

Whatever Happened to Information Theory in Psychology?

R. Duncan Luce
University of California, Irvine

Although Shannon's information theory is alive and well in a number of fields, after an initial fad in psychology during the 1950s and 1960s it no longer is much of a factor, beyond the word *bit*, in psychological theory. The author discusses what seems to him (and others) to be the root causes of an actual incompatibility between information theory and the psychological phenomena to which it has been applied.

Claude Shannon, the creator of information theory, or communication theory as he preferred to call it, died on February 24, 2001, at age 84. So, I would like to dedicate this brief piece to his memory and in particular to recall his seminal contribution "A Mathematical Theory of Communication," which was published in two parts in the *Bell System Technical Journal* in 1948 and rendered more accessible in the short monograph by Shannon and Weaver (1949).

Let me begin by saying that information theory is alive and well in biology, engineering, physics, and statistics, although my conclusion is that for quite good reasons it has had little long-range impact in psychology. One rarely sees Shannon's information theory in contemporary psychology articles except to the extent of the late John W. Tukey's term *bit*, which is now a permanent word of our vocabulary. If we look at the table of contents of J. Skilling's (1989) *Maximum Entropy and Bayesian Methods*, we find the pattern of chapters on applications shown in Table 1. You will note none are in psychology.

Because it is doubtful that many young psychologists are learning the subject, a few words are necessary to set the stage. As mathematical expositor extraordinaire Keith Devlin (2001, p. 21) stated: "Shannon's theory does not deal

with 'information' as that word is generally understood. Instead, it deals with data—the raw material out of which information is obtained." Now, that makes it sound akin to what we normally think to be the role of statistics, which is correct. It begins with an abstract, finite set of elements and a probability distribution over it. Let the elements of the set be identified with the first m integers and let $p(i)$, where $i = 1, \dots, m$, be the probabilities that are assumed to cover all possibilities, that is, $\sum_{i=1}^m p(i) = 1$. *Uncertainty* is a summary number that is a function of the probabilities; that is, $U(\{p(i)\}) = U(p(1), \dots, p(m))$. What function?

Shannon's Measure of Information

To get at that, Shannon imposed a number of plausible properties that he believed such a number should satisfy. Others subsequently proposed alternatives, among them Aczél and Daróczy (1975); Aczél, Forte, and Ng (1974); and Luce (1960). Probably the best alternative was that offered by Aczél et al. (1974), which, with less than total precision, was as follows:

1. The labeling of the elements is totally immaterial; all that counts is the set of m probabilities.

2. For $m = 2$, $U(\frac{1}{2}, \frac{1}{2}) = 1$. This defines the unit of uncertainty, the bit.

3. $\lim_{p \rightarrow 0} U(p, 1 - p) = 0$.

4. $U(p_1, \dots, p_m, 0) = U(p_1, \dots, p_m)$.

5. If P and Q are two distributions and $P * Q$ denotes their joint distribution, $U(P * Q) \leq U(P) + U(Q)$, with $=$ holding if P and Q are independent.

The mathematical conclusion is that

This article was prepared at the instigation of Alan Borneau as a contribution to a Division 1 session at the 2001 Annual Meeting of the American Psychological Association in San Francisco. I have benefited from interchanges on the topic with Peter Killeen and Donald Laming.

Correspondence concerning this article should be addressed to R. Duncan Luce, School of Social Sciences, Social Science Plaza 2133, University of California, Irvine, California 92697-5100. E-mail: rdluce@uci.edu

Table 1
Pattern of Application Topics in Skilling's (1989)
Table of Contents

| Topic | No. of articles |
|--------------------------------------|-----------------|
| Thermodynamics and quantum mechanics | 5 |
| Physical measurement | 6 |
| Crystallography | 5 |
| Time series, power spectrum | 6 |
| Astronomical techniques | 3 |
| Neural networks | 2 |
| Statistical fundamentals | 17 |

$$U(\{p(i)\}) = - \sum_{i=1}^m p(i) \log_2 p(i). \quad (1)$$

Maximal uncertainty about what will be selected occurs when all probabilities are equally likely, that is, $p_i = 1/n$, in which case $U(\{1/n\}) = \ln_2 n$.

Uncertainty measured in this way is thus a single number associated with a finite probability distribution. In some physical contexts, U is identifiable with thermodynamic entropy, and some authors call it that and use the physical notation H rather than U . Other authors use I for information.

This measure is to be distinguished sharply from familiar statistics such as the mean and variance, which are associated not with a probability distribution over a set of elements but with a random variable that maps the set into numbers and, in the process, introduces an ordering of the elements not available in Shannon's context.

