

On Measurements in User Studies

Charles T. Meadow
Professor Emeritus
Faculty of Information Studies
University of Toronto

The formal measure of information most often found in the literature is that of Shannon [1] {Fig 1} which has to do with the probability of a symbol being selected or transmitted. The average information content of a *system*, not a single symbol or message, is the sum of all $p_i \log p_i$ for an entire set of possible symbols. OK, easy to compute. Now suppose we are trying to measure how much information a library user got from his or her last visit. We can measure the probability of each book being selected from the entire collection but that tells nothing about what that user gained from the books checked out. Ask the person and you are likely to hear something like, "Well, I got three good books." Ask for more detail and you might get, "One is really great, one doesn't look so valuable now that I've had a second look, and one is OK but not great." We could, in fact make a numeric scale out of *great*, *OK*, and *not so great*, sum the scale values for all books selected and use this as measure of information retrieved. But of course, this is more *relevance* than *information content* in the usual meaning of these words. But it serves as an example of what we face in trying to measure *users*, *user performance*, and *search outcome*.

I see this as the major problem in this discipline of ours - user studies. We keep doing projects, taking measurements, but not using anything like standard definitions of our dimensions or metrics, leading to . . . what? What follows is an unabashedly personal view of the field of user studies and what I think needs to be done to make us a respectable science. Many of my citations are quite old. I do this to stress for how long we have talked about some of these problems without doing much about them.

The announcement for this program mentioned a 1997 paper of which I was co-author. [2] My colleague, Weijing Yuan, and I had been working on a project whose overall goal was to develop ways to measure the "impact of information on development." The context was development in the sense of developing countries. The Canadian government spends a fair amount of money supporting this kind of development and also on research into the development process. The question arose of how to tell what benefit derives from efforts to provide information or information systems to developing countries. We did a small project along that line, one of several, and then, with a little bit of left over money, tried to tackle the question of what variables matter, what variables other researchers used in their studies of the use of information systems, and what the variables meant, what phenomena they measured.

Why this second project? At one meeting of six or eight of the project heads, reporting on our respective progress, it became apparent to some of us that the meanings of the expressions *information*, *impact*, and *development* varied enough that we were convinced that we were not all working on the same issue. So, we wondered, what do these words mean and how do you measure them? Dr. Yuan and I first did a study of how *information* and related words were used. [3] We found that this question has arisen in many (most?) disciplines and that when careful definitions were made, while rarely identical, they were usually compatible. We were hardly the first to have undertaken to look into this question and a list of information science people who have had a crack at it is in our paper.

Unfortunately, most authors of otherwise careful user studies research do not trouble to define the terms or refer to standard definitions of terms they are using. Also, we almost all use the word *information* to have a variety of meanings in casual conversation or writing, rarely distinguishing among knowledge, information, and data. But when we write carefully, we tend toward the concept of information as a message containing news (something we did not know before), which is of significance, or which changes the system receiving it. One form of a receiving system is a human's knowledge base. Dr. Yuan and I became convinced that, although we could never, in our life times, get the world to adapt a single

meaning for each of *information, data, intelligence, and perhaps relevance*, it should be possible to get journal editors, referees, and perhaps research granting agencies to insist that each author make clear which meaning of these terms is being used.

Next, we turned our attention to variables used. This was a primitive form of meta-analysis, but instead of comparing the *values* of variables appearing in a set of papers, we had to content ourselves with looking just at the *names* of the variables used in information retrieval user studies. This was necessitated by the huge variation we found in meanings, means of measuring, or scales used in measuring variables. In effect, our work was more a form of bibliometrics, using referral to variables rather than to other documents as the basic unit of analysis. It was not truly meta-analysis which generally compares results of similar phenomena found in different studies. I have to stress that ours was a small task with end-of-project money and there were only two of us. What we needed ideally was a team of at least a dozen graduate students reading several hundred papers, noting all the variables used, in what combinations, and the popular ones not used, and then communicating with the authors about what definitions they used for these variables. Typically, the definitions were not present in the paper. We did invite comments by some authors about why they used certain variables and what they meant, but got little information from them. Perhaps it was not the most tactful thing for us to have done.

Some Background

We were hardly the first to have recognized this problem. Samuel Hayakawa wrote, as early as 1939, that "no word ever has exactly the same meaning twice." [4] This seems unhappily to apply to the selection of names of variables as well as in general communication. Leon Brillouin wrote in 1956 that to formulate a scientific theory of information, "the first requirement is to start from a precise definition. *Science begins when the meaning of the words is strictly delimited.*" [5] (Emphasis added.) William Goffman wrote in 1970 that the word *information* is used in so many contexts that no single definition is possible. [6]

Earlier, Gary Marchionini, Joan Cherry and I had done a paper [7] expressing some of our frustration with variables used or mentioned in user studies, but rarely well defined. In mathematics we even have to define what is meant by a well formed formula in order to accept a formula or statement. In our work in information science we, as a whole, have been remarkably unwilling to do anything like this and I believe it hurts our research and also our new product developments. Further, at least partially because none of us ever seems to use the same definitions or the same measurement techniques, we rarely can build upon data collected by others or proofs others have established. That last point is critical. Imagine publishing research, say in biology, without stating the number of subjects tested. Granted, that omission is pretty rare, but I have seen it, and in a respectable journal in our field.

