HOW PEOPLE INFLUENCE EXPERIMENTAL RESULTS [1]

June Roberts Cornog, [2] National Bureau of Standards

Information is stored so that people can retrieve it when there is need. If people cannot easily learn to manipulate the system by which information is stored when they want to know something, or if the system is undiscriminating enough so that it returns a good deal of unpertinent information when queried, or if the desired information isn't present in what is retrieved, then the whole storage process has been in vain.

If we extend the line of thinking begun in the above paragraph, certain other truisms must also be stated. Any system that people are to use can be validly tested: a) only by having people use it; b) by observing the users under some kind of controlled conditions so factual data about their performance can be gathered instead of just asking their opinions; and, c) by assigning them an experimental task characteristic of those for which the stored information would normally be required. These are obvious truths but their application to any testing situation is much less patent.

If we examine the problems connected with any experimental investigation that fulfills the requirements stated above, one major difficulty at once appears: people under observation do not behave as they normally would! The mere fact that they know they are being watched, or measured, changes their behavior. Considerable thought has been devoted over the years to how best to cope with the "human nature" of the people who must be used in experimental studies, but so far it is doubtful that any research administrator has ever been able to control his Subjects completely.

Changes in behavior due to observation may take many forms but in an experimental study such as that described below, five sources of behavioral variation are very likely to occur: (Refs. 2, 4, 5, 6, 7)

1. People try to help!

   They usually decide what they think the outcome of the study is likely to be and often they consciously or unconsciously weight their efforts in that direction!

2. They set up their own private operating rules.

   Every experimenter confidently lays down the rules of the game, with every expectation that they will be observed by all participants. But rules practically never cover every eventuality and strict inquiry soon reveals that participants are handling the unincluded, or are perhaps just interpreting the rules, in their own individual ways. Systematic error may creep in through loss of uniformity in operation!

3. The Subjects learn as the study progresses.

   People don't stay naive very long. If one of the factors in the experimental task is measurement of learning time, or if the performance of naive Subjects is to be compared with that of experienced people, such individuals can be used only once. The performance or working habits of experienced people often change as they continue to work at jobs they do all the time. If the objective is to measure the effectiveness of a particular method, any changes in work habits or any learning on the job will introduce error.

4. People are never able to be totally consistent in a judgment task.

   The point does not need to be expanded. Every reader will recognize the minor daily variations in his own decisions.

5. People under observation usually show signs of stress.

   They become more careful, give more attention to detail, or, in the case of a manual task such as typewriting, commit an unusual number of mechanical errors.

The U.S. Patent Office research and development program is concerned with the testing of mechanized information storage systems. The information stored is that identifying the inventions in patents already granted. Retrieval of pertinent information may occur when the already patented ideas are searched to determine the novelty of the idea contained in the new application. The data storage systems are usually unique to the "art" they are designed to accommodate so that storage of information about, say electronic transistors, must be quite differently handled than information about chemical insecticides.

After a storage system is designed and at least partly applied, its true applicability to the art it is supposed to store, as well as its

usefulness to the patent examiner, must be assessed. This can be done only by having the future users use it - and, of course they are people! Patent examiners are, in addition well above average in intelligence, education and analytical skills so that the kind of "cover story" usually used in experimental situations is useless. In the study described below it was necessary, therefore, to state the purpose straightforwardly and simply to appeal to the good sportsmanship of the participants, by asking them to behave as normally as possible.

The influence of people on the experimental results may be illustrated by a Patent Office research project. One of the chemical "arts" is known as the organometallic group. A project to store the information from these patents in a mechanical system was undertaken.

After the system of encoding and storage was designed and partially carried out, a study of the encoding process was devised in order:

1. To examine the effects of coding on ultimate retrieval of patents (would the encoding used produce all pertinent references and only the pertinent ones upon demand. Discussion of this objective is not included in this paper) and,

2. To derive a method for establishing a criterion for satisfactory coding before full-scale work in this area was begun.

