0.6 and 0.8 seconds where the tolerance range is somewhat less than 10% of the value of the stimulus duration (e. g., 10% of 0.6 seconds delay between clicks). With intervals longer than 4.0 seconds such as 6.0 to 30 seconds, the equivalent tolerance ranges were shown to be 20 to 30%. Substantially the same relationships held where the interval was started and stopped with a pulse of light.

The foregoing results were based on carefully controlled stimuli and full attention to the interval by the subjects. Where the stimulus changes from one display to another, and where there is subjective variability introduced by the human operator making a control response that initiates a machine delay, it is likely that response time variations may exceed these tolerances substantially. By exactly how much would require empirical data from subjects in simulated task environments.

Of indirect significance to this report are the findings by a number of investigators (cited by Stevens) that the time interval that bounds what is subjectively felt as the "psychological present" is between 2.3 to 3.5 seconds, although under some special conditions the boundary may extend to 12 seconds. This interval contains "the physical time over which stimuli may be spread and yet all perceived as present . . . the maximal physical time over which may extend a temporal stimulus pattern . . . which is perceived as a whole."

*See in S. S. Stevens, *Handbook of Experimental Psychology*, Chap. 32, "Time Perception" by H. Woodrow.

Basis of response-time estimates

The estimates of delay times offered in the following pages are the best calculated guesses by the author, a behavioral scientist, who has specialized in task behavior, including thinking and problem solving. These estimates are based on rationales, some of which are cited above and others in context. They should, indeed, be verified by extended systems studies—not in artificial laboratories using abstract tasks—but in carefully designed, real life task environments and problems. The human subjects in these studies must have had many dozens of hours practice in acquiring relevant task skills (and not merely in manipulating the controls at the console) in order for the findingst o be useful. Novices have their short-term memory registers heavily filled with what they are trying to learn; therefore, they are not guides as to what the problem-solving user (or other user) will be able to do and want to do when he is highly skilled. Traditional research practices in psychological laboratories would delay answers on these questions for years, however, and perhaps provide them after a new generation of large data-base systems are already on the market.

Nevertheless, the reader should accept the parameters cited as indicative rather than conclusive. It is relatively easy to arrange demonstrations for the skeptic about short response times that will impress him with how long four seconds can seem to be. The demonstration requires merely that he become absorbed, motivated, and emotionally aroused by the demonstration task.

Definitions of response time

Response to human inquiry, with few exceptions, serves as feedback to a continuity of thought. Human behavior occurs at a variety of information handling levels. Different kinds of response and response delay will be appropriate at different behavior levels. The definitions that follow depend in part for their time estimates on psychological rationales given in Part I of this report.

Topic 1. Response to control activation

This is the indication of action given, ordinarily, by the movement of a key, switch or other control member that signals it has been physically activated. The click of the typewriter key, or the change in control force after moving a switch past a detent position are examples. They indicate responsiveness of the terminal as an object. This response should be immediate and perceived as a part of the mechanical action induced by the operator. *Time delay:* No more than 0.1 second. See also Topic No. 13, "Graphic Response from Light Pen."

A second form of feedback to the user at a keyboard is evidence of the key's being struck. In a typewriter, this is given by the printed character on the paper. This appears practically simultaneously (to the user) to striking or activating the key. Even if printed feedback of text being entered by the user goes through the computer before it is printed on the platen or CRT, the delay between depressing the key and the visual feedback should be no more than 0.1 to 0.2 seconds.

(Note that this delay in feedback may be far too slow for skilled keyboard users. These people are able to attend to the display, not the keyboard, while activating keys, and they will be aware of an out-of-synchronization relationship between eye and hand. Some adaptation can be made—the mechanical pipe organ had delays estimated at between 0.1 and 0.2 seconds. Part of the organist's skill was learning to adapt to this delay. Recognize, however, that the sense of hearing is more time-dependent than the sense of vision.)

If the light pen is used to select characters for a message, confirmation by brightening the selected character should be identifiable by the user within 0.2 second.

