New Organic Compounds in *Chemical Abstracts* and *Current Abstracts of Chemistry and Index Chemicus*: A Comparison

ARTHUR ANTONY* and JANET STEVENS

Sciences-Engineering Library, University of California, Santa Barbara, California 93106

Received August 3, 1979

On the basis of a sample of 200 compounds indexed as new in the 1974 volumes of Current Abstracts of Chemistry and Index Chemicus (CACIC), the indexing of those compounds in CACIC and Chemical Abstracts (CA) is compared. The results of the research reaffirm the value of both CA and CACIC for the library supporting advanced research in organic chemistry. In spite of the existence of these two high-quality retrieval tools, in a few instances authors had failed to reference previous reports of a compound. These cases are discussed in detail.

INTRODUCTION

Two of the most important English language literature indexes which focus on chemistry are Chemical Abstracts (CA), published by Chemical Abstracts Service in Columbus, Ohio, and Current Abstracts of Chemistry and Index Chemicus (CACIC) published by the Institute for Scientific Information in Philadelphia. CACIC can be considered a retrieval tool for organic chemistry, whereas CA comprehensively covers all areas of chemistry and chemical engineering. CACIC is much more restricted than CA in the extent of its literature coverage in that the CACIC indexers concentrate upon a core of approximately 100 research journals, while the CA indexers scan over 14000 journals in chemistry as well as other scientific fields for relevant source articles. CA covers patents, books, and other document types not included in CACIC coverage.

Both publications are noted for their depth of indexing for specific chemical substances. CACIC displays structures for virtually all of the chemical compounds reported in an article in the CACIC surrogate of that article, and includes an entry in the formula index for each new organic compound reported. The entries in the CACIC subject index are often for compound classes rather than specific compounds. Relatively few structures are drawn in the CA abstracts, and specific compounds often go unmentioned there; CA includes entries for specific substances in both its Chemical Substance and Formula Indexes, but does not distinguish new compounds from others. The depth of substances indexing in CA and CACIC is a most valuable asset to a chemist who is trying to find everything on a particular substance, or is anxious to note if any report at all has been made on a substance. Ross¹ has pointed out that this depth of indexing can, on occasion, be annoying, especially when a computer search dumps a large number of citations, many of which are based on a rather minor role of the compound being searched.

There have been very few reports in the literature comparing CA and CACIC, but there have been several reported evaluations of CA, some of them based on the computerized versions.²⁻⁷ Some of the features of CACIC and CA, such as the arrangement and the types of indexes, were compared by Luthra in 1971.⁸ Luthra noted that CACIC compares favorably with CA with regard to having a shorter time lag between publication of the original article and coverage by the secondary service. Kiehlmann⁹ also summarized some important features of CA, CACIC, and other major chemical indexing and abstracting publications.

Leggate, Rossiter, and Rowland¹⁰ remarked that CACIC exhibits a much lower recall than CA, reflecting the limited coverage of CACIC. Although some of the profiles upon which their research was based were clearly outside the scope of CACIC coverage, 70% of a sample of articles that had been

Table I

 	A	В	Δ, %	
English	65.00	70.76	-8.14	
Russian	13.50	10.16	+32.87	
German	10.00	9.69	+ 3.20	
French	8.50	6.48	+31.17	
Japanese	1.50	1.82	-17.58	
Italian	0.50	0.47	+ 6.38	
Others ^a	1.00	0.62	+61.29	

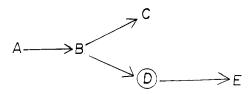
^a There were two "others" in our sample. Both were in Polish.

judged relevant but were not retrieved from CACIC were in the field of organic chemistry, and nine of the articles (15% of the sample) reported new organic compounds or syntheses. These nine articles were in journals which were not scanned by CACIC staff. The study indicated that 61.5% of the relevant articles retrieved (using both CA and CACIC) were not covered by CACIC. It is important to note that many of these were articles outside the parameters for CACIC coverage. However, the results suggest that about 15% of that 61.5%, or 9.2% of all the articles judged relevant, did report new organic compounds or syntheses, but were not covered by CACIC. This is reasonably close to the report by Garfield, Revesz, and Batzig¹¹ that 100 journals cover 96% of the synthetic chemical literature.

