GLORIA S. CLINE

Application of Bradford's Law to Citation Data

This study serves as a test of the two formulations of Bradford's law, verbal and graphical, using 5,628 citations to journal literature referenced in College & Research Libraries *and* Special Libraries, *1940 through 1974. The data are divided into seven five-year spans so that comparisons can be made between the calculated percentage errors for each journal during each five-year period. In addition, trends in citation patterns are identified. Neither the verbal nor the graphical formulation provides results that are clearly more consistent with the practical situation.*

INTRODUCTION

The intellectual base of any discipline is revealed in its journal literature, which serves, among other things, as a vehicle for disseminating information, introducing innovations, and reporting the findings of research in the field. In recent years, bibliometric techniques have been used widely to identify the characteristics of the journal literature of many different subject areas, but most often researchers have concentrated the use of these techniques on the literature from various branches of science. Library periodical literature has seldom been subjected to such intense scrutiny and when appraisals have been made, comments generally have been negative. In 1967, for example, Katz conducted a survey of the attitudes of library science faculty and students toward their professional literature and summarized their comments by saying, "Library literature is timid, rotten, unimaginative, vague, repulsive, and debased."[1] Moon described the literature as a "stream of garbage."[2] With these thoughts in mind, it seems obvious that librarians must look back at their literature and produce evaluative accounts of any changes that may have occurred in the principal library science journals to refute these negative statements. Bibliometric techniques present themselves as a key to objective evaluation.

Since the basic data for many of the bibliometric techniques are citations, the question rapidly arises concerning which of the library periodicals are likely subjects for investigation. Until recently, the national general-interest journals such as *Library Journal* and *American Libraries* published articles having no references whatever. Others like *Library Quarterly* were thought to be too scholarly to have wide appeal. Appearing between the two extremes were *College & Research Libraries (C&RL)* and *Special Libraries* (SL), both of which had been published for many years, thus permitting retrospective analysis, had solid reputations, and had wide readership. Bibliometric techniques could certainly be applied to the citations listed in these journals in order to identify changes that had occurred through the years.

Although one of the leading bibliometric techniques, Bradford's law, lends itself to such a study, it will require some explication.[3] First published in 1948, Bradford's law has been used to test the completeness of a bibliography, to describe the characteristics of various subject literatures,[4]

Gloria S. Cline is assistant director, University Libraries, the University of Southwestern Louisiana, Lafayette.

and even to describe the interaction between book users and books available for use in a library.[5] Librarians generally relate the Bradford distribution to "zones" of productivity and appreciate the fact that the nuclear zone identifies the most productive journals in a subject area. Here in practical terms Bradford's law aids in selection and collection development. However, when one goes beyond this point to discuss the difference between the verbal and graphical formulations of Bradford's law, most librarians become confused, uncertain as to just what the difference is. This study will attempt to explain the two formulations of Bradford's law, verbal and graphical, and then, utilizing Wilkinson's formulas, will test the two to determine which provides a better fit for citation data taken from *C&RL* and *SL* over the thirty-five-year period 1940 through 1974.[6] By dividing the data into five-year spans, it will also be possible to identify changes in the literature published in *C&RL* and *SL* during this time.

BACKGROUND

Basically, Bradford's law states that a quantitative relation exists between journals and the papers they publish. The oft-quoted law, now recognized as the verbal formulation, represents Bradford's theory rather than his observations and reads:

> If scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject, and several groups or zones containing the same number of articles as the nucleus, where the number of periodicals in the nucleus and succeeding zones will be 1:a:a². . . .[7]

In other words, only a small number of journals will be needed to supply the nucleus of papers on a given topic, assuming that the topic is a narrow scientific subject. Beyond the nucleus or first zone, however, the number of journals required to produce the same number of papers increases dramatically. For example, if two journals supply 300 articles on a topic, then four additional journals will be needed to supply the next 300 articles, and sixteen journals the next 300 articles.

