Cumulative Advantage and Success-Breeds-Success: The Value of Time Pattern Analysis

John C. Huber
Institute for Invention and Innovation, 500 E. Anderson Ln., Suite 238-X, Austin, TX 78752-1207.
E-mail: jchuber@ccsi.com

Many different theoretical models can be made to fit empirical informetric data. For the case of the distribution of papers across authors, the Success-Breeds-Success or Cumulative Advantage model is a popular candidate. This article shows that examination of the time pattern of production allows independent evaluation of the component processes that generate the distribution of papers across authors. Specifically for inventors, the Cumulative Advantage model for increasing rate of production with experience is not confirmed. Furthermore, the distribution of individual production is Poisson and the distribution of the rate of production across the population fits the Gamma distribution. Thus, the non-uniform giftedness model is more appropriate for inventors.

Introduction

The principle of Cumulative Advantage or Success-Breeds-Success has a long tradition in informetrics. Simon (1957) first proposed such a model, later named the Simon-Yule process. Many other authors (Chen, 1989; Chen, Chong, & Tong, 1994; Fedorowicz, 1982; Price, 1976) added to this tradition. This process produces mathematical functions showing that the primary informetric laws (Lotka, Price, Pareto, Zipf, etc.) are related (Bookstein, 1990; Fedorowicz, 1982; Tague, 1981). Most recently Egghe and Rousseau (1995) and Glanzel and Schubert (1995) have developed more generalized Success-Breeds-Success models. Clearly the principle of Success-Breeds-Success and Cumulative Advantage is alive and well as a model for general studies in informetrics.

The Simon-Yule framework is often expressed in such a way as to model the growth of the population of authors publishing in a field. However, Schubert and Glanzel (1984), Allison, Long, and Krauze (1982), and Rao (1980) have expressed it explicitly as a model for the increasing rate of publication for individual authors. This model implies that the shape of the distribution of production across a population arises from the distribution of production within each individual's career. It is this application which will be investigated in this article. In the general case, the Simon-Yule framework is expressed as sources which produce items. For clarity, we will express it as authors who produce publications in keeping with the application chosen here.

Burrell and Fenton (1993) clearly described the two component processes that must be present to generate data that fits the distribution of publications across authors. Readers who are unfamiliar with the typical distribution of publications across authors may wish to examine Figure 1. One of the component processes is the time pattern of individual production (publications counted over the years of a career). The other is the distribution of productivity over the population of authors, analogous to the distribution of IQ over the general population.

However, there are models other than Cumulative Advantage which also fit the empirical laws. One popular model (Allison, 1980a; Burrell & Fenton, 1993) assumes that the rate of publication for an individual is stable over his or her career and that the distribution of the rate of publication over the population determines the distribution of publications over authors. This model generally has been nameless; here it will be denoted ''non-uniform giftedness.'' Thus, the Cumulative Advantage and non-uniform giftedness models reflect nearly opposite views or opposite ends of a continuum. One objective of this article is to demonstrate a method of determining which model is a better fit for a given set of empirical data, or if a mixture of the models is appropriate.

Lotka (1926) established the field of investigating the distribution of production of scientific publications over authors. Lotka found that the number of authors fell as the square of the number of papers produced, since known as Lotka's Law. It has since been generalized to be an inverse power function with exponent ranging from about 1.8 to 3.8 (Pao, 1986). Since then, there has been consid-

Received April 29, 1997; revised July 18, 1997; accepted July 18, 1997.

© 1998 John Wiley & Sons, Inc.

erable focus on creating models which can generate mathematical functions that fit the empirical observations. At various times, negative binomial (Allison, 1980b; Land, McCall & Nagin, 1996), Poisson, (Land, McCall, & Nagin, 1996), Contagious Poisson (Allison, Long, & Krauze, 1982), Generalized Poisson (Consul, 1989, pp. 129–135), and Generalized Inverse Gaussian Poisson (Burrell & Fenton, 1993) distributions have been fit to various new and old data sets with varying degrees of success. The causes of these discrepancies have not been resolved. Another objective of this article is to suggest some potential causes for these discrepancies.