Some Pragmatic Consequences

One consequence of incomplete or ineffective measurement is that system designers do not have reliable data upon which to base design parameters. For example, I believe that the query languages or whatever we may wish to call them, of modern Web-based search engines are inferior to those we had ten or twenty years ago with Dialog, BRS, or MEDLINE. To a large extent, because most Web documents are not well indexed, there is not a great deal of benefit to a language capable of distinguishing the role of a term, rather than its mere occurrence. An exception to language deficiencies of today, to me, is in ranking but there are flaws with it, too. Mainly, this is due to lack of information about how it works and lack of user ability to change it. Yet, of course, the databases are hugely different in size and getting 100,000 hits is no rarity today.

Why are modern systems inferior? I can answer only with an anecdote. A year or two ago I visited a software developer who had produced an intermediary system that would analyze a question, try to improve it and then submit it in turn to a selection of search engines. It would combine the results from these engines, then use its own ranking technique to produce the list of ranked items displayed to the

user. The query improvement was done by finding the frequency of occurrence of query terms in a reference collection in which the reference works were presumably related to the field of the searcher, who had a hand in selecting which were to be used. And the referral to search engines was also at the user's choice.

I was quite impressed. However, the system lacked one feature the same company had put into a more specialized version that I had used earlier, a feature I liked. It allowed users to change the weight assigned to individual words for ranking purposes. When I asked why it wasn't in the newer, expanded system, I was told too many users had complained that it was too complicated, hence distracting. But, it would have been easy to make the feature optional. The company rejected it, based on anecdotal evidence, not hard facts. Are we using systems designed to be attractive to novices and lazy users, rather than those that can produce the best results?

Another aspect of this over-simplification of systems was brought to me by the late Jeffrey Katzer. He told me of his concern from seeing business web searchers (not students) seemingly accepting the first three items they see that appear to be on the desired subject. Quality, reliability, even truthfulness were not at issue. Further, everything they might want, in the way of information, is assumed to be on the Web and available free. Maybe your securities analyst was doing this kind of searching when the hi-tech stocks began to melt down. Jeff was convinced that few of these people were skilled searchers, but they would not acknowledge it, and as a result would take no steps to become better. Our co-chair today once told me of having been invited to give a few lectures on searching to a business school class. Afterwards, she received a thank-you note from the inviting instructor who pointed out that the students seemed not to have understood what she meant by source evaluation. How can people search for information but have no idea what it means to evaluate what they find?

How do these stories relate to measurement? We researchers in the field cannot provide system designers with meaningful data about users, what they know, how they work, what will work best for them, what help they may need, and what constitutes a successful outcome for them. And we need to be able to recognize differences not just among individuals but among cultures and environments. Users who do not know much about what they are doing are likely to want the simplest-appearing systems, not the systems likely to produce the best results. If a document has the right key words in it, it must be useful and true. We have ranking systems used by many search engines, but how often are users informed of what the algorithms are, let alone asked what algorithm they want? When we get 50,000 hits, it is important to know why we are seeing the first ones first. What is the best ranking system for a system designer to use? Would any engine designer dare ask the user what ranking procedure should be used?

Might we be seeing a form of Gresham's Law ("Bad money drives out good") applied to information? Daniel Boorstin first noticed this [8], in these terms, but he was referring to content - the easily obtained information of short-bite television news, for example, rather than a book or magazine article. The law may also apply to systems: pick the system that seems to make life easier. If you're buying books, buy the ones whose titles end with "for Dummies." Avoid investments of time and money in learning to use more complex systems that may offer far higher long-term pay-off in terms of results. Best to deny that anything better is possible.

Our Problems

Let me try to summarize where I think we stand:

1. We do not have established definitions for the appropriate descriptors or dimensions for

Characterizing users, in terms of their knowledge, skill, intent, and limitations imposed on their searching. It is necessary to recognize differences among groups as well as individuals and the

effect of local environments, such as search time restrictions.

Characterizing the tasks users are performing or goals they are aiming for, although there has been some work along this line by Bourne, Belkin and Saracevic.

Characterizing the actions taken toward achieving their goals, and again there has been work done but no generally accepted conclusions.