All the available coding people were divided into two as nearly matched groups as possible. All persons were trained simultaneously and their understanding of the work tested in advance. All Subjects worked at their own desks under as normal conditions as possible, each participant was reassured that only mass data would be used - that no one's record would affect his job standing in any way. Only one part of the task was a little out of the ordinary - each coder was asked to keep a record of the time he spent on each case.

For a description of the way the study materials were handled and how the criterion for accuracy in the study was established, I quote the paper delivered by Mr. King a few minutes ago, Ref. (1), Sect. B.

> Briefly, 201 documents were drawn at random from the total file of 3,625 documents. Each of the documents had, during the construction of the file, been coded and reviewed by an experienced analyst, the reviewed encoding being included in the file. Twenty-four of the 201 documents were drawn at random for an intensive coding experiment. Each of the remaining 177 documents was encoded again by random assignment to one of four experienced analysts. Then the two most experienced analysts, with the original encoding, the reviewed encoding and the second encoding before them, selected what, in their combined opinion, was the "correct" encoding. Half of the 24 documents were assigned, at random, to three inexperienced analysts, and half to three experienced analysts. ("Experience" refers to length of time analysts had previously devoted to indexing organometallics documents. All were experienced chemical analysts.) Then the same two senior analysts who selected a correct encoding for the 177 documents made the same determination for these 24, having before them the original encoding, its review, and three more independent encodings. Thus, for 177 documents one can compare actual and experimental coding against a coding defined to be correct. For each of these, measures of consistency between two analysts (the original and the experimental analyst) are available. For each of the 24 documents one can compare the original coding and three experimental codings against a standard and can compute measures of consistency among three experimental codings and an original coding.

The difficulty of a patent was judged by two standards - The chemical compounds included and the sheer thickness of the document. All Subjects were asked not to confer with each other (and asked later to affirm that they had not) when they were told that members of each group of three persons would be duplicating each other's work. Measurement was in terms of:

1. Amount of time spent on each patent;

2. Number of terms encoded;

3. Conditional probabilities as a measure of accuracy;

4. Two indices of indexing consistency.

The work covered a period of four days. Other details of statistical methods may be found in Ref. (1).

Analysis of the data gathered showed some interesting contrasts which illustrate the principles stated earlier. The summary statements were based on both the factual data and on observations made by the administrator as the study progressed. Even though these sources of error were noted, they did not significantly affect the outcome of the study.

1. There was more difference in performance within one of the two experienced coder groups than between the experienced and inexperienced groups, as measured by total number of codes produced, accuracy of those codes and time required in encoding the document. Even though some of the differences among the experienced coders may have been due to variance in the difficulty of the patents worked on, learning among the inexperienced coders occurred so quickly that the pre-test training sessions evidently reduced experience as a test factor. The worst coder in terms of accuracy was one of the experienced people!

2. The time spent on single cases varied between 10 minutes and 3 1/2 hours! The difference in the thoroughness characteristic of the individual coders was undoubtedly strongly felt in these time measures, plus some influence from the variations in their educational backgrounds. The most obvious implication is the previously cited difference in the difficulty of the individual patents encoded, of course.

3. Some Subjects consistently encoded larger numbers of terms than others did but this measure was not related to amount of previous experience. Furthermore, when work performed under the close observation of experimental conditions was tested against the numbers and kinds of codes produced in the course of normal work, additional differences in performance were found although these were not statistically significant.

4. When some reviewers knew which coder's work they were reviewing, they were inclined to become more or less critical according to which coder was involved. Criticism was indicated by a larger number of codes added or deleted.

5. Adequacy of coding was necessarily a matter of judgment but the reviewer tended to go along with the work the coder had done, especially if he didn't know who the coder had been. From the standpoint of omitted terms, two independent analyses or encodings were shown to give better results than one analysis and one review.

Error attributable to the sources listed above was not observable in the data from the groups, but existed in the scores of individuals according to which coder was involved. It is probable that a good deal of the variance among individuals was due to personality variables since no correlation existed between scores and either education or amount of experience.