In general, an individual's time pattern of publication will not be exactly constant. Burrell & Fenton (1993), Sichel (1992), Allison (1980b), and Bookstein (1976), have explicitly assumed that an individual's time pattern of publication follows a Poisson distribution. But there appears to be no published empirical confirmation of the Poisson distribution until Huber's (in press, 1997) recent work. Whether Poisson or not, the parameter(s) of the distribution (e.g., mean and variance) may be constant in time and experience (homogenous, in mathematical statistic nomenclature) or may vary (heterogeneous). Individual Cumulative Advantage having increasing rate of publication is an example of a heterogeneous distribution. This article will demonstrate a method to determine which one fits an empirical data set.

In addition, there is the distribution of these individual-descriptive parameters across the population of authors. Some studies have assumed a Gamma distribution of the Poisson parameter for individual production. This is a convenient assumption since the result from this mixture is the negative binomial distribution with parameters that can directly produce those of the component Gamma distribution (Johnson, Kotz, & Kemp, 1993, p. 204). However, this is one of the few such combinations of component distributions which are so nicely behaved. More commonly, such combinations lead to complicated distributions with weaker estimation procedures and weaker connection to the component distributions (Johnson, Kotz, & Kemp, 1993, pp. 326–335). Also, there appears to be no published empirical confirmation of the Gamma distribution describing informetric data.

In this article, we will show that measuring the time pattern of individual production is vital to resolving the conflicting claims of various models and their application to the general Lotka Law empirical data. Firstly, measuring the time pattern of individual publication can independently test the goodness-of-fit of the Poisson distribution, or any other distribution. It is straightforward to determine if the distribution is constant over time and experience or not. Thus, Success-Breeds-Success and Cumulative Advantage can be evaluated directly. Secondly, once this distribution has been determined for each author, then the distribution of those parameters across a sample can be calculated and parameters for the population can be estimated. Thus, goodness-of-fit of the Gamma distribution, or any other distribution, can also be evaluated directly.

The empirical data set evaluated here is a random sample of inventors. Tague and Nicholls (1987) have pointed out that many historical, legacy data sets are not random samples. Therefore arguing for a particular model based on the number of historical data sets having a good fit is not generally persuasive. For the inventor data set, the distribution of individual publication is found to fit the Poisson distribution. Also, the Poisson parameter is constant over the time interval studied. Thus, Success-Breeds-Success and Cumulative Advantage can be rejected without recourse to fitting a mathematical function to the distribution of papers over authors. Also, for this data set, the distribution of the Poisson parameter (average rate of production) over the population fits a Gamma distribution.

Method

In informetrics, most of the empirical analysis has been on scientific publications. What is the advantage of analyzing inventors? There are four reasons why inventors should be included in informetric analysis.

In any empirical study, it is important to avoid sources of uncontrolled variance. Such excess variance obscures relationships among the variables being studied and leads to ambiguous conclusions. One of the concerns about informetric studies of scientific publications is the variable quality of different journals (Price, 1963, pp. 67–69). The quality standards may vary across journals within the same field and also across fields. However, for a patent to be issued, the invention must be new to the world, useful in a practical sense, and unobvious to persons reasonably skilled in the art, i.e., the specific technology (Burge, 1984, pp. 43–46). This standard is applied uniformly to all patents. Thus, there tends to be a more consistent standard than for scientific publication. However, we do not know at this time if this truly reduces variance.

Another reason to study patents is the abundant biographical information provided about the inventors. Those who perform index searches know that it is difficult to determine if the Jane Q. Jones of a particular publication is the same one who has other known publications. The movement of academic authors between universities and between fields aggravates this discrimination problem. On the other hand, inventors tend to stay in one place and in one field. Firstly, most inventors are employed in industry, since it takes substantial investment to develop and market new products. Secondly, they tend to stay with one employer because of non-competition and confidential information issues. Thirdly, they tend to stay in the same field, or evolve slowly. Accurately identifying individual inventors is not a minor issue. There are 27 different Robert L. Smith's in the U.S. Patent database. More-

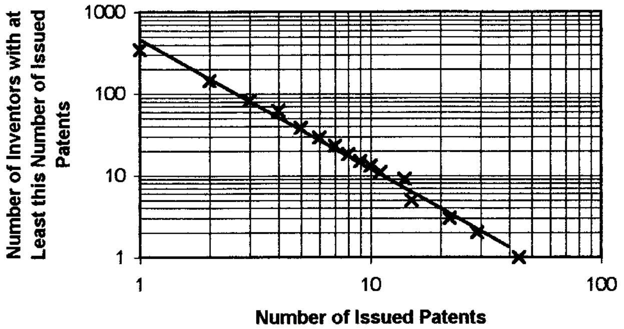

FIG. 1. Pareto diagram for inventors (regression line slope = 1.6; correlation = 99.5%.