Topic 2. Response to "System, are you listening?"

The hum of the dial tone is the response the telephone gives to this implicit query. No dial tone means: "There's no point in trying to do anything further on this channel now."

Time delay: Up to three seconds. The time for onset of this response may be variable, but at some cost in user confidence. Confidence will, of course, be highest if the response signal begins within a second after activating the ON switch.

Comment: These statements apply only to the condition in which the user is becoming "initialized" in a session with the console. If he is actively engaged in a working conversation with the console, he must get immediate (as perceived by him) attention for making an input to the system, such as pressing a control key or other form of

entry. Having to wait four seconds, or even half a second for any reason when he wishes to enter information is violently disrupting to thinking.

This question has two levels. On the first level, the user wants to know if the system is available to work for him. After favorable acknowledgment, the user—depending on his task—will specify the programs and data he requires for his "private working area" at this session.

Topic 3 Response to "System, can you do work for me?"

A. Because in many cases a "yes" or "no" to the question of system availability may depend on the kind of work to be done, the user must key in a request for a given service. As the user does this, he is becoming psychologically locked into a conversation, and his capacity for annoyance with the quality of service is increased.

Time delay: For a routine request (as defined by the user performing a task) the acknowledgment should be within two seconds. A routine request is likely to be a demand for an information image in the store. For an impromptu, complex request, the delay may extend to five seconds.

B. The loading of the programs and data called for by the user should be within 15 seconds, although delays of up to one minute should be tolerable. The user will spend his time during this delay in arranging whatever notes he has, and in organizing his thoughts preparatory to work.

C. Response to the user requesting "Set up my job from where I left off yesterday" should be within 15 seconds for most favorable acceptance, up to one minute for acceptance.

Topic 4. Response to "System, do you understand me?"

This implicit query may precede Topic 3, or be concurrent with it. Assume the user has entered a 7-digit telephone number as a single, meaningful operation. If he has made an error that the system can detect, he should be allowed to complete his segment of thought before he is interrupted or told he is locked out. After two *seconds* and before four *seconds* following completion of keying in his "thought," he should be informed of his error and either "told" to try again, or told of the error he made.

Comment: It is rude (i.e., disturbing) to be interrupted in mid-thought. The annoyance of the interruption makes it more difficult to get back to the train of thought. The two-second pause enables the user to get his sense of completion following which an error indication is more acceptable.

Topic 5. Response to Identification

Assume a badge-reader type of terminal. The user is on his way to his work station or is at his work station.

He inserts the card, badge, or other identifying medium. Ideally, he should have two kinds of feedback.

1. *Feedback to correctly positioned card.* This should be in the order of direct mechanical response, such as activating a detent or producing a click or snap, with a delay of less than 0.4 to 0.5 second. If failure to position the card properly occurs rarely, this form of feedback is unnecessary.
2. *Feedback saying the equivalent of "OK, I've read you."* This response time should be within two seconds, and be a fixed length of time. In general, people on their way to an activity experience mild annoyance at having their progress interrupted in order to be identified as an employee. The annoyance may be mitigated by making the interruption brief, simple, and standardized so that it can be accomplished practically by a series of reflex actions. That is why the confirmation of the identification should be made to the user in a standard length of response time. When a user clocks out, he is apt to be even more impatient with impediments. Then, a two-second delay will seem four times as long as a one-second delay.

   Another factor in identification speed is the bottleneck likely to exist at entrances to work locations where many employees arrive at about the same time. Small lines of employees were informally observed as they punched in at time clocks. Cycle time per employee—when he had his time card in his hand—was about three seconds at the clock. The clock itself had a response time of about one second after the time card was seated. Cutting this response time to 0.5 second would reduce the cycle time per employee by 16%, assuming other factors remained constant. But if the response time was four seconds, and it were to be cut to one second (and the other factors remained constant), people would pass through the line twice as fast with a one-second delay as with a four-second delay imposed by the action of the mechanism.