When Bradford applied his formulation to bibliographies on lubrication and geophysics, he found that three zones of productivity resulted. Unfortunately, Bradford did not conclude his study by simply stating his law verbally, but instead went on to express it graphically using experimental data, not noting himself that the graphical expression was not mathematically identical to the verbal formulation. He plotted $R(n)$ (cumulative total of relevant papers) against log n (natural logarithm of the total of productive journals) and found that the data revealed an elongated S-shaped curve, the general form of which is shown in figure 1. Part one of the curve, the initial concave portion, represents the higher density of the nuclear zone. Part two, the linear portion of the curve when data are plotted on a semilog scale, is equivalent to the Zipf distribution,[8] hence the commonly used expression the *Bradford-Zipf distribution*. Part three, often called the *Groos droop*, shows the departure from linearity for higher values of n, the reason for which is not yet fully understood.[9] Brookes thought that the droop was observed when there were omissions from the relevant literature.[10] However, Praunlich and Kroll thought it was an intrinsic factor of the distribution.[11]

In the years following the publication of Bradford's law, papers by eminent researchers such as Vickery, Brookes, and Leimkuhler contributed to a partial understanding of the Bradford distribution—partially because these contributors did not interpret the law in mathematically identical terms. Vickery extended the verbal formulation to show that it applied to any number of zones of equal yield, not to only the three zones that Bradford had used for his data.[12] Later Leimkuhler expressed the verbal formulation mathematically as is shown in equation 1.[13]

$$R(n) = j \log (n/t + 1) \; (n > n_m) \qquad [1]$$

where

$R(n)$ = cumulative total of relevant papers found in the first n journals when all periodicals are ranked 1, 2, 3 . . . n in order of decreasing productivity;

n = cumulative number of journals producing $R(n)$ relevant papers;

j and t = constants defined in terms of other variables; see equations (4) and (6); and

n_m = the value of n beyond which the curve becomes linear.