The role of information transmission is to reduce uncertainty. Suppose $\{p(i)\}$ is the distribution before information is transmitted and $p(i, j)$ denotes the probability of the joint state of i being transmitted and j received. The conditional probability is defined in the usual way as

$$p(j|i) = \frac{p(i, j)}{p(i)} \quad (p(i) > 0).$$

The posterior uncertainty $U(\{p(j|i)\})$ can be shown to be the original uncertainty less the transmitted information; that is, $U(\{p(j|i)\}) = U(\{p(i)\}) - U(\{p(i, j)\})$. This can be thought of as the posterior uncertainty.

There clearly is at least a conceptual relation to Bayes's theorem (which was really worked out in the form we know it by P. S. LaPlace [1812/1820]). In fact, the relation is far more than just conceptual and has been well developed for a variety of statistical concepts in many papers and books. I return to this linkage later.

Shannon's theory then went on to consider the limitations of channels to transmit information, for which he defined a concept of channel capacity and, given a noisy transmission line, the (often quite elaborate) coding necessary to achieve near-to-perfect transmission. As noted earlier, Shannon strongly preferred the term *communication theory* to *information theory*, but in psychology at least, *information* became the standard term.

Introduction of Information Theory in Psychology

During graduate school at the Massachusetts Institute of Technology and during a brief stint (1950–1953) at the Research Laboratory for Electronics (RLE), I was surrounded by a flurry of ideas that had matured in research laboratories during World War II and shortly thereafter and that have, in fact, all played a significant scientific role since then. They were as follows: information theory, feedback, and cybernetics; networks of various sorts and automata theory; the beginnings of artificial intelligence and of both analogue and digital computers; Bayesian statistics and the theory of signal detectability; Chomsky's developing theory of linguistics; and game and utility theory. Some psychologists had been close to one or another of these developments, and so different ones latched onto different ideas.

A substantial group of psychologists and engineers at RLE focused on information theory and held a very well attended weekly seminar on the topic, which, among its incidental consequences, provided me with an education on the topic. Later in 1954, some of these and related developments were reported at the "Conference on the Estimation of Information Flow," organized by Henry Quastler and held in Monticello, Illinois. The conference was summarized in Quastler's 1955 edited volume *Information Theory in Psychology*.

Perhaps for psychologists, the two defining articles to come out of those meetings were the late William J. McGill's (1954) "Multivariate Information Transmission" and George Miller's (1956) "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." The latter article addressed several phenomena that seem to exhibit information capacity limitations, including absolute judgments of unidimensional and multidimensional stimuli and short-term memory. Comprehensive, but quite different, summaries of the ideas and experiments were given by Attneave (1959) and Garner (1962). Later articles expanded the interest to the relation between mean response times and the uncertainty of the stimuli to which participants were responding. In early experiments, mean response time appeared to grow linearly with uncertainty, but glitches soon became evident. The deepest push in the response time direction was Donald Laming's (1968) subtle *Information Theory of Choice-Reaction Times*, although he later stated: "This idea does not work. While my own data (1968) might suggest otherwise, there are further unpublished results that show it to be hopeless" (Laming, 2001, p. 642).

The enthusiasm—nay, faddishness—of the times is hard to capture now. Many felt that a very deep truth of the mind had been uncovered. Yet, Shannon was skeptical. He is quoted as saying "Information theory has perhaps ballooned to an importance beyond its actual accomplishments" (as cited in Johnson, 2001). And Myron Tribus (1979, p. 1) wrote: "In 1961 Professor Shannon, in a private conversation, made it quite clear to me that he considered applications of his work to problems outside of communication theory to be suspect and he did not attach fundamental significance to them." These skeptical views strike me as appropriate.

Why Limited Application in Psychology?

The question remains: Why is information theory not very applicable to psychological problems despite apparent similarities of concepts? Laming (2001) provided a very detailed critique—far more specific than this one—of a variety of attempts to use information, in particular the appealing concept of channel capacity. He pointed out (p. 639), however, that

Shannon's way of defining the concept requires that not individual signals be transmitted but rather very long strings of them so as to be rid of redundancies. That is rarely possible within psychological experiments. This is definitely part of the answer.

But in my opinion, the most important answer lies in the following incompatibility between psychology and information theory. The elements of choice in information theory are absolutely neutral and lack any internal structure; the probabilities are on a pure, unstructured set whose elements are functionally interchangeable. That is fine for a communication engineer who is totally unconcerned with the signals communicated over a transmission link; interchanging the encoding matters not at all. By and large, however, the stimuli of psychological experiments are to some degree structured, and so, in a fundamental way, they are not in any sense interchangeable. If one is doing an absolute judgment experiment of pure tones that vary in intensity or frequency, the stimuli have a powerful and relevant metric structure, namely, differences or ratios of intensity and frequency measures between pairs of stimuli. And that structure has been shown to matter greatly in the following sense. Substantial sequential effects exist between a stimulus and at least the immediately preceding stimulus–response pair, but with the magnitude of the correlation dropping from close to one for small signal separation in either decibels or frequency to about zero for large separations (Green, Luce, & Duncan, 1977; Luce, Green, & Weber, 1976). Similarly, if one does a memory test, one has to go to very great pains to avoid associations among the stimuli. Stimulus similarity, although still ill understood and under active investigation, is a powerful structural aspect of psychology.