Comment: The delays proposed in this section are intended to apply only to that kind of identification implied by the statement, "Here I am and ready to go to work." Where the user sits at an inquiry terminal and says, symbolically, "This is who I am and I want to use your facility," a longer delay in acknowledgment is likely to be acceptable—say, up to five or seven seconds. (Note that this estimate is consistent with that of "Response to, 'System, can you do work for me?' " when the user is initiating an impromptu, complex request. See Topic 3.)

Topic 6. Response to "Here I am, what work should I do next?"

This inquiry is that of a production worker in a factory who has completed an assignment, acknowledged its completion, and requests from the terminal his next assignment. It is likely that this will be displayed to him in the form of a printed slip or card prepared at and by the terminal. Acceptable delays could range from 10 to 15 seconds.

This condition does not apply to the user in conversation with a terminal, such as in computer-assisted instruction. If the student has completed a segment of study and wishes to continue into another topic, the delay should be less than five seconds.

Topic 7. Response to simple inquiry of listed information

This form of inquiry presumes that the query addresses an existing record, or record-string, which can be directly retrieved and displayed. Example: Part #123456: give physical description. Or, Richard R. Roe: give man number. Or, Standard circuit #12345: give description.

If a terminal is frequently used by an employee for this kind of inquiry (say, more than once an hour), the response should be within two seconds. The employee is likely to have in mind some specific issue which the display response may resolve. It is also likely that the employee may have to scan several responses to his queries before hitting on the frame that fits his intent.

Topic 8. Response to simple inquiry of status

An example would be: "Current order status of inventory Part Number 123456." This is a simple inquiry because it asks for one category of information about an unambiguously identified object. The system may have to do some searching and processing from several storage locations to assemble the response. Where the user recognizes this requirement, the two-second delay limit may be relaxed to seven to ten seconds.

The user will be holding an idea in mind while waiting for the response, but it will be a single idea rather than a complex one. For example, "Can I or can't I take an order for 2000 items of this part number?"

Topic 9. Response to complex inquiry in tabular form

A complex inquiry is one which requires collecting and displaying data on the basis of logical relationships among categories. It assumes an "image" of the displayed response does not preexist in the system. An example: "How many orders for Product X, placed since January 1, 1967, have been cancelled to date?" Assume that master records are filed by customer name to which details of the order are added as attributes. These attributes include "date that order was placed," and "status" of which "cancelled" is a subcategory. The system must search these records (perhaps via indexes) and pull out the relevant items. (This is a simple example of complex inquiry.)

The user will certainly have a continuity of ideas in mind when he makes complex inquiries. This particular inquiry should get a complete response within four seconds.

Assume, however, the user had asked the same question for Product X, Y and Z. It would now be acceptable to display the answer about Product X within four seconds, about Product Y within four seconds after that, and about Product Z within the next four seconds.

The principle here is that it takes time for the user to assimilate the elements in a complex pattern. In many situations, four seconds per item would be longer than necessary, and two-second delays would in all cases be preferable.

If the display is graphic rather than tabular in format, additional considerations will apply. (See Topics 13 through 16 on graphics.)

Topic 10. Response to request for next page

Assume a graphic or high-speed printer output at the display. The user has completed reading or skimming a section of text which overruns into another "frame." The user activated the "Next Page" control.

Here, time delay should be no more than one second until (at least) the first several lines of text on the new page appears. You can test the annoyance of longer delays by becoming engrossed in some text and, when you are about to turn the page, be restrained from doing so to a slow count of four by an associate.

Delays of longer than one second will seem intrusive on the continuity of thought.

There is another page-turning condition. This is when the user is searching for some item of content which may lie on any of several pages or frames. A half second is a relatively long time, subjectively, for getting a page turned while searching for items of information.

A problem may be created when the user wants to scan through category indexes and therefore would like to flip pages quickly, unless the index already exists as an "image." In some cases, however, the index may have to be custom-built on the basis of the user's specific request. Where this occurs, the user must be informed that he can expect two-second delays when requesting the next frame of index terms.