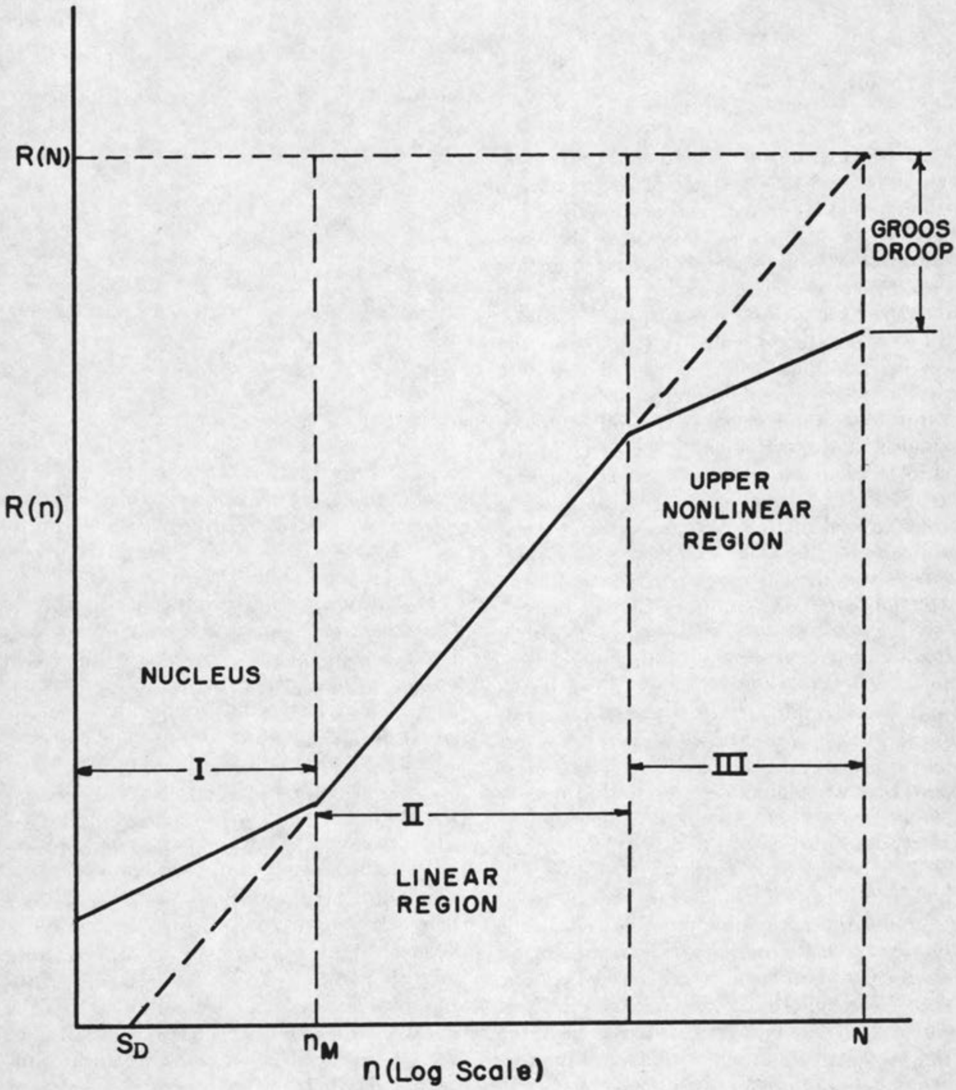

Fig. 1
General Form of Bradford Distribution

Still later, Brookes expressed the formula for the graphical version of Bradford's law beyond the nuclear zone and for N large as is shown in equation 2.[14]

$$R(n) = N \log (n/s)\ (n > n_m) \qquad [2]$$

where

N = total number of journals estimated to contain articles relevant to the subject of the search; and

s = a constant calculated using experimental data.

For some time, it seemed that only Vickery had noted that the verbal and graphical formulations were not mathematically identical. Once the disparity between the two formulations was recognized, the question arose concerning which of the two was more practical to apply to empirical data. Wilkinson devised a comparative test between the two formulations utilizing the same bibliographic data for four different subjects (agricultural economics, muscle fiber, schistosomiasis, and mast cells).[15] The test did not require calculation of the nucleus (region I in figure 1). Instead it utilized simple formulas for calculating N (the estimated total number of journals containing articles relevant to the subject of the search) and $R(N)$ (the estimated total number of papers produced by N). Only p (number of journals) and S (the corresponding cumulative number of papers) had to be known in order to apply the formulas. Both p and S were obtained from a plot of the empirical data on semilog paper. Although the value of p could be chosen anywhere in the linear portion of the curve, the point at which the initial concave portion of the curve turned into the linear region ($n = n_m$) was arbitrarily chosen to be equal to p and was used in determining the corresponding value of S. By identifying on the plot $2S$ papers, the corresponding number of journals required to supply $2S$, called q, was ascertained. The values obtained for S, p, and q were then used to calculate N and $R(N)$ for both the verbal and graphical expressions of Bradford's law (equations 3–6*). Wilkinson's test revealed that, for the data she considered, the graphical rather than the verbal formulation was more consistent with the practical situation.

*Note that the discrepancy in equations 4 and 6, when they are compared with those given by Wilkinson, was attributed to a printer's error in the original article.

Verbal Formulation

$$N = \frac{S}{\log \alpha} - \frac{p}{\alpha - 1} \qquad [3]$$

where $\alpha = \dfrac{q - p}{p}$

$$R(N) = \frac{S}{\log \alpha} \cdot \log \left[\frac{S}{\log \alpha} \cdot \frac{(\alpha - 1)}{p} \right] \qquad [4]$$

Graphical Formulation

$$N = \frac{S}{\log \beta} \qquad [5]$$

where $\beta = \dfrac{q}{p}$

$$R(N) = \frac{S}{\log \beta} \cdot \log \left[\frac{S}{\log \beta} \cdot \frac{\beta}{p} \right] \qquad [6]$$

Purpose

The purpose of this investigation is threefold. First, the study serves as a test of the two formulations of Bradford's law, verbal and graphical, to determine which better fits citation data from two selected journals in library science. Wilkinson's formulas are used, thus permitting comparisons to be made, not only between the two library periodicals, but also with the results of her study. Two basic differences exist between the Wilkinson study and the current one. The first is that citations are used rather than a bibliography. This means that the physical significance of N as representing the exhaustive search of a complete bibliography does not apply here since there is no restriction on the cumulative sum of citations that can be made. The second difference is that the citations were drawn from two library science periodicals, *C&RL* and *SL*, rather than from journals representing a narrow scientific subject. Thus, another purpose of this study is to test the appropriateness of applying Bradford's law to works in broad subject fields such as library science. The final purpose of this investigation is to briefly identify any changes and trends in adherence to the Bradford distribution that may have occurred in library literature during the thirty-five-year period 1940 through 1974.

