PHYLLIS A. RICHMOND

Systems Evaluation by Comparison Testing

Most evaluation thus far of systems for "identifying and storing the information content of bibliographic materials for future search and retrieval" has been attempted by comparison with other systems. Care, however, has not often been taken to eliminate the flaws that normally may be expected to accrue in such comparisons from the operation of variable factors. It is hoped that future comparison testing of systems will not be attempted without well stated conditions and criteria and unless the systems are essentially comparable.

COMPARISON TESTING tends to be based on the "one best way" principle. This principle holds that for any given set of circumstances there is one best way of doing something. The principle is fallacious because very few methods or circumstances are so alike that they can safely be compared without danger of distortion. In most respects, distortion and false conclusions result when obvious differences between things being compared are ignored.

A case in point is the testing by comparison of different kinds of systems for the intellectual organization of information. The word "systems" here is used to denote the end products of distinct methods for identifying and storing the information content of bibliographic materials for future search and retrieval. These may be alphabetical indexing systems, formal and informal classification systems, alphabetico-classed indexes, graphically- or statistically-derived classificatory or indexing systems, assemblages of related words—hierarchically, probabilistically, structurally, or linguistically defined—and so on, all designed for different purposes and operating, for the most part, in different kinds of situations. Their common denominator is that all organize information-bearing material for search and retrieval and most have some kind of classification features, but here the likeness ends. In some cases, subject matter may be held in common, but the total system approach is sufficiently different to make the results suitable for various purposes, not necessarily identical.

The use to which a system is to be put to a considerable degree affects the choice of system, as well as its application. A universal classification, for example, covers the whole of recorded knowledge but does so in such a way that all of its parts are interdependent. If one section is selected for special treatment or expansion or realignment, the ramifications are soon felt throughout the rest of the system, which then needs the same kind of attention so that it will continue to function as an organic whole.[1] One may use a selected part of such a classification for a limited field, but experience indicates that no field stays limited for long and that eventual-

Dr. Richmond is Supervisor of River Campus Science Libraries, University of Rochester.

[1] The author is indebted to Mrs. Marie Henshaw, Decimal Classification office, Library of Congress, for clarification of this point.

ly one must make decisions vis-à-vis the whole. The special vocabulary of a coordinate indexing system, which has undefined or minimally defined terms, is dependent for successful operation upon its use in a limited, homogeneous subject field.[2] Inside this field, words may have to be combined in a special order or with indicated relations to be consistently meaningful. Outside the field, the homograph factor demands exact definition of terms.[3] The simplicity of the system, which is the basis for its effectiveness[4] in the narrow field, then disappears, because with exact definition it necessarily follows that one must have exact rules for application and very exact designation of which kind of relationships may be made between terms.

Let us consider, in some detail, what happens when distinction is not made among the purposes and operations of different kinds of systems and they are treated as equals. The first ASLIB-Cranfield Research Project is a case in point.[5] In this experiment, distinction was not made between those systems designed with a universal approach to the intellectual organization of information and those designed for limited use in parts of the whole. The former, when one comes to a specialized subject like aeronautics, is a dilute approach, while the latter is a concentrated one. At Cranfield, the dilute approach was made through the Universal Decimal Classification, and through alphabetical subject headings, which are generalized-concept index terms. The concentrated one was made through a faceted classification tailor-made for the subject, and through Uniterms, which had a vocabulary composed of words taken directly from documents dealing with the subject. All four were applied, as if each were equally qualified, in a situation that called for the concentrated type of approach. One would expect, in such a situation, that the concentrated approach would yield much better results in terms of recall than the dilute approach. Swanson's interpretation of the results suggests this is what happened.[6]

The opinion of the director of the Cranfield Project, Cyril Cleverdon, that the dilute approach is almost as effective as the concentrated one in answering a group of questions based on the source documents is open to doubt because of the rather irregular nature of the statistical reporting, the type of questions asked, the method by which they were compiled, the preponderance of title word indexing, the switching of indexers from system to system, and the type of subject analysis used.[7]

In the Cranfield comparison of sys-

[2] The impressive logic of Donald Hillman suggests that even this conclusion is too favorable to coordinate indexing based on Boolian algebra; *cf.* "Two Models for Retrieval System Design," *American Documentation,* XV (July 1964), 217-25; Alan Rees, "Why Are Information Centers Successful?" *Proceedings of the American Documentation Institute,* vol. 1, *Parameters of Information Science* (Philadelphia; 1964), p. 175; Arthur D. Little, Inc., *Centralization and Documentation: Final Report to the National Science Foundation,* Report C-64469 (Cambridge, Mass.: Arthur D. Little, Inc., July 1963), *passim.*