If delays in advancing from a previous frame to a next frame in a viewing series are more than two seconds, it is increasingly unlikely that the user will use this medium for scanning and searching. It seems possible that adequate design of the application, however, can minimize the need for impromptu organizations of new indexes on immediate demand.

Skipping a number of pages or frames should be manageable with the help of a displayed index on one segment of the screen. The user should be able to skip ten pages all at once, as rapidly as the next page would appear.

Topic 11. Response to "Now run my problem."

Assume that an engineer or scientist has written a short program to solve a specific equation. He has written the program at the terminal. He presses the GO button.

(a) How long he will wait with patience will be partly a function of how long he took to write the program and enter the data.

(b) His patience will also depend on the number of additional data runs or changes he expects to make before selecting a particular set of parameters.

(c) His patience will also depend on how anxious he is to get back to other work for which the calculated result is a step towards solution.

If the result is returned to him within 15 seconds, he may remain at the terminal "in the problem-solving frame of mind." If the delays are longer, he will, to a corresponding degree, tend not to think of the terminal and system as in-line with his thinking, and attempt to fill in the wait times with secondary activities—probably an unsatisfactory arrangement to him, but less so than staring at a blank screen, or waiting hours for a response from the Computation Center. These interruptions may also *tend* to make him satisfied with a result after less experimentation than if he could continue uninterruptedly. (We assume he wants to see an "answer" before he tries another hypothesis.) This is a net loss to both system utilization and a user's problem-solving potential.

Topic 12. Response to delay following keyboard entry vs. light-pen entry of category for inquiry

Let us distinguish between light-pen entry of a category of information (such as a request for a given image or format by touching the light pen to a code name), and using the light pen as a stylus or drawing instrument. In this topic only the use of the light pen as category or function-selector is relevant.

Because it is easier for a nontypist to select instructions by light pen than by keyboard, he will expect a faster response to light pen. The difference may be that between the two-second response time to the light pen, and three-second response time to the keyboard. We can also expect a one to one-and-a-half second adaptation time required by the user for shifting his attention from the keyboard to the display.

This distinction disappears, however, when the user is activating a "page-turning" function on the display he is viewing. If he is continuing the reading of text (graphic or perhaps even tabular material) from one displayed frame to another, one-second delay after activating the control (light pen or function key) is a maximum. This is too long if he is scanning pages while searching for some specific content. (See Topic 10 which calls for less than one-second response time.) The user who is scanning a series of frames will keep his finger (or the stylus) poised over the "Advance to Next Frame" control, and activate it without shifting his attention from the screen.

Topic 13. Graphic response from light pen

There are two major ways in which the light pen is used as a stylus (as contrasted with its use as a control selector or alphameric message composer). One is that of drawing lines on the scope face where the direction and shape of the line have significance. That is, the actual path travelled by the light pen is the input to the system.

Where the lines are drawn with deliberation by the user—relatively slowly as compared with slashing sketch strokes—a delay of up to 0.1 second seems to be acceptable. There must not be variability perceived by the user in this delay.

Another way of using the light pen for graphics is to compose an image from a "menu" of image parts. For example, a glossary of references at the side of the image frame may be symbols of resistor, diode, transistor, and so forth. The user places his light pen over one of these symbols and moves the light pen to the position on the frame that he wants the symbol to be. A copy of the symbol follows the light pen. The response delay in the image following the light pen may be as much as one second because the user is not tracing a line but positioning an image that, for him, is completed when his stylus touches the destination for the image.

Similar delays of up to one second would be acceptable when the user is constructing the format for a graphic display of, say, a bar chart or line graph from a menu of symbols.

Topic 14 Response to complex inquiry in graphic form

Assume the same kind of inquiry as described in Topic 9 "Response to Complex Inquiry in Tabular Form" except that the response will be a display of bar chart, schematic, or graph.

The graphical response should begin within two seconds and certainly be completed within ten seconds if the user is to maintain thought continuity in an ongoing task—example, localizing the cause of an exception by means of category search. Other examples of such continuity in thinking would be the use of historical files during problem-solving sessions where the outcomes of these sessions would result in plans and hypotheses for organizational changes (operations research) or for growth (systems analysis).