METHODS

The data base for this study consists of 5,628 citations to the journal literature referenced in the articles published in *C&RL* and *SL* from 1940 through 1974.[16] The data were divided into seven five-year totals. *C&RL* began publication in December 1939; this single issue from 1939 is included in the study. The bibliographic information recorded for each citation included journal title and date. Journal titles were sorted in descending order of productivity. Graphs were plotted on semilogarithmic paper, with the vertical axis representing the cumulative number of citations and the horizontal axis being the natural logarithm of the number of journals producing these citations. On the resulting graphs, the linear region was extended as a dashed line for use in calculating the deviation from the Bradford distribution. Great care was taken in ascertaining the slope of the extended portion of the linear region, because it was found that even slight inaccuracies strongly affected the percentage-error calculations. Figures 2 and 3 represent the extreme cases encountered in this study, both of which by coincidence involved citation data from *C&RL*. Note that the dashed line in figure 2 deviates dramatically from the solid line, illustrating that the calculated number of papers was far greater than the observed number of papers. The percentage error here is 144.0 percent, the largest of the study. On the other hand, the dashed line in figure 3 deviates very little from the solid line and indicates only a small difference between the calculated and observed num-

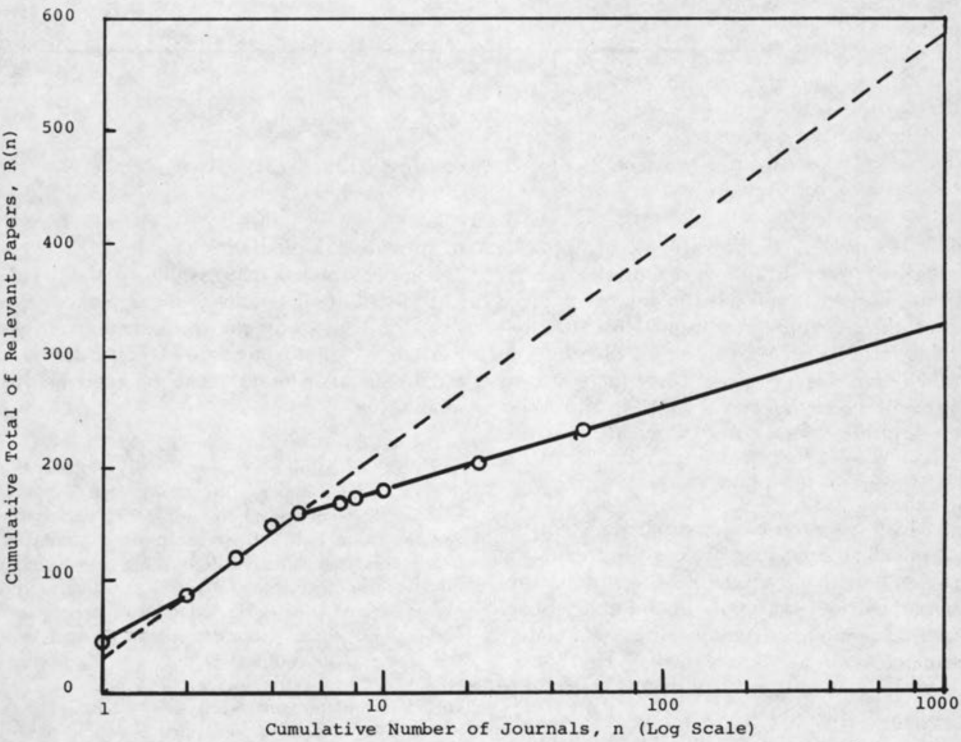

Fig. 2
Bradford Distribution of *C&RL* Cumulative Citation Data, 1940–44

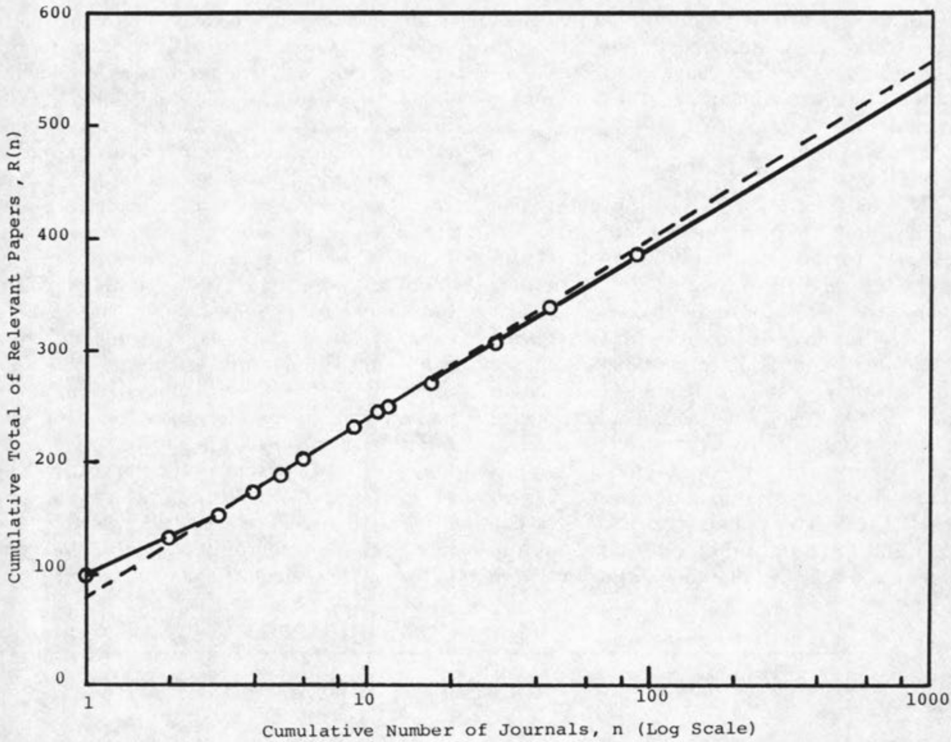

Fig. 3
Bradford Distribution of *C&RL* Cumulative Citation Data, 1955–59