THE SAMPLE, AND DATA GATHERED FROM IT

"New" organic compounds were selected at random from CACIC for the year 1974. For that year the CACIC citation numbers ranged from 212168 to 224854. Two hundred random digits were chosen from "A Million Random Digits",12 starting with number 18109, and proceeding in sequence using the numbers (up to the 200th) that were between 12168 and 24854. By placing a "2" before these 200 five-integer numbers, a random sample of 200 CACIC abstracts was chosen. An extra classification number (between 1 and 50) was then used to obtain the sample compound to be studied for each citation. New compounds are distinguished in CACIC from those previously reported by having underlined compound numbers. The extra classification numbers were assigned sequentially to the citation numbers (i.e., first 1, then 2, then 3, etc.), and the underlined compound numbers counted in increasing order until the appropriate extra classification number was reached. A slight variation was used when there was a large number of underlined compounds and the numbering was scattered over the page. In these cases, the sample compound was chosen by counting from the upper left hand corner across the row, just as one reads a book.

The language distribution of the articles cited by CACIC in 1974 has been recorded.¹³ The language distribution for our sample (column A) is compared with that distribution for



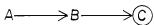


Figure 1. Examples from an unproductive attempt to classify compounds on the basis of reaction schemes.

the entire 1974 CACIC volume (column B) in Table I.

The following information was recorded for each compound selected from CACIC: structure of compound, article title, authors, language, and the CACIC issue date. Additional information, based in most cases on an analysis of the original article, was also recorded and will be discussed later. The compound "name" (which in many cases was generic) as it appeared in the subject index to CACIC was also recorded.

The author's name was then used to find the Chemical Abstracts (CA) citation, and the CA volume and abstract number, and the date of that issue of CA, were recorded. One of the 200 articles in the sample was not reported at all in CA. This particular article, in an Italian journal but in the English language, reports the preparation of a number of compounds of biological interest. The journal, Annali di Chimica, is included in the list of those monitored for CA coverage. There is thus no apparent reason that this paper should not have been covered by CA, and its absence from that publication must be judged as an oversight. Chemical Abstracts Service, which has a procedure for correcting serious errors in its entries, and presumably also will add entries that have been inadvertently overlooked, has been informed of this omission.

One of the papers selected did not report any new compounds. In 22 other selections, the compound chosen was a labeled compound. The 23 items were excluded from further analysis in this study. The three Japanese language papers were only partly analyzed. Hence, unless otherwise noted, the rest of the discussion in this report pertains to the 173 compounds selected from the 173 remaining papers.

COMPOUNDS NOT COVERED IN CA

Each of the 173 compounds was searched via the CA Formula Index, in order to obtain the CA assigned name and registry number. For CA entries that were not readily located in this manner, a printout of the full record from Lockheed Retrieval Service's file 30¹⁷ was obtained, and each of the registry numbers that was printed was checked in the "Registry Handbook" in order to ascertain if there was a match. For 32 compounds, the CA record lacked information. This is interesting, because these are 32 compounds which CACIC reported as new compounds, and for which the publications were covered in CA, but for which CA did not provide index entries in the Formula or Chemical Substance Indexes. As will be discussed later in this paper, most of these instances can be accounted for as consistent with CA indexing policy.

In order to explore factors that might account for the exclusion of entries for compounds in the CA indexes, a characterization of all 173 compounds was made. At first, an attempt to characterize the compounds according to a reaction scheme as illustrated in Figure 1 was made. However, it soon

Table II

- A target product of a synthesis
- A' product used in a study, e.g., a model compound in a physical chemical study, or a compound prepared for the express purpose of studying a reaction mechanism
- B isolated intermediate in a synthesis
- C nonisolated but proven intermediate in a synthesis
- D postulated intermediate
- E starting material for a synthesis
- F formed in the study of a reaction, or in a mechanism study
- G natural product, isolated (including products isolated in metabolism studies)
- H derivative of a target product in a synthesis, and used to help identify the target product.
- derivative of an isolated intermediate in a synthesis, and used to help identify the intermediate
- J byproduct, not used for further study
- K mentioned from another paper in connection with a study, and not actually involved in experiments at all
- L other

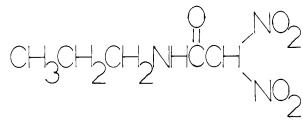


Figure 2.

became evident that this was a very unsatisfactory approach. For example, the same pattern, $A \rightarrow [B] \rightarrow C$ would occur if B were an un-isolated intermediate or the target compound of a preparation with C as a derivative of B. Two other methods of classification appeared to be more successful, although they were not completely free of ambiguity. The original papers were examined in each case in order to assign the compounds into the two classifications described in the following paragraphs.