over, the distribution of output is highly skewed, with most inventors having few patents, similar to scientific publications. Therefore, most individuals with identical names will have few publications. When they are inappropriately aggregated as one ‘‘individual’’ there will be an erroneous decrease in the number of authors with few publications and an erroneous increase in the number of authors with moderate publications. Thus, mistakes in discriminating among authors leads to a ‘‘hump’’ in middle output, deviating from the straight line representation of the log of papers versus the log of authors representative of informetric laws. Another reason to study patents is that the total inventory of approximately one million patents issued by the U.S. Patent Office since 1975 can be searched electronically on CD-ROMs, and on the Internet. And last, but by no means least, patents are an important informetric source that has been little studied (Acs & Audretsch, 1989; Narin, 1994; Narin & Breitzman, 1995; Schmoolker, 1966, pp. 197–215; Shockley, 1957).

How the Sample Was Obtained

Unfortunately, the electronically-accessible U.S. Patents database does not lend itself to drawing random samples of inventors. Rather, it is designed to retrieve individual patents based on matching specific information. However, this problem was addressed by assuming that the distribution of inventors in the patents database is the same as the distribution of names in the U.S. population. Another way of saying this is that there is no special relationship between an inventor’s name and his or her patent output. A representative sample of the U.S. population may be taken from a telephone book. The telephone book chosen was the residential White Pages from Austin, Texas. Austin was chosen because the vast majority of its residents are immigrants from elsewhere in the United States. Thus, any ethnic bias is minimized, since many other cities have large ethnic populations and there may be some relationship between ethnicity and proclivity to patent, though none is known at this time. Ninety-nine names were drawn at random by choosing the name in the upper right corner of every tenth page in the telephone book. Middle initials, since they were rarely present, were ignored. Then the U.S. patents database since 1975 was searched for inventors with those names. Inventors working for companies outside the U.S. were not included because companies may be more selective about filing patents outside their home markets. Thus, a non-U.S. inventor may have only a fraction of his or her patents filed in the U.S. The resultant sample contained 346 inventors who had 854 patents.

The Method of Analysis

For each individual inventor, the output of patents was separated into the years since the first patent. The patents in each year were counted. This resulted in a time sequence of patents, such as 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, for a career duration of 10 years.

A requirement for most statistical procedures is that the observed data be obtained from random samples. In addition to the random sample of inventors discussed above, we are also interested in the randomness of production within individual careers. Here, the number of patents for each year is a sample for that individual inventor. A common method of determining randomness is the test of runs (Hamburg, 1977, pp. 352–354). A few examples will help to clarify this procedure. Consider three inventors, each with exactly five patents spread over 10 years duration. Inventor A has the following sequence of patents per year: 1, 1, 1, 1, 0, 0, 0, 0, 0, 1. In this simplified example, a run is a sequence of one or more years with the same number of patents. Thus, inventor A has three runs. Inventor B has the following sequence of patents per year: 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, exhibiting nine runs. Inventor C has the following sequence of patents per year: 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, exhibiting seven runs. For five patents over 10 years, the expected number of runs is six, and the probabilities of other values can be calculated directly (Brownlee, 1965, pp. 225–226; Lehmann & D’Abrera, 1975, p. 314). Inventor A’s three runs has 0.03 probability, being too few runs and indicating a non-random initial surge pattern. Inventor B’s nine runs also has 0.03 probability, being too many runs and indicating a non-random consistent pattern. Inventor C’s seven runs has probability 0.19, indicating a random pattern. Yet another example is to consider the same sequences above as coin-tossing, with the 1s and 0s representing heads and tails.

The generalized runs test requires two-value input data. In this analysis, the two values were determined by whether or not the number of patents in each year exceeded the inventor’s average yearly production. For most of these inventors, when the observed number of runs differs from the expected value by 20%, the probability drops below 0.10. This is equivalent to having about 10% of the years with above average patents being moved either adjacent to another above-average year, or being separated from a previously-adjacent above-average year.

This is clearly a common occurrence in the course of a career. Having twice as many such deviations may be judged to be a strong indication of non-randomness and, for most of the inventors in these samples, results in the probability dropping below 0.025. Thus, a randomness selection criterion of 0.025 is reasonable.