[3] This point *for all descriptor usage* has been emphasized by Calvin N. Mooers, "The Indexing Language of an Information Retrieval System," *Information Retrieval Today: Papers Presented at the Institute Conducted by the Library School and the Center for Continuation Study, University of Minnesota, September 19-22, 1962.* Ed. by Wesley Simonton (Minneapolis: Center for Continuation Study, University of Minnesota, 1963), p. 34. Mooers' original term "descriptor" had a limited meaning, but it has become generic for "index term" and the meaning broadened considerably in the process.

[4] The terms *effective* and *efficient* are treated as synonyms in this paper *(cf. Webster's Third New International Dictionary . . . Unabridged).* This is in contrast to their definition as separate terms by John A. Swets, "Information Retrieval Systems," *Science,* CXLI (July 19, 1963), 245.

[5] Cyril Cleverdon, *Report on the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems* (Cranfield, Eng.: The College of Aeronautics, September 1960); *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems* (Cranfield, Eng.: 1962).

[6] Don R. Swanson, "The Evidence Underlying the Cranfield Results," *Library Quarterly,* XXXV (January 1965), 13.

[7] Phyllis A. Richmond, "Review of the Cranfield Project," *American Documentation,* XIV (October 1963), 307-11; John R. Sharp, review of Jean Aitchison's and Cyril Cleverdon's *A Report on a Test of the Index of Metallurgical Literature of Western Reserve University, et al.,* in *Journal of Documentation,* XX (September 1964), 170-74; Swanson, *op. cit.,* 1-20.

tems, the means for making the comparison was highly significant. The experiment was expected to reveal differences between the four systems for the intellectual organization of knowledge and to show which was the most effective in dealing with a corpus of aeronautical material. Since the four systems were treated as if each were equally applicable to the given test situation, each document was analyzed for all four systems in one operation within a definite time limit. Regardless of which system was used for the *initial* analysis (the four were rotated), its result was then matched to the terminological or structural pattern of the other three. No attempt was made to break the train of thought occasioned by the first analysis. For example, if this were Uniterms, the analyst, unconsciously if not consciously, must have used these terms in searching the indexes of the other three systems. Each time he made four analyses from the brain work on one system, *each time translating the initial analysis into the language of the other three*. The test was, in effect, a consecutive three-part conversion of each system for one-quarter of the documents, rather than a test of each system on each document.

To make the point even clearer, suppose the conversion of the first analysis of a document into the other three systems had been done automatically by computer. Theoretically, the differences between system terminology could have been minimized and the human error factor virtually eliminated. Such a course would have forced prior decisions on compatibility of terms and class descriptions which in themselves would have shown up major differences between the systems. Such a machine conversion process would not have eliminated an error in analysis in the first place; it would merely have transferred it to all systems. This may have happened anyway in the Cranfield Project. Was an error in the initial analysis corrected in the other three systems, or was an error an error throughout? The time factor suggests the latter, since there is no evidence that the individual document was re-analyzed for each successive system after the initial mental work was done.

If each indexer-classifier had worked with one system alone, presumably the differences between systems would have been maximal since the source documents would have been analyzed each time by someone thoroughly familiar with the system and using its viewpoint, without reference to the other three. The use of a single analyst for each system might have shown better how each system operated in the given test situation. This brings out another point. There was no test to show how well each system operated within the framework for which it was designed. This, perhaps, was most apparent in the case of the faceted classification, which came off rather badly because the stop-watch nature of the experimental environment put speed of access to a system at a premium, and this kind of classification is not designed for quick reference.

This detailed discourse indicates some of the difficulties involved in treating all systems for the intellectual organization of information as equals in a given subject area. Such a course has its origin in confusion over the nature of varied methodologies used in the different systems tested. One means of looking objectively at methodology in a field is to stand off and view the field with the eye of a stranger. This is easier to do if one actually is a stranger. Another way is to choose to look at a similar field as an analogous situation. This path is dangerous in that no two fields are exactly alike and extrapolation from one to another should be done sparingly, if at all. Analogy, however, is the basis for making mathematical, mechanical, or other models in scientific research and has proved quite helpful in providing explanations in relatively intractable situations.[8] The analogous procedure to be used here is not a formal model, but merely citation of similar instances in the second field as parallel illustration of the organization of the first.