The compounds were classified as to whether or not a chemist reading the paper could actually prepare the compound on the basis of information contained in the paper. The synthesis was considered described if the paper included an experimental procedure for the preparation of the compound, or if it directed the reader to some other published source that contained such a description. The authors made no value judgments regarding the adequacy or reasonableness of the experimental descriptions.

Of the 32 compounds *not* indexed in CA, 9 (29.0%) were classified as having the synthesis described. Of the 141 compounds that were indexed in CA, 119 (84.4%) were classified as having the synthesis described.

Table II illustrates the other classification scheme that was used. While there is a (perhaps somewhat hazy) conceptual distinction between A and A', in actual practice it occasionally was very difficult to choose between them, and so even though our statistics treat these as two different categories, the somewhat arbitrary distinction between them should be kept in mind. Though in most cases the assignment of a compound to one of the classes was rather straightforward, the authors are aware that this assignment was subjective.

One compound had to be assigned to class L (other). The compound (Figure 2) was mentioned in the paper¹⁹ as an example of a type of compound which, with one exception, had not been prepared before. However, with regard to this particular compound, the article indicates only that the potassium salt (Figure 3) was prepared. CACIC identifies the

Figure 3. Table III

class	no. of compds indexed in CA (%)	no. of compds not in- dexed in CA	total no.	x ²	
A	61 (96.8)	2	63	55.2	
\mathbf{A}'	15 (83.3)	3	18	8.0	
В	18 (90.0)	2	20	12.8	
C	1 (25.0)	3	4		
D	0 (0)	11	11	11.0	
E	6 (85.7)	1	7		
F	18 (90.0)	2	20	12.8	
G	10 (90.9)	1	11	7.4	
H	7 (87.5)	1	8		
I	3 (100)	0	3		
J	2 (40.0)	3	5		
K	0 (0)	2	2		
L	0 (0)	1	1		

compound in Figure 2 as a new compound, but CA does not include it in its index at all. The potassium salt, however, is indeed in CA.

In all cases where a sufficient number of compounds fell into the classification for a χ^2 test to be made, the difference between those that were and were not indexed by CA was significant at the 99% level.²⁰ Table III supports the use of CA for locating information about compounds that were isolated (A, A', B, F, G) whether they were actually targets for synthesis or not. On the other hand, the chemist who is searching for information on un-isolated intermediates is advised to use CACIC rather than CA. Of course, under some circumstances the chemist may want information about a compound whether it has been isolated or not, and may thus want to use both sources, for although CA has a much wider base of publications from which to index, its coverage of unisolated intermediates is poor.

These results are not surprising and, in fact, reflect the published policies of the index producers.21,22 The results merely confirm that there is a high level of agreement between the expected and actual performance of the indexes.

Prior to 1973, intermediates were indexed in CA to only a very limited extent, but coverage of substances was extended in 1973 to include intermediates for which sufficient detail is supplied.²³ Nonetheless, CACIC appears to be more thorough than CA in its coverage of intermediates.

In our results, there were two target products that CA failed to index. In one case²⁴ the authors clearly specified that they had prepared the target compound (Figure 4), and, although they did not give the details for its preparation, they did for a very similar compound (with a benzene ring in the place of the naphthalene entity) which was offered as an example of a general preparative reaction that was observed for several variant products. In the other case²⁵ the preparation of the target compound (Figure 5) was described in detail.

CURRENCY OF COVERAGE

For the 199 papers that were covered by both CA and CACIC, the time difference of the coverage was determined, based on the dates printed on the CA and CACIC issues for

Figure 4.

Figure 5.

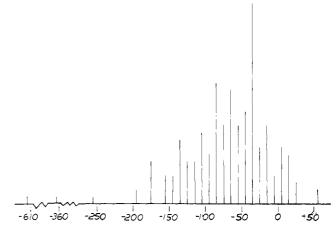


Figure 6. Time difference (in days) between CA and CACIC coverage.

each of the references. The time differences, in days to the nearest ten days, are shown in Figure 6. A negative value indicates that a reference appeared in CACIC before it appeared in CA. For four articles, coverage in CA occurred more than 200 days after coverage in CACIC: two of them from Tetrahedron (-217 days and -257 days), one from the Journal of the Chemical Society, Perkin Transactions 1 (-365 days), and one from the Chemical and Pharmaceutical Bulletin of Japan (-614 days). One hundred thirteen (more than half) of the articles were indexes in CACIC at least 50 days before their coverage in CA, whereas for two articles, one from Carbohydrate Research (+51 days) and one from Acta Chemica Scandinavica (+58 days), CA coverage occurred more than 50 days before CACIC coverage.