The distribution of patents was tested for goodness-of-fit to the Poisson distribution by the Chi-Square method (Rice, 1995, pp. 299–310). For most of these inventors, moving 20% of his or her patents into an adjacent year that already contains a patent (or moving them out) causes the probability to drop below 0.10. This is clearly a common occurrence in the course of a career. Having twice as many such deviations may be judged to be a strong indication of non-Poisson-ness and, for most of the inventors in these samples, results in the probability dropping below 0.025. Thus, a goodness-of-fit selection criterion of 0.025 may be reasonable. The Poisson distribution also describes a number of other common processes, including radioactivity, automobile traffic, occurrence of telephone calls, and occurrence of accidents (Rice, 1995, p. 43–44), which are all discrete events without negative values.

Results

Informetric Behavior

As mentioned above, the sample has many inventors with few patents and few inventors with many patents. The cumulative distribution of patents over inventors is shown in Figure 1. This highly-skewed distribution was also observed by many other studies (Nicholls, 1986; Pao, 1986; Price, 1963, p. 40; etc.) for published scientific papers. It is not surprising that the output of inventors should have the same skewed distribution as shown by others (Narin, 1994; Narin & Breitzman, 1995; Shockley 1957). It is variously named the Pareto, Zipf, and Zeta distribution, and is also used to describe distributions of individual incomes, word frequency, machine reliability, city size, and many other phenomena (Johnson, Kotz, & Kemp, 1993, pp. 465–471). As mentioned in the introduction, it has been shown that the general empirical laws (Lotka, Price, Pareto, Zipf) of informetrics are related.

Selected Inventors

Clearly, inventors with just one patent cannot be examined for randomness, or fit to a Poisson distribution or Cumulative Advantage. Even for inventors with a career duration of 4 years, almost half of them exhibited a time pattern of 1, 0, 0, 1 which is not very satisfactory for such tests. Therefore, only inventors with career durations of 5 years or more were selected for further analysis. There were 70 inventors meeting these criteria.

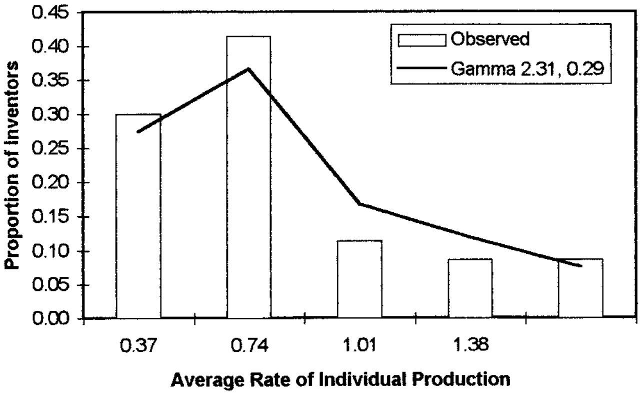

FIG. 2. Distribution of Poisson parameter.

Discussion

For a test statistic of 0.025, 100% of the inventors exhibited random runs, and 97% (68/70) had time patterns of production that fit the Poisson distribution. This means the Poisson parameter was stable over the inventors' career covered by the 21-year time interval. Thus, there is no significant evidence of Success-Breeds-Success or Cumulative Advantage for this sample. This conclusion should not be taken to extend to other samples, domains, and fields. All it does is demonstrate that these conclusions can be arrived at by the direct analysis of a reasonable number of authors.

A natural question is whether there is a small effect. Schubert and Glanzel (1984) calculated a small Cumulative Advantage coefficient of 3–5% per year. However, those results were obtained from distribution parameters estimated from a goodness-of-fit test of a three parameter model and did not use time pattern information. For a comparison using time pattern information, the yearly productions of the inventors with 10 years of data were added together and fit with a regression line. The slope was −0.013 with standard error of 0.017, t-statistic of −0.77, and $p = 0.46$, indicating no significant Cumulative Advantage. Many authors (McCrae, Arenberg, & Costa, 1987; Horner, Rushton, & Vernon, 1986; Simonton, 1997; Stephan & Levin, 1992; and others) have made extensive studies of age and production, and have found a general curvilinear trend (upwards in early adulthood and declining after middle age), but the trend only explains 7–10% of the variance of yearly production. These analyses detected no general upwards trend. Simonton (1997) offers specific analysis refuting the Cumulative Advantage principle.

