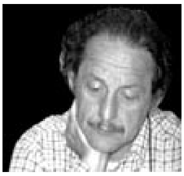

"Current Comments"

Is the Ratio Between Number of Citations and Publications Cited a True Constant?

February 9, 1976 Number 6

The first experimental citation indexes of scientific literature were compiled almost twenty years ago. By the time we had completed the 1964 *Science Citation Index®* (*SCI®*), we were aware that there was a surprising near-constancy in the ratio of 1.7 between references processed each year and the number of different items cited by those references. Very early we began to call the 1.7 ratio *the citation constant.* We have frequently discussed this number at ISI® , and used it for various estimates informally. I have never attempted any rigorous analysis of it, though it has been mentioned in some of these essays.

If you examine any annual *SCI* Guide, this 'constant' is readily apparent in the chronological statistical analysis provided. As the number and type of journals covered by *SCI* has grown, the ratio has changed slightly. Perhaps my own mathematical and statistical naiveté has made it possible for me to suffer in silence so many years while I wondered about the probability of a true constant. This does not mean that a number of people have not concerned themselves with regularities in citation data. Probably Derek Price was the first in recent times to publish on the subject, though he himself often cites the pioneer studies of A.J. Lotka in this connection.[1] Using *SCI* data, Price showed in 1965 how many papers will be cited *n* times.[2]

Authors often ask me how significant it is that a paper has been cited ten times in one year. They are surprised to learn that less than 25% of all papers will be cited ten times in all eternity! How categorical can you get? As you have seen in the various lists of highly cited papers we have published in *Current Contents®* (*CC®*), any paper cited ten times in one year is *ipso facto* significant. Occasionally there is an anomaly. But a paper cited ten times in each of two successive years is well on its way to citation stardom. Whether the author is on the way to immortality depends on how well he or she does in other papers.

But let us return to *Garfield's 'constant'*. Why is it 1.7? A Dutch scientist, M.C. Gomperts, studied the constancy of citation in the special field of Chladni's plates,[3] but he failed to take note of our magical 1.7. A.E. Cawkell reviewed Gompert's paper in some detail, and reminded him of the *SCI* statistics cited above.[4] At that time our 'constant' was given as 1.65.

A few years ago I decided to explore this mystery. I was scanning a yearly report that shows the average number of references per article for each journal we cover. It may surprise some readers to learn that there is substantial variation--by discipline, and as recently noted,[5] by geographical origin. The average chemistry or physics article, for example, contains about twenty references, while math articles contain less than ten. Disregarding language and discipline, the average article in 1974 contained thirteen references. The average article in a journal published in France contained 8.8 references. Review articles naturally contain more references than others. Indeed the addition of a substantial number of review-type journals to our database could affect the ratio significantly. A recent Soviet article gives some interesting comparative data on number of references, with particular attention to math journals.[6]

It is not at all obvious why there should be a big difference between math and physics, between medicine and chemistry, between French and non-French. In the case of math--as is likely to be the cause when the field is new or small--it is not because there is comparatively little literature to cite. The fewer references in math articles may say something about the superspecialization in mathematics that is impossible in disciplines like biochemistry. Or it may say something about the literary style of mathematicians, engineers, and others who write more esoterically and possibly with less needful regard for their topic's history.

One can construct a simple model of the literature with certain reasonable assumptions about the size of the existing literature, the length of papers and number of references, etc. This is the sort of thing Price did when he concluded that forty papers make up a research front.[2]

Assume that the total citable literature in a field consists of 10,000 papers. Assume that the literature is growing about 10% a year--that is, that about 1000 new papers appear each year. If the average number of references per paper is ten, these 1000 new papers will produce 10,000 references. However, we know or can assume that only about half the papers ever published will be cited in any particular year. Thus in our model, 5000 papers will be cited 10,000 times and Garfield's 'constant' would be 2.0.

If you want more realistic numbers for the model, you'll find that the extant literature is closer to 7.5 million items, that 3.0 million (about 42%) are cited each year, that the average number of references per paper is about twelve or thirteen, and that the annual growth rate is about 7%.