ARE "NEW" COMPOUNDS REALLY NEW?

CA does not distinguish "new" compounds from "old" ones in its indexing. CACIC makes a point of highlighting new organic compounds in its indexing. However, some compounds

Table IV. Characteristics of the Compounds Which Were Not New

Language: English 9, Russian 4; Polish 1
Synthesis described?: yes 11; no 3
Time difference between CA and CACIC coverage in days: +23,
+9, -5, -33, -40, -47, -47, -61, -68, -68, -75, -89, -124,
-131
Class (see Table II): A5, A'2, B1, E2, F1, G2, I1

labeled in CACIC as "new" prove to actually be "old".

Lawlor and Batzig²⁶ reported that at least 7.2% of the compounds that are indicated as new in CACIC are not really new, and that approximately half of those cases resulted from the failure of an author to cite his or her own previous report of the compound.

In this investigation 14 of the 173 compounds (8.1%) were definitely not new. Compounds were judged as not new if there were earlier reports of them located via the 9th Collective Index to Chemical Abstracts, either through a manual search of the index or an online search of Lockheed's File 3. Some of the other compounds might also not be new, since they may have been reported in the literature earlier than the time covered by the 9th Collective Index. In order to specify that a compound was definitely not new, the actual date of publication had to be ascertained, either from the date on the journal issues, or, in the case of patents, the date of grant as given in Chemical Abstracts. There were four additional compounds for which there were other publications that may or may not have been earlier; in effect, the publication date of one or another of the relevant documents in each case could not be obtained with sufficient precision to come to a definite conclusion as to which was first. These four compounds are not included in the discussion which follows.

The characteristics of the compounds which were not new are recorded in Table IV. The language distribution of the papers reporting these compounds is rather similar to that of the entire sample, although there appears to be a disproportionately large number of Russian publications. There does not appear to be any class of compound (as defined in Table II) which is strikingly overpopulated by compounds which are not new.

For two of the compounds (one in class A and the other in F) there were so many previous reports that it might be assumed that the authors considered them so well known as to need no citation. Yet in both cases the molecular structures are complex enough that a CACIC could easily be excused for not having identified them as compounds not requiring a citation. In one additional case²⁷ the authors do cite an earlier publication at an appropriate point within the text, so the compound should not have been considered as new by CACIC.

In five cases, at least one author was the same for both the earlier and later documents. These can be considered examples of an author failing to cite himself or herself. In three of these cases, the earlier publication was a patent. Maynard has remarked on the tendency of many chemists to underrate the patent literature.²⁸ Since it is reasonable to assume that authors are aware of their own previously published patents, one wonders if perhaps some scientists discount patents altogether from what they consider to be part of the "literature".

Table V displays the characteristics of the 11 documents in which the authors failed to cite earlier reports of a compound selected in our sample. In the column labeled "Earlier Authors", "d" means that the earlier and later documents had no author in common, whereas "s" means that they had at least one author in common. "Earlier document type" (the second column) indicates the nature of the document that failed to be cited. The letter "p" is used if the earlier document is a patent and "dj" or "sj" if it is a journal article. Different and same earlier journal titles are distinguished by "dj" and "sj", respectively. Zhurnal Obshchei Khimii (twice) and Zhurnal

Table V

earlier authors	earlier docu- ment type	com- pound class	earlier authors	earlier docu- ment type	com- pound class
d	dj	G	s	dj	G
d	dj	\mathbf{A}'	\$	sj	Α
đ	dj	E	S	p	Α
d	sj	Α	S	р	Α
đ	sj	\mathbf{A}'	S	p	В
d	p	I			

Organicheskoi Khimii (once) accounted for the three instances of "sj" in the second column.