Gradually, as the importance of this reality began to set in, one saw fewer—although still a few—attempts to understand global psychological phenomena in simple information theory terms. Of course, the word *information* has been almost seamlessly transformed into the concept of "information-processing models" in which information theory per se plays no role. The idea of the mind being an information-processing network with capacity limitations has stayed with us, but in far more complex ways than pure

information theory. Much theorizing in cognitive psychology is of this type, now being more or less well augmented by brain imaging techniques.

Before going on, let me note that this incompatibility between either a summary measure of a probability distribution or the distribution itself and the psychological structure of sets of related stimuli is an issue not only for information theory. The unstructured elements of probability theory and the structure of psychological stimuli have made it very difficult indeed to add, in a principled fashion, probabilistic aspects to models of behavior of any complexity at all. The melding, for example, of utility theory and randomness has proved to be as elusive as it is important.

Information Theory and Statistics

As noted earlier, there is a close linkage between Shannon's information measures and statistics that has been very well explored (Diamond, 1959; Jaynes, 1979, 1986, 1988; Justice, 1986; Levine & Tribus, 1979; Mathai, 1975; Skilling, 1989; Zellner, 1988). As Jaynes (1988, p. 281) remarked, commenting on Zellner's linking the information measure directly to Bayes's theorem, "entropy has been a recognized part of probability theory since the work of Shannon 40 years ago. . . . But now we see that there is, after all, a close connection between entropy and Bayes' theorem." Laming (2001) also emphasized the tight connection.

But in contrast to their wide use in economics, Bayesian statistics have not yet been a roaring success in psychology despite many years of being promoted, notably by Ward Edwards. For more than 40 years he has run an annual (February) Bayesian conference that, in recent years, has been held at the Sportsman Lodge in Studio City, California. The major place where Bayesian ideas have flowered quite well in psychology is in decision theory, in which there is quite a natural melding of utility theory ideas with Bayesian updating of the probabilities underlying subjective expected utility. But as an inferential engine in psychology proper, which continues to be haunted by the hypothesis testing paradigm of agriculture and clinical trials, Bayesian methods have so far played very little

role. This may well be unfortunate given their successes in other fields.

There is one notable exception to what I have just stated about the incompatibility of psychological structure and information theory. This is when the number of stimuli m in information theory and the number of hypotheses being considered in a Bayesian analysis is two. In that case, the role of stimulus structure largely disappears, and one can sometimes get away with treating the two stimuli as without structure beyond probabilities of choice. Here Bayesian ideas have played a major role in the widely used theory of signal detectability (Egan, 1975; Green & Swets, 1966; Macmillan & Creelman, 1991; Swets, 1964). Recall that one has among the variables a probability distribution over signal presentations (i.e., p and $1 - p$); one has two conditional probabilities of response, one for each of the possible stimuli; and one has payoffs that can be modeled by the simplest of utility theories. The main role of the theory is to provide a decomposition of the process into two parts, one attributable to inherent sensory properties and the other to decision-making criteria. This is similar in some ways to the information theory decomposition. So, it comes as no great surprise that people have approached the problem from the Bayesian perspective, in the form of classical signal detectability theory, and from that of information theory, and have attempted to relate the two.

Perhaps the most elaborate effort in this direction has been that of Kenneth H. Norwich (1993), who summarized his approach in *Information, Sensation, and Perception*. Although he states his hypothesis quite generally that subjective sensation is basically the measure of uncertainty reduction that occurs when the stimulus is presented, his detailed explorations are largely confined to binary situations, wherein his style closely resembles that of classical 19th-century physical thermodynamics. A more recent comparison of signal detection and information theories is a manuscript authored by Peter R. Killeen and Thomas J. Taylor (2001) titled *Bits of the ROC: Signal Detection as Information Transmission*. And as I suggested earlier, most decision theory in which Bayesian ideas play a role involves binary decisions based on binary data with distinct sources of information being dealt with independently. Laming (2001, pp.

462–463) pointed out, however, that careful analyses of the criterion in signal detection show systematic shifts depending on previous outcomes, which rules out any simple use of information or Bayesian theory in this context; indeed, simple probability matching of responses to presentations is far more predictive.