ber of papers. This graph illustrates the smallest percentage error of the study, −0.7 percent. Note also in figures 2 and 3 that these curves represent Zipf distributions rather than true Bradford distributions since there is essentially no nuclear zone present in either case. As East and Weyman pointed out, this is because citation data are used.[17]

RESULTS

Recall from the earlier discussion that the physical significance of N as representing the exhaustive search of a complete bibliography does not apply in the case of citations, since no restriction on the cumulative sum of citations can be made. This study, therefore, deals only with the estimated values of $R(N)$. However, the value of N still represents a point on the extension of the linear portion of the curve and as such can be calculated using Wilkinson's formulas to provide values of $R(N)$ for both the verbal and graphical expressions of the Bradford distribution. Tables 1 and 2 present the data for this study for the seven five-year periods, 1940 through 1974. The data for *C&RL* given in table 1 can be read in this manner:

> During the 1940–44 time span, two journals (p) produced a total of 86 papers (S). By plotting the data on semilog paper, it was found that 5.7 journals would be required to supply 172 papers (or $2S$). For that entire five-year period, a total of 234 papers was actually cited. When the formula for the graphical expression was applied to the observed data, the estimated total of papers was found to be 391.1, a sum considerably larger than the observed total of 234. Therefore, the percentage error was rather large, 67.1 percent. The verbal formulation provided an even larger estimated total of papers, 571, and, of course, a still greater percentage error, 144.0 percent.