The distribution of the Poisson parameter for this sample is shown in Figure 2 as vertical bars. The skewed-right shape resembles the Gamma distribution, and maximum likelihood estimates of the parameters were calculated by standard methods (Johnson, Kotz, & Balakrishnan, 1994, pp. 360–365). The estimated distribution is plotted as a broken line. Chi Square Goodness-of-Fit test confirms the Gamma model fits the distribution of rate over the population at $p = .018$. Thus, the Gamma distribution can be supported by the direct analysis of a reasonable number of authors. Similarly, the uniform distribution is rejected at $p = .54$. This large difference in goodness-of-fit lends strong support to the non-uniform giftedness model. If the contrast were not so marked, it is obvious that a mixture of models can also be tested for goodness-of-fit. This conclusion should not be taken to extend to other samples, domains, and fields. All it does is demonstrate that these conclusions can be arrived at by the direct analysis of a reasonable number of authors.

The author is indebted to an anonymous reviewer who cautioned that lacking an observable increasing rate does not necessarily lead to rejecting the Cumulative Advantage model. There may be debilitating processes that offset Cumulative Advantage. Some examples of debilitating processes are—Declining physical and mental powers with age, senior faculty being drawn on for administrative roles, transition from journal articles to books.

Conclusions

Many informetric processes, especially the distribution of papers across authors, arise from the interaction of two or more component processes. The goodness-of-fit of a model's mathematical function to the author–paper production distribution is a necessary, but not sufficient condition for acceptance. Since it has been shown that a number of different models have adequate goodness-of-fit to the empirical laws, it is important to distinguish between the candidate models on more than goodness-of-fit to the distribution of papers across authors.

This article has shown that the method of time pattern of publication for individual authors can be used to discriminate between Cumulative Advantage and non-uniform giftedness models. Furthermore, each component distribution can be tested independently and directly, without needing to include assumptions about the other component distributions.

In general, historic data sets have not collected time pattern data, and so this technique often cannot be used retrospectively. However, since many of these legacy data sets are not random samples and/or their methods may be suspect (Pao, 1985), their applicability is questionable anyway. An important part of any empirical method, including time pattern analysis, is to ensure a random sample is being studied. The effort to create an appropriate-sized random sample is not excessive. The sample in this study was created in less than 30 hours.

The generality of the model described by Burrell and Fenton (1993) makes it clear that wide varieties of distributions of production should be the norm, rather than the exception. For an active population of contributors, there are two component distributions. One is the consistency of individual production, shown here to be Poisson for inventors. Another is the distribution of individual capacity, shown here to be Gamma for inventors. Of the many possible explanations for the discrepancies in empirical studies, an important candidate is the distribution of individual capacity. Thus, deviations from Lotka's law, assuming random samples and other correct procedures, may be caused by differences in the component distributions. As shown in this article, the component distributions can be, and should be, examined independently, thereby deriving their shapes directly from the time pattern of production.

References

Acs, Z. J., & Audretsch, D. B. (1989). Patents as a measure of innovative activity. *Kyklos, 42,* 171–180.

Allison, P. D. (1980a). Estimation and testing for a Markov model of reinforcement. *Sociological Methods & Research, 8,* 434–453.

Allison, P. D. (1980b). Inequality and scientific productivity. *Social Studies of Science, 10,* 163–179.

Allison, P. D., Long, J. S., & Krauze, T. K. (1982). Cumulative advantage and inequality in science. *American Sociological Review, 47,* 615–625.

Bookstein, A. (1976). The bibliometric distributions. *Library Quarterly, 46,* 416–423.

Bookstein, A. (1990). Informetric distributions, part II: Resilience to ambiguity. *Journal of the American Society for Information Science, 41,* 376–386.

Brownlee, K. A. (1965). *Statistical theory and methodology: In science and engineering.* New York: John Wiley & Sons.

Burge, D. A. (1984). *Patent and trademark practices (2nd ed.).* New York: John Wiley & Sons.

Burrell, Q. L., & Fenton, M. R. (1993). Yes, the GIGP really does work—And is workable! *Journal of the American Society for Information Science, 44,* 61–69.

Chen, Y. S. (1989). Analysis of Lotka's law: The Simon-Yule approach. *Information Processing & Management, 25,* 527–544.