Note: Many variables cited in previous topics also apply here.

Topic 15. Response to graphic manipulation of dynamic models

It is, of course, possible to animate a diagrammatic representation of a logical system (such as a computer), or a process system (such as a factory or inventory), or a topological system (such as transportation routings and flow). Pulses can simulate messages or transactions, and the thickness of a bar at the input to a symbolic work station may represent the size of a queue. Dynamic changes in the distributions of wait times at each of many stations can be shown on bar charts, whereas changes in the profiles of the bars show different patterns of queues or delays.

Experience with this kind of display is not sufficiently widespread to suggest the limits of analytical perception of human viewers of this kind of graphical simulation. We can expect that after many hundreds of hours of studious effort with this form of display, great improvements in perceptual sensitivity, retention, and interpretation will be achieved by at least some individuals with talent for it.

The problem-solving user will want at least three special properties in this kind of display. One is that of enlarging a segment of a display field. A second is that of selectively suppressing details in the representation of action or structure—similar in principle to going from lower levelt o higher level diagrams of a mechanism. A third will be an easy means of visually enhancing some given path or paths in a complex representation, while suppressing the remaining content into visual phantoms.

Response-time limits for these functions are not even readily conjectured. The serious problem solver will, of course, be prepared to spend many hours planning and executing the design, optimization, or simulated test of a complex system facility. Flexibility in his ability to get the display

to shift rapidly from one degree of time compression or expansion in simulated system behavior, or from one level of detail to another, will be important. This flexibility will determine how much and how well he can perceive, interpret, hypothesize, control, and modify. But putting minimum limits to the words "flexibility" and "shift rapidly" in the preceding sentences would be premature beyond the guess that whatever "scenario" of events the user must comprehend and work with should be compressible into 50-minute periods of time. Even this may be 10 times greater than the chunk of information that even a problem-solving specialist can hold in mind and work with as a designer or evaluator.

It is here that we need inventive, developmental studies somewhat similar to that conducted by the RAND Corporation in the early 1950's about how mucha SAGE operator could assimilate—and under what conditions.

Statements about response times for graphic simulation of dynamic models will, therefore, not include even guesses at this point in knowledge.

Topic 16. Response to graphic manipulation in structural design

Examples of structural modelling are a highway engineer's designing a bridge, or an engineering architect's designing a building.

When the designer adds an element to the design, one system requirement is that of applying sets of algorithmic rules to that design element. For example, "Only one physical body can occupy a given space at onet ime," or "Building codes require that. . . ." Another system requirement is remembering what the designer has already done. A third requirement is translating sketch responses into the equivalent of appearance renderings and engineering renderings.

The intensity of design conceptualization demands rapid response from the medium on which the designer is working. But the designer will have to accept some constraints (disciplines) in how he attacks and sequences or stages his design effort in order to obtain reasonable system response (i.e., two-second response time, to be informed that he just sketched in a dimension that violates a rule, or type of rule).

During creative effort, idle time beyond a couple of seconds by the designer, while he waits to see the consequence of a unitary action, will be inhibiting and intolerable. But, after the designer has completed working out an idea—a chunk made up of a number of individual actions—he will be inclined to wait a minute or two, while the system "catches up to him."

Comment: People engaged in creative activities recognize the relatively large amounts of work that can be executed during concentrated and continuous "mental heat" in a single session. This heat can cool off in interruptions lasting less than a minute. It is this heat of attention that the system should attempt to preserve.

Graphic motion that the designer perceives as relevant to the design task will help keep his attention and state of arousal, at least if it continues for no longer than ten seconds in consummating some design action. In other words, it is possible to present artifacts to the designer that will maintain his psychological "coupling" to the system. The concept precludes setting fixed response time limits to various response functions, except that their limits will be in seconds (usually) rather than in minutes.