For those of you who like equations, let L be the extant literature, U the utilization factor (the percentage of the literature used or cited each year), R the number of references per paper, and G the growth rate. GL will then be the number of new papers published each year, UL will be the number of papers cited each year, GLR will be the number of references processed each year, and $GLR/UL = C$, the citation ratio. Since L cancels out of numerator and denominator, we have $C = GR/U$.

If the growth rate (G) is 7%, the number of references per paper (R) is 12, the utilization factor (U) is 0.5, then C turns out to be 1.7. Since we can't be certain about the size of the extant literature, we can only speculate on the accuracy of U. For cocktail-party conversation, one can assume that about half the literature must be cited each year if the growth rate is 7%, or possibly a third of the literature is cited if the growth rate is 5% a year.

Using the simple model, what might one expect for fields like mathematics and molecular biology? Certainly they are quite different. Doubling the growth rate will double the 'constant'. Doubling the number of references per paper will have a similar effect. Thus, a combination of these effects will inevitably cause the average paper in molecular biology to have a higher impact than a paper in mathematics. However, if a field consistently shows more references per paper (thus increasing R), there is perhaps a corresponding growth in the utilization factor. If a fast-moving field tends to ignore the older literature, then its utilization factor could even decline while having a higher than average number of references per paper. This would show up in what we have called the *immediacy factor.*

If one wants to obtain the appropriate ratio for a particular field, it is critical to define the limits of that field. The cluster represented by a particular journal or group of journals ought to be sufficient for most purposes. However, we can get similar results when we segregate papers by means of other objective clustering methods. In this way, one can compare the citation records of individuals and of papers within that field or cluster.

This presumably helps answer the question, "Did scientist X or paper X have as much impact on the field as did scientist Y or paper Y?"

It may seem only a parlor game to ask, regardless of citation analysis, whether Einstein had as much impact on physics as did Mendel on biology. But I'm told that questions like these are constantly debated by philosophers and historians of science. After all it isn't that long ago that 'scholars' were debating how many angels could dance on the tip of a needle.

Obviously a changing number cannot be called a constant. But if the *SCI* were a real random sample of the total literature or achieved 'complete' coverage, we then would observe a constant, I believe, or at least be able to explain why we didn't.

The same kind of discussion above on one year's data can be applied to longer periods. For example, consider the period covered by the *SCI Five-Year Cumulation 1965-1969*. For the period as a whole, the ratio described above turns out to be 2.55[7] Over the course of those five years, 1000 source journals were added to the coverage of the *SCI,* and there was a 50% increase in the number of articles covered. During that time almost 6.5 million different articles and books were cited in the processing of almost 17 million references. Despite all this, the annual values of the ratio for the five years 1965-1969 were 1.65, 1.65, 1.66, 1.67, 1.67. Can you blame me for suspecting that behind all this there lurks a constant, whatever you choose to call it.

1. **Lotka A J.** The frequency distribution of scientific productivity. *J. Wash. Acad. Sci.* 16(12):317-23, 1926.

2. **Price D J D.** Networks of scientific papers. *Science* 149:510-15, 1965.

3. **Gomperts M C.** The law of constant citation for scientific literature. *J. Documentation* 24(2):113-17, 1968.

4. **Cawkell A E.** Documentation notes: citation practices. *J. Documentation* 24(4): 229-304, 1968.

5. **Garfield E.** Journal citation studies. 23. French journals--what they cite and what cites them. *Current Contents* (*CC*) No. 4, 26 Jan 1976, p. 5-10.

6. **Klement'ev A F.** O nekotorykh kharakteristikakh tsitiruemosti publikatsii po matematike (po materialam otechestvennykh izdanii) [Citation characteristics of mathematics publications (based on Soviet sources)]. *Nauchno-Tekhn. Informatsiya Ser. 1* (7):33-34, 1975.

7. *Science Citation Index Five-Year Cumulation 1965-1969.* 17 vols. (Philadelphia: Institute for Scientific Information, 1971), vol. 1, p. 17.