Conclusions

Laming (2001), despite his detailed critique of almost all attempts to apply information theory in psychology, ended on a surprisingly optimistic note. He wrote: “Information theory provides, as it were, a ‘non-parametric’ technique for the investigation of all kinds of systems without the need to understand the machinery, to model the brain without modelling the neural responses” (Laming, 2001, p. 645).

I am not so optimistic. The fact is that generalizations of either Bayesian or information theory ideas to situations with more than two hypotheses or stimuli remain unfulfilled, and my conjecture is that this is not likely to change soon. This is my brief answer to the question “Whatever happened to information theory in psychology?”

References

- Aczél, J., & Daróczy, Z. (1975). *On measures of information and their characterizations*. New York: Academic Press.
- Aczél, J., Forte, B., & Ng, C. T. (1974). Why the Shannon and Hartley entropies are “natural.” *Advances in Applied Probability*, 6, 131–146.
- Attneave, F. (1959). *Applications of information theory to psychology*. New York: Henry Holt.
- Devlin, K. (2001). Claude Shannon, 1916–2001. *Focus: The Newsletter of the Mathematical Association of America*, 21, 20–21.
- Diamond, S. (1959). *Information and error: An introduction to statistical analysis*. New York: Basic Books.
- Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic Press.
- Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.
- Green, D. M., Luce, R. D., & Duncan, J. E. (1977). Variability and sequential effects in magnitude production and estimation of auditory intensity. *Perception & Psychophysics*, 22, 450–456.
- Green, D. M., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Jaynes, E. T. (1979). Where do we stand on maximum entropy? In R. D. Levine & M. Tribus (Eds.), *The maximum entropy formalism* (pp. 15–118). Cambridge, MA: MIT Press.
- Jaynes, E. T. (1986). Bayesian methods: An introductory tutorial. In J. H. Justice (Ed.), *Maximum entropy and Bayesian methods in applied statistics* (pp. 1–25). Cambridge, England: Cambridge University Press.
- Jaynes, E. T. (1988). Discussion. *American Statistician*, 42, 280–281.
- Johnson, G. (2001, February 27). Claude Shannon, mathematician, dies at 84 [obituary]. *New York Times*, p. B7.
- Justice, J. H. (Ed.). (1986). *Maximum entropy and Bayesian methods in applied statistics*. Cambridge, England: Cambridge University Press.
- Killeen, P. R., & Taylor, T. J. (2001). *Bits of the ROC: Signal detection as information transmission*. Unpublished manuscript.
- Laming, D. R. J. (1968). *Information theory of choice-reaction times*. New York: Academic Press.
- Laming, D. (2001). Statistical information, uncertainty, and Bayes’ theorem: Some applications in experimental psychology. In S. Benferhat & P. Besnard (Eds.), *Symbolic and quantitative approaches to reasoning with uncertainty* (pp. 635–646). Berlin: Springer-Verlag.
- LaPlace, P. S. (1820). *Theorie analytique des probabilités* [Analytic theory of probability] (2 vols.; 3rd ed., with supplements). Paris: Courcier. (Original published 1812; reprints available from Editions Culture et Civilisation, 115 Avenue Gabriel Lebron, 1160 Brussels, Belgium)
- Levine, R. D., & Tribus, M. (Eds.). (1979). *The maximum entropy formalism*. Cambridge, MA: MIT Press.
- Luce, R. D. (1960). The theory of selective information and some of its behavioral application. In R. D. Luce (Ed.), *Developments in mathematical psychology* (pp. 5–119). Glencoe, IL: Free Press.
- Luce, R. D., Green, D. M., & Weber, D. L. (1976). Attention bands in absolute identification. *Perception & Psychophysics*, 20, 49–54.
- Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user’s guide*. Cambridge, England: Cambridge University Press.
- Mathai, A. M. (1975). *Basic concepts in information theory and statistics: Axiomatic foundations and applications*. New York: Wiley.
- McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, 19, 97–116.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.

- Norwich, K. H. (1993). *Information, sensation, and perception*. San Diego, CA: Academic Press.
- Quastler, H. (1955). *Information theory in psychology*. Glencoe, IL: Free Press.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656.
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.
- Skilling, J. (1989). *Maximum entropy and Bayesian methods: Cambridge England 1988*. Dordrecht, the Netherlands: Kluwer.
- Swets, J. (Ed.). (1964). *Signal detection and recognition by human observers*. New York: Wiley.
- Tribus, M. (1979). Thirty years of information theory. In R. D. Levine & M. Tribus (Eds.), *The maximum entropy formalism* (pp. 1–14). Cambridge, MA: MIT Press.
- Zellner, A. (1988). Optimal information processing and Bayes' theorem. *American Statistician*, 42, 278–284.

Received February 5, 2002

Accepted March 15, 2002 ■