In the tables note that the percentage errors

TABLE 1

COMPARISON OF RESULTS FOR TOTAL NUMBER OF PAPERS R(N) USING DATA FROM *C&RL*

Time Period	p	q	S	Observed Total	Graphical Formulation Estimated Total	Graphical Formulation Percentage Error	Verbal Formulation Estimated Total	Verbal Formulation Percentage Error
1940–44	2	5.7	86	234	391.1	67.1	571.0	144.0
1945–49	2	6.7	156	435	693.7	59.5	879.0	102.1
1950–54	1	3.0	105	391	540.8	38.3	760.5	94.5
1955–59	3	25.0	153	385	382.5	−0.7	390.7	1.5
1960–64	5	35.0	203	496	519.9	4.8	535.9	8.0
1965–69	3	12.3	297	993	1191.8	20.0	1368.6	37.8
1970–74	3	18.0	258	797	815.4	2.3	860.0	7.9

p = particular value of the cumulative number of journals, n, arbitrarily selected = n_m.
S = estimated cumulative number of papers, $R(p)$, corresponding to $n = p$.
q = cumulative number of journals, n, required to supply $2S$ papers.

$$\text{Percentage error (for both graphical and verbal)} = \frac{\text{Estimated} - \text{Observed}}{\text{Observed}} \times 100 \text{ (at corresponding } N\text{)}.$$

TABLE 2

COMPARISON OF RESULTS FOR TOTAL NUMBER OF PAPERS R(N) USING DATA FROM *SL*

Time Period	p	q	S	Observed Total	Graphical Formulation Estimated Total	Graphical Formulation Percentage Error	Verbal Formulation Estimated Total	Verbal Formulation Percentage Error
1940–44	1	2.9	40	175	176.2	0.7	251.0	43.4
1945–49	4	54.0	80	159	142.7	−10.3	142.9	−10.1
1950–54	4	29.0	50	119	96.5	−18.9	97.6	−18.0
1955–59	6	46.0	97	215	195.7	−9.0	198.2	−7.8
1960–64	3	16.4	82	251	216.1	−13.9	227.3	−9.5
1965–69	4	30.3	158	369	389.8	5.6	399.5	8.3
1970–74	3	14.0	170	609	567.9	−6.8	622.3	2.2

p = particular value of the cumulative number of journals, n, arbitrarily selected = n_m.
S = estimated cumulative number of papers, $R(p)$, corresponding to $n = p$.
q = cumulative number of journals, n, required to supply $2S$ papers.

$$\text{Percentage error (for both graphical and verbal)} = \frac{\text{Estimated} - \text{Observed}}{\text{Observed}} \times 100 \text{ (at corresponding } N\text{)}.$$

for each set of citation data from *C&RL* and *SL* generally decrease in the latter years of the study, and rather dramatically for the *C&RL* data.

Whereas Wilkinson's results plainly indicated that, for her data, the graphical formulation more closely adhered to the practical situation, the present investigation reveals no such clear-cut picture. In all seven cases for the *C&RL* data, the graphical formulation does provide smaller percentage errors for estimated values of $R(N)$. On the other hand, the verbal formulation provides smaller errors in five of seven cases for the *SL* data. For these sets of citation data, then, neither formulation appears to be distinctly superior to the other.

Wilkinson's study utilized complete bibliographies for four narrow scientific fields, while the current investigation uses citation data from two widely circulating library science periodicals. The question thus arises concerning which group of data more closely adheres to the Bradford distribution. Since, at the present time, there is no agreed-upon percentage of error that determines adherence to Bradford's law, one can state only that the smaller percentage errors indicate closer adherence. For Wilkinson's data, the errors ranged from 1 percent to slightly more than 5 percent in six of eight cases for both graphical and verbal formulations. The errors were much higher for the *C&RL* and *SL* data, with fewer than 50 percent of the cases calculated as about 10 percent or less. Only five of twenty-eight cases had errors of less than 5 percent. This gloomy picture is lightened, however, by a closer inspection of the percentage errors given in tables 1 and 2, which reveals an important fact—that the higher errors generally occurred in the early years of the study, 1940 through 1954. This is particularly evident in the *C&RL* data. After 1955, in six of eight cases, the *C&RL* data had errors of less than 10 percent. This obvious trend toward smaller errors in the latter years of the study is not present in the *SL* data. Instead, smaller percentage errors exist for both graphical and verbal formulations of *SL* data, with ten of fourteen cases having errors of about 10 percent or less. Thus, while the percentage errors for both formulations of Bradford's law and for both library journals overall were not as small as Wilkinson's, the majority are in the range of acceptability for the latter years of the study.