Topic 17. Response to "Execute this command into the operational system."

An example of such a command is a manager's intervening in an automatic ordering process and designating an alternate vendor. Or, the manager may insert a command which, when effected, results in a change in scheduling of some manufacturing operation. Or, as a result of simulation and modelling of certain activities of the business, a revised operating budget is introduced and its implications for a number of affected departments are exploded and disseminated.

Although the user should be informed by the system within four seconds that it has understood and can interpret the command, its execution and final confirmation to the user that the command has been executed may have long and variable delays of minutes. The user has terminated one level of activity when he enters the command. It will be psychologically incomplete only to the degree that he expects a feedback telling him of interference with its execution. These delays, however, are partly dependent on operating activities outside the scope of the automatic system, such as a remote manager's being unable to accept a budget cut or change in schedule.

Postcript 1

Discontinuity of waiting time at 15 seconds

Assume an inquiry of any kind has been made. The user—and his attention—is captive to the terminal until he receives a response. If he is a busy man, captivity of more than 15 seconds, even for information essential to him, may be more than an annoyance and disruption. It can readily become a demoralizer—that is, a reducer of work pace and of motivation to work.

If, therefore, response delays of more than 15 seconds will occur, the system had better be designed to free the user from physical and mental captivity, so that he can turn to other activities and get his displayed answer when it is *convenient to him* to do so.

A possible, but doubtful, exception may arise when the user is in series with some process or continuity that demands (as soon as possible) the answer from him, which, in turn, depends on information he is trying to get at the terminal. In this case, the operating demands dictate acceptable time delays.

In any event, response delays of approximately 15 seconds, and certainly any delays longer than this, rule out *conversational* interaction between human and information systems.

Postcript 2

Time recovery from errors and failures

A dimension of response time is the question, "How quickly can I get going on my task again after something goes wrong?" What may have gone wrong could have been a machine failure, a failure in an operating program, an operator error, or an error by the user in mid-task.

The design of the system, including the application, should simplify both the effort and shorten the time required for recovery as perceived by the user.

If the user was in simple-inquiry mode, he will probably have a record of his last inquiry at the terminal, and can input the inquiry again.

If the user was in complex-inquiry mode, the last *index of categories* that he was using before the failure should have been retained and made available to him, so that he can pick up his inquiry from a position of good context.

If the user was in conversational problem-solving mode, there should have been retained a copy of all the parameters and starting structure of the model he constructed. Reconstructing this model would, from the user's standpoint, be the most arduous and unreliable of activities. (As an example of this almost universal dread of work getting lost, many writers and engineers save their yellow-sheet draft sketches in desk drawers until the job is entirely completed.) One can tolerate the loss of a machine run, which can be rerun later, but the loss of even an hour's creative work is obviously demoralizing. Rarely does one feel confidence that the reconstruction has all of the magic contained in the original.

When a system failure occurs, from whatever cause, the user is likely to feel an irrational sense of failure if his job has been lost. In some degree, it will be remembered as personal failure, and various psychological defenses will be inevitable. (One form of defense is to avoid the cause of the threat in the future.) It is therefore desirable, for motivational reasons as well as operating reasons, to attempt to restore the system as quickly as possible so that he can pick up and continue. "As quickly as possible" means "while he is still in dialogue (or work session) with the system"—and that means within 15 seconds, or failing that, within less than five minutes. The system should tell him how long he may have to be patient, and it should do so immediately after the failure, whatever it may be.

BIBLIOGRAPHY

W BLAKELY
The discrimination of short empty temporal interval
PhD disertation University of Illinois Library 1933

J R NEWMAN
Extension of human capability through information processing and display systems
System Development Corp Santa Monica December 1966

H SACKMAN
Experimental investigation of user performance in time-shared computing systems
System Development Corp Santa Monica May 1967

H SIMON R K MELLON
Reflections on time-sharing from a users point of view.
In Computer Science Research Review
Carnegie Institute of Technology 1966

S S STEVENS
Handbook of experimental psychology
Wiley N Y 1951