Consistency among judges was unusually good -- of all the terms encoded inconsistently, only 12% were considered to be ambiguous by the judges. In one other Patent Office information coding and retrieval study, however, there was as much disagreement among the judges as among the coders!

Some further explanation is in order here as to why ambiguous terms are unusually likely to occur in Patent Office work. Inventions always deal with new ideas, with advances in science or technology, with material which has not yet had time to become standardized. Inventors frequently must invent not only the gadget or process but the technical language to describe its function or purpose. If several applicants arrive at nearly the same idea at approximately the same time, they don't necessarily use the same terms to describe it. The patent examiner and coder must look to the concept involved and try to reach a standard language for themselves. Ambiguities are therefore likely to be more numerous than the "outsider" would expect.

The organometallics study yielded one important observation as well as some unusually high consistency measures:

1. It was observed that people who seemed to have greater patience for detail did the best job of encoding patents.

   They were more accurate -- they missed a smaller number of terms and carried through the analysis of chemical compounds more thoroughly. They apparently had a goodly share of the well-known compulsion to "get it right!" Neither the amount of experience nor the educational background of the coder correlated with the quantitative measures taken.

2. Participants showed high levels of accuracy when their work was compared with the criterion codes (3). (See Table I.)

TABLE I

ESTIMATES OF ACCURACY AND CONSISTENCY AND OF LINEAR CORRELATION BETWEEN ACCURACY AND CONSISTENCY BY TERMS FOR SAMPLES OF 24 AND 201 DOCUMENTS.**

	Linear correlation coefficients	
	24 documents	201 documents
a. A measure of accuracy vs. a measure of consistency*	.88	.86
b. A measure of accuracy vs. the Consistency Coefficient*	.90	.95
	Estimates of these characteristics	
c. A measure of accuracy*	.823	.816
d. First measure of consistency*	.862	.726
e. Accuracy squared	.698	.675
f. Consistency Coefficient*	.729	.665

The Standard Errors of these estimates do not exceed .02 in any instance.

*See Ref. (1), Sec. B, for explanation of measures used.

Inspection of data in Table 4, Ref. (1), Sec. B, shows that accuracy and consistency measures of the encoded terms were closely related. The same personality trait probably influenced both measures.

From the data gathered in the organometallics coding experiment, the investigators were able to predict the parameters of the model in the small study and to test the prediction model against the actual, observed retrievals. Apparently the sources of human error described here did not significantly affect the experimental factors.

**Data taken from Tables 3 and 4, and from explanation for Tables 3 and 4, Sec. B, Ref. (1).

Bibliography:

1. Bryant, E.C., King, D.W. and Terragno, P.J., "Designs of Experiments in Information Retrieval," Proceedings of the Social Statistics Section, American Statistical Association, 1963.

2. Edwards, Allen L., "Experiments: Their Planning and Execution," in Handbook of Social Psychology, Vol. 2, pp. 259ff, ed. by Gardner Lindzey, Addison-Wesley Publishing Co., Cambridge, Mass., 1954.

3. Jacoby, J. and Slamecka, V., "Indexer Consistency under Minimal Conditions," Documentation Inc., Bethesda, Md., November, 1962.

4. Jahoda, M., Deutsch, M. and Cook, S.W., Research Methods in Social Relations, Part One, pp. 92-127, Dryden Press, New York, 1951.

5. Jahoda, M., Deutsch, M. and Cook, S.W., Research Methods in Social Relations, Part Two, pp. 463-487, Dryden Press, New York, 1951.

6. Sidman, M., Tactics of Scientific Research, pp. 341-393, Basic Books, Inc., New York, 1960.

7. Underwood, R.J., Psychological Research, pp. 17-48, Appleton-Century-Crofts, New York, 1957.

Footnotes:

1/ Prepared in connection with a special project sponsored by the U.S. Patent Office, on procedures for orienting patent examiners toward non-manual searching methods.

2/ Research Psychologist.