Conclusions

The purpose of this study was threefold: (1) to ascertain whether the graphical or verbal formulation of Bradford's law more closely adhered to the practical situation when applied to citation data; (2) to test the appropriateness of applying Bradford's law to works in broad subject fields such as library science rather than to data from narrow scientific fields; and (3) to identify any trends in adherence to the Bradford distribution that occurred in popular library literature during the thirty-five-year span, 1940 through 1974. The results were mixed to such an extent that few firm conclusions can be reached.

For the data analyzed, neither verbal nor graphical formulation of Bradford's law provided strong enough evidence to indicate its superiority for use with citations. While the graphical formulation provided smaller percentage errors for the citations from *C&RL* in all seven cases, the verbal formulation did so for the *SL* data in five of seven cases. Thus, the obvious conclusion is that further testing of citation data is needed, with the stipulation that the same methodology and formulas be used.

The appropriateness of Bradford's law as a test of data from a broad subject field like library science is another question addressed in this investigation. The evidence supports a positive response. In almost all cases for the *SL* data, the percentage errors were not so high that either the data or the test could be called invalid. The reasons for the consistently closer adherence of *SL* citations to Bradford's law are not known. It may have been that *SL* had a narrower subject coverage than did *C&RL*. On the other hand, the very large errors observed in the *C&RL* data for the first three time spans, 1940 through 1954, indicate that either the test is inappropriate or the data are invalid. The latter reason seems to be correct since much smaller and more acceptable errors are observed after 1954. This fact may indicate that library literature as reflected in *C&RL* and *SL* changed or, at the very least, began to change into something more substantive and more worthy than "garbage." Regardless of the reasons for the behavior of the citation data from *C&RL* and *SL*, however, the overall evidence indicates that Bradford's law can be applied with a relatively high degree of confidence to data from sources not generally considered to be "narrow" or "scientific"—sources such as library science periodicals.

References

1. William Katz, "Publications," *Drexel Library Quarterly* 3:176 (1967).
2. Eric Moon, "The Library Press," *Library Journal* 94:4104–9 (1969).
3. S. C. Bradford, *Documentation* (Washington, D.C.: Public Affairs Press, 1950).
4. See, for example, W. Brace, "A Citation Analysis of Doctoral Dissertations in Library and Information Science, 1961–1970" (Ph.D. diss., Case Western Reserve Univ., 1975).
5. S. Bulick, "Book Use as a Bradford-Zipf Phenomenon," *College & Research Libraries* 39:215–19 (1978).
6. E. A. Wilkinson, "The Ambiguity of Bradford's Law," *Journal of Documentation* 28:122–30 (1972).
7. Bradford, *Documentation*, p.116.
8. G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Reading, Mass.: Addison-Wesley, 1949).
9. O. V. Groos, "Bradford's Law and the Keenan-Atherton Data," *American Documentation* 18:46 (1967).
10. B. C. Brookes, "The Derivation and Application of the Bradford-Zipf Distribution," *Journal of Documentation* 24:247–65 (1968).

11. P. Praunlich and M. Kroll, "Bradford's Distribution: A New Formulation," *Journal of the American Society for Information Science* 29:51–55 (1978).
12. B. C. Vickery, "Bradford's Law of Scattering," *Journal of Documentation* 4:198–203 (1948).
13. F. F. Leimkuhler, "The Bradford Distribution," *Journal of Documentation* 23:197–207 (1967).
14. Brookes, "Derivation and Application."
15. Wilkinson, "The Ambiguity of Bradford's Law," p.125–26.
16. G. S. Cline, "A Bibliometric Study of Two Selected Journals in Library Science, 1940–1974" (D.L.S. diss., Univ. of Southern California, 1978).
17. H. East and A. Weyman, "A Study in the Source Literature of Plasma Physics," *Aslib Proceedings* 21:160–71 (1969).