Evaluating how the task was performed and how well the goals were achieved. We must be able to distinguish among the performance of the user, the database and the search system.

In evaluation, we have to consider who does the judging about user qualifications, tasks assigned or undertaken, actions taken, and outcome. Also, what are appropriate qualifications for doing judging. We know that evaluation studies were done as far back as the 1950s, notably the Cranfield studies in England [9, 10] and these of Kent et al [11] in the U.S., but for years we all used precision and recall, based on the will-o'-the-wispy measure, *relevance*. Saracevic, back in 1975 [12], showed the ambiguity in this measure - so many different meanings. Yet, we still tend to use it. Rarely have I seen a research paper in which it is made clear what meaning of relevance judges (whether the users, themselves, or "experts") are expected to use. And if no one tells them, we must accept that they make a random decision, some choosing one definition, some another, probably without recognizing it.

2. Lacking these clear definitions and the requirements that investigators be clear, we have no compendia of data or meta-analyses upon which new investigators can base their work. Hence, we generally go back to scratch on every experiment. Here is an example of the result of some meta-analysis done long before this term was coined. {Fig 2} You can see a beautifully smooth curve fitted to the data, but also a great deal of variation in individual measurements. But, it can only be done if everyone agrees on what, in this case, *thermal conductivity*, means.
3. One reason we have no such compendia or meta analyses is that research funding agencies have been loath to invest in such unglamorous undertakings. Why is this? Now, I'll admit to be guessing. One reason is that it's not glamorous for the funders. Another is that it's not glamorous for tenure committees. Another could be that no one is asking. I made a couple of overtures some years ago and admit to being daunted by the complete lack of funder interest. On the other hand, the U.S. Government used to fund entities called *information analysis centers* back in the 1960s, which did some of this kind of work but not, to my knowledge, in information science. They funded the Office of Standard Reference Data in the former National Bureau of Standards, which did meta-analysis, but again not in our field.

What We Need to Do

We have to go back to fundamentals and work toward standards. This does not mean everybody must do same thing or do it in the same way, but that everybody must use definitions that are generally understood and, if non-standard, can be related to the standards. If Americans insist on measuring distance in miles at least they are precisely convertible to what the rest of the world uses. Editors, referees, research granting agencies must insist on adherence to standards in reporting work.

We need a way to share numeric research findings. That would require some standardization in the way in which data are recorded.

We have to learn to build on each other's work as is conventionally done in science, but not in ours. Yes, we have citations, but rarely do you see a research paper in which it states that the work of A has established the following values for x , y , and z and we proceed from there.

We do suffer from working with variables that are difficult to define and measure. Distance is so much easier to deal with than relevance. That's an excuse for not having got as far as the physicists have but not an excuse for not having started.

References

1. Shannon, Claude E.; Weaver, Warren. *The Mathematical Theory of Communication*. Urbana, Illinois The University of Illinois Press, 1959.
2. Meadow, Charles T.; Yuan, Weijing. A study of the use of variables in information retrieval user studies. *Journal of the American Society for Information Science*, 50(2) 1999, 140-150.
3. Meadow, Charles T.; Yuan, Weijing. Measuring the impact of information: defining the concepts. *Information Processing and Management*, 33(6) 1997, 697-714.
4. Hayakawa, S. I. *Language in Thought and Action*. New York: Harcourt, Brace & World, 1939, 60.
5. Brillouin, Leon, *Science and Information Theory*. New York: Academic Press, 1956,
6. Goffman, W. Information science: discipline or disappearance? *ASLIB Proceedings*, 222, 589-595.
7. Meadow, Charles T.; Marchionini, Gary; Cherry, Joan. Speculations on the measurement and use of user characteristics in information retrieval experimentation. *Canadian Journal of Information and Library Science*, 19(4) 1994, 1-22.
8. Boorstin, Daniel. *Gresham's Law: Knowledge or Information?* Remarks at the White House Conference on Library and Information Services. Washington: Library of Congress, 1979.
9. Cleverdon, Cyril; Keen, Michael. *Factors Affecting the Performance of Indexing Systems*, Vol 2. ASLIB, Cranfield Research Project. Bedford, UK: C. Cleverdon, 1966, 37-59.
10. Cleverdon, C.W.; Thorne, R.G. *A brief experiment with the uniterm system of coordinate indexing for the cataloging of structural data*. RAE Library Memorandum #7, AD 35004. Farnborough UK: Royal Aircraft Establishment, 1954.
11. Kent, A., M.M. Berry, F.U. Leuhrs, and J.W. Perry. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation*, 6(2), 1955, 93-101.
12. Saracevic, Tefko. Relevance: a review of and a framework for the thinking on the notion in information science. *JASIS*, 26(4) 1975, 321- 343.