The third column indicates compound class according to the categories in Table II. Those cases with a "d" in the first column and either A, A', or G in the third are deserving of further comment, because they each suggest that an author failed to give proper recognition to another author for the preparation of a target molecule or the isolation of a natural product. Because these cases came to light from a random sample, the authors of this paper do not consider it fair to expose the authors in those four cases to approbation from their peers, and so the specific papers involved will not be cited.

Case I. The authors of a 1974 paper in Arzneimittel-Forschung reported the isolation of a compound in the metabolic study of a parent drug compound. The authors failed to cite an earlier French paper, published in 1971, which focused on the therapeutic properties of that compound, and which reported its synthesis in the laboratory.

Case II. The preparation of a target compound was described in the *Zhurnal Organicheskoi Khimii*. Two years earlier, in the same journal, this compound was reported as an isolated intermediate in the preparation of something else. The two preparations used different methods, with the later paper reporting a higher yield. The first paper appeared in print a couple of months after the second one was first submitted.

Case III. An unexpected product for a particular kind of reaction was reported in *Zhurnal Obshchei Khimii*. This compound was published in the same journal three years earlier, as a product obtained in the identification of something else. It is very unlikely that the earlier report of the compound would be of any use to anyone subsequently interested in it (although the paper itself conceivably could be, for other reasons). The second paper was submitted before the earlier one had appeared in print.

Case IV. In a paper in the Journal of Organic Chemistry authors reported the identification of a small amount of a compound in the study of a rearrangement reaction. The preparation of this compound had been reported earlier in a section of the Journal of the Chemical Society. The earlier paper was published about three months before the later one was received for publication.

In none of the four cases was there a serious infringement on the priority claims of the earlier authors. In three of the cases (excepting case III) there was some information in the earlier paper that might have been of some use to the authors of the later paper, as well as the readers of that paper, and a reference to the earlier paper should have been given. For case II and case III the long delay between submission and printing of a paper might have contributed to the authors' failure to cite the earlier papers.

CONCLUSION

The results of the research reported here reaffirm the value of both CA and CACIC for the library supporting advanced research in organic chemistry. Not only do the two publications serve to back each other up, to take care of the inevitable

occasional mistake or misprint in one or the other, but each has important unique features. For example, new compounds are explicitly highlighted in CACIC, but not in CA. This study showed that some compounds covered by CACIC were not included in CA, but the techniques employed were such that the opposite (coverage in CA but not CACIC) could not be shown. However, Leggate et. al¹⁰ have already shown that.

In spite of the existence of two high quality secondary services monitoring the literature for reports of new organic compounds, chemists occasionally fail to learn about new compounds that are relevant to their research. Such failures could lead to costly duplication of research efforts, or to research that is less effective than it might otherwise have been. The subject of unintentional duplication of research has been studied.²⁹⁻³² The results of this study suggest that the problem is less serious in the field of synthetic organic chemistry than it is in other fields.

ACKNOWLEDGMENT

The authors of this paper gratefully acknowledge having received a General Research Grant from the University of California, Santa Barbara, which supported this research.

REFERENCES AND NOTES

- (1) J. C. Ross, "Searching the Chemical Literature via Three On-Line Vendors: A Comparison", J. Am. Soc. Inf. Sci., 30, 103-6 (1979). B. Charton, "Searching the Literature for Concepts", J. Chem. Inf.
- Comput. Sci., 17, 45-46 (1977).
 C. Oppenheim, "The Performance of Chemical Abstracts Subject and
- Formula Indexes in Retrieving Compounds Disclosed in Chemical Patents", Inf. Sci. (London), 9, 107-11 (1975).
- R. J. Rowlett, "Symposium on User Reactions to CAS Data and Bibliographic Services. Concluding Remarks", J. Chem. Inf. Comput. Sci., **15**, 186-9 (1975)
- J. H. Schwartz, "Some Observations Concerning Chemical Abstracts
- Formula Indexes", J. Chem. Doc., 9, 169-74 (1969). M. A. Simkins, "A Comparison of Data Bases for Retrieving References to the Literature on Drugs", Inf. Process. Manag., 13, 141-53
- M. Williams, "Analysis of Terminology in Various CAS Data Files as Access Points for Retrieval", J. Chem. Inf. Comput. Sci., 17, 16-20
- (8) K. L. Luthra, "Comparative Use of Chemical Abstracts and Index Chemicus", Herald Libr. Sci., 10, 122-30 (1971).
- E. Kiehlmann, "Organization and Efficient Manual Searching of the Major Chemical Title and Abstract Publications", J. Chem. Doc., 13, 78-82 (1973)
- P. Leggate, B. N. Rossiter, and J. F. B. Rowland, "Evaluation of an SDI Service Based on the Index Chemicus Registry System", J. Chem. Doc., 13, 192-203 (1973).