Chen, Y. S., Chong, P. P., & Tong, M. Y. (1994). The Simon-Yule approach to bibliometric modeling. *Information Processing & Management, 30,* 535–556.

Consul, P. C. (1989). *Generalized Poisson distributions: Properties and applications.* New York: Marcel Dekker, Inc.

Egghe, L., & Rousseau, R. (1995). Generalized success-breeds-success principle leading to time-dependent informetric distributions. *Journal of the American Society for Information Science, 46,* 426–425.

Fedorowicz, J. (1982). The theoretical foundation of Zipf's law and its application to the bibliographic database environment. *Journal of the American Society for Information Science, 33,* 285–293.

Glanzel, W., & Schubert, A. (1995). Predictive aspects of a stochastic model for citation purposes. *Information Processing & Management, 31,* 69–80.

Hamburg, M. (1977). *Statistical analysis for decision making.* New York: Harcourt, Brace, Jovanovich.

Horner, K. L., Rushton, J. P., & Vernon, P. A. (1986). Relation between aging and research productivity of academic psychologists. *Psychology and Aging, 1,* 319–324.

Huber, J. C. (in press). Invention and inventivity as a special kind of creativity, with implications for general creativity. *Journal of Creative Behavior.*

Huber, J. C. (in press). Invention and inventivity is a random, Poisson process: A potential guide to analysis of general creativity. *Creativity Research Journal.*

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions (2nd ed., Vol. 1).* New York: John Wiley & Sons.

Johnson, N. L., Kotz, S., & Kemp, A. W. (1993). *Univariate discrete distributions, (2nd ed.).* New York: John Wiley & Sons.

Land, K. C., McCall, P. L., & Nagin, D. S. (1996). A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models. *Sociological Methods & Research, 24,* 387–442.

Lehmann, E. L., & D'Abrera, H. J. M. (1975). *Nonparametrics: Statistical methods based on ranks.* San Francisco: Holden-Day.

Lotka, A. J. (1926). The frequency of distribution of scientific productivity. *Journal of the Washington Academy of Science, 16,* 317–323.

McCrae, R. R., Arenberg, D., & Costa, P. T., Jr. (1987). Declines in divergent thinking with age: Cross-sectional, longitudinal, and cross-sequential analyses. *Psychology and Aging, 2*, 130–137.

Narin, F. (1994). Patent bibliometrics. *Scientometrics, 30,* 147–155.

Narin, F., & Breitzman, A. (1995). Inventive productivity. *Research Policy, 24,* 507–519.

Nicholls, P. T. (1986). Empirical validation of Lotka's law. *Information Processing & Management, 22,* 417–419.

Pao, M. L. (1985). Lotka's law: A testing procedure. *Information Processing & Management, 21,* 305–320.

Pao, M. L. (1986). An empirical examination of Lotka's law. *Journal of the American Society for Information Science, 37,* 26–33.

Price, D. J. S. (1963). *Little science, big science.* New York: Columbia University Press.

Price, D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27,* 292–306.

Rao, I. K. R. (1980). The distribution of scientific productivity and social change. *Journal of the American Society for Information Science, 31,* 111–122.

Rice, J. A. (1995). *Mathematical statistics and data analysis.* Belmont CA: Duxbury Press.

Schmoolker, J. (1966). *Invention and economic growth.* Cambridge MA: Harvard University Press.

Schubert, A., & Glanzel, W. (1984). A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics, 6,* 149–167.

Shockley, W. (1957). On the statistics of individual variations of productivity in research laboratories. *Proceedings of the IRE, 45,* 279–290.

Sichel, H. S. (1992). Anatomy of the generalized inverse Gaussian-Poisson distribution with special applications to bibliometric studies. *Information Processing & Management, 28,* 5–17.

Simon, H. A. (1957). *Models of man: Social and rational.* New York: John Wiley & Sons.

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review, 104,* 66–89.

Stephan, P. E., & Levin, S. G. (1992). *Striking the mother lode in science: The importance of age, place, and time.* New York: Oxford University Press.

Tague, J. (1981). The success-breeds-success phenomenon and bibliometric processes. *Journal of the American Society for Information Science, 32,* 280–286.

Tague, J., & Nicholls, P. (1987). The maximal value of a Zipf size variable: Sampling properties and relationship to other parameters. *Information Processing & Management, 23,* 155–170.