- (11) E. Garfield, G. S. Revesz, and J. H. Batzig, "The Synthetic Chemical
- Literature from 1960 to 1969", Nature (London), 242, 307-9 (1973). The RAND Copporation, "A Million Random Digits with 100,000
- Normal Deviates", The Free Press, Glenco, IL, 1955.
 "Index Chemicus. Annual Cumulative Index to Current Abstracts of Chemistry and Index Chemicus", Vol. 56-59, 1975, Institute for Sci-
- entific Information, Philadelphia, Pa., 1976, p ii.

 (14) G. Donelli, M. Di Ciommo, and C. Ravazzoni, "Synthesis of 1-Hexadecanoyl-2-acyl-sn-glyco-3-phosphorylcholines", Ann. Chim. (Rome), 63, 209-18 (1973)
- "Chemical Abstracts Service Source Index", Chemical Abstracts Service, Columbus, Ohio, 1975, Part 1, p 117.
- (16) D. H. Sears (Chemical Abstracts Service Customer Service Representative), letter to SDF customers, dated March 20, 1979, enclosure A, p 2.
- (17) At the time this search was done, File 30 corresponded to CASIA for 1973 to 1976. Since that time, Lockheed Retrieval Service has rearranged files, and File 3 would now have to be used.
- American Chemical Society, Chemical Abstracts Service, "Registry Handbook: Number Section", Columbus, Ohio, 1965/71-.
- (19) I. V. Tselinskii, and V. K. Krylov, "Anions of Dinitromethyl Compounds. 34. Synthesis of N-Substituted Dinitroacetamides", J. Org. Chem. USSR, 9, 2495-7 (1973); translated from Zh. Org. Chim., 9, 2474-7 (1973).
- (20) The χ test was based on formulas and a table in "An Introduction to Quantitative Research Methods for Librarians" by Taverekere Srikantaiah and Herbert H. Hoffman, Headway Publications, Santa Ana, Calif., 1977.
- "Chemical Abstracts", Vol. 76, Part 12, Paragraph 14.
- (22) "How to Use Current Abstracts of Chemistry and Index Chemicus" broadside issued by the Institute for Scientific Information, Philadelphia, Pa., 1975.
 (23) "Indexing of Reactants, Intermediates Expanded", CAS Report, No.
- 4, 2 (Aug 1975).
- (24) W. J. S. Lockley, V. T. Ramakrishnan, and W. Lwowski, "A Convenient Thermal Route to N,N-Dialkylaminoisocyanates", Tetrahedron Lett., 2621-24 (1974).
- (25) G. A. Howie, P. E. Manni, and J. M. Cassady, "Synthesis of Alkyl-Substituted α, β -Unsaturated γ -Lactones as Potential Antitumor Agents", J. Med. Chem., 17, 840-43 (1974).
- (26) H. A. Lawlor and J. P. Batzig, "Duplication in the Chemical Literature: II. A Systematic Study for the Years 1971-1974", Abstracts of Papers, 172nd National Meeting of the American Chemical Society, San Francisco, Aug 29-Sept 3, 1976, paper 23, Chemical Information
- (27) B. Stridsberg and S. Allenmark, "Intramolecular Transfer of Chirality from Sulfur to Carbon: Dehydrocyclization of Optically Active o-Benzylsulfinylbenzoic Acid with Dicyclohexylcarbodiimide", Acta Chem. Scand., Ser. B 28, 591-3 (1974).
 J. T. Maynard, "Understanding Chemical Patents. A Guide for the Inventor", American Chemical Society, Washington, D.C., 1978, p 7.
 J. Martyn, "Unintentional Duplication of Research", New Sci., 21, 338

- (30) D. R. Masson, "The South African Research Worker's Approach to
- (31) G. J. Brockis, and P. F. Cole, "Evaluating the Technical Information Function", Chem. Br., 3, 421-3 (1967).
 (32) I. Hirsch, W. Milwitt, and W. J. Oakes, "Increasing the Productivity of Scientism", III.
- of Scientists", Harvard Bus. Rev., 36, 66-76 (March/April, 1958).