The field of intellectual organization of information is composed of a complex of systems. It may be likened to the transportation field, which is also a complex of systems. In one instance recorded knowledge is transferred from head to head. In the other, heads (with bodies) and goods are transported from place to place. Both kinds of complexes may be evaluated in terms of multiple factors studied with an unlimited degree of refinement: speed, safety, convenience, reliability, comfort, ease of use, ability to transport directly from point to point, accessibility, time limits, switching, interchangeability, flexibility, modernity, and so forth. The ideal system in either complex would take anyone anywhere without switching, with speed, safety, reliability and comfort, on schedule, and by any route the user wished to take. Needless to say, there is no single transportation system for all these purposes, and it seems most unlikely that there will be a single information system for all purposes.

This is an easy point to make with transportation systems because individual preferences and needs are taken for granted. Aunt Maud would not be caught dead on an aeroplane, while Cousin John will go thousands of miles to ride on a train drawn by a steam engine. One may, however, become a little more subtle than this. It is possible to get to Los Angeles by train, aeroplane, automobile, ship, roller skates, and by many other means. The Santa Fe Railroad and Pan American Airways are equally effective ways of getting there, but the routes, scenery, travel time, equipment, and such are considerably different. One cannot take Pan American to Chicago or the Santa Fe to Honolulu because they do not go there, though both go to Los Angeles. It seems almost silly to mention such obvious differences in transportation systems, but similar differences in systems designed for the intellectual organization of information go unnoticed.

Suppose the Cranfield experiment had been made with transportation systems. One might have had four, as follows: the Canadian Pacific Railway, a transcontinental system corresponding to the Universal Decimal Classification; the New York Central Railroad, a regional system corresponding to a faceted classification; Air France, a worldwide system corresponding to alphabetical subject headings; and Ozark Air Lines, a regional system corresponding to Uniterms. All four transport passengers and freight. All four have similarities in operating detail, such as ticketing, rate structure, terminals, baggage handling, guarantees, schedules, etc. There are also a few major differences between them. The New York Central takes to the air once in a while, and Air France has been known to come to ground unexpectedly, but in general they operate in different milieus. All four systems operate successfully in their respective areas, but to expect one to do what another does and to compare them in effectiveness in transporting passengers depends entirely on the wishes of the passengers, not on the system. All are equally effective in reaching their destinations; it just depends on how those destinations are sought.

The important factor in comparing systems is *not* to show whether one is better than another (more efficient in a specific set of circumstances) because *better* is a qualitative term, always a subjective judgment and relative to an

[8] Mary Hesse, "The Role of Models in Scientific Theory," in Dudley Shapere, ed., *Philosophical Problems of Natural Science* (New York: Macmillan, 1965), pp. 102-109; Toulmin discusses a model as a metaphor which suggests further questions and can be systematically deployed, noting that "a model can only be used to explain the behaviour of things which are distinct from it." Stephen Toulmin, *The Philosophy of Science: an Introduction* (New York: Harper and Row, 1960), pp. 38-39, 165.

intuitively-derived set of ideals or frame of reference. The importance factor is to show whether each system does what it was designed to do, where it fails, and where it could be improved. Note that the comparison of unlike systems, in particular, should be made on the basis of clearly stated criteria and should lead to possible adoption or adaptation of desirable features from one to another, rather that to a choice between them, with implied condemnation of the "losing" system.[9]

Comparison of a generalized system with a highly specialized one is like comparing the New York Central with the cog railway that runs up Mt. Washington. Unless the criteria for comparison are very clearly given, the two have little in common to make comparison valid. "Effectiveness" without saying effectiveness in precisely what way is not enough. The uneven comparison is easy to see with this example. It is not so easy to see when comparing one part of a universal classification with the index to a highly specialized publication in the same subject area. Though there are similarities in modern classification and indexing, the concept-structure approach is still very different from the word approach, even when the relationships between words are classified. If the comparison is done to test the hierarchical chain procedure needed to reach a given concept against an alphabetical listing to get the same thing, there should be consideration of the time and entry factors involved. A system, such as an index, which can be entered directly by means of the *correct* word or words, would be much faster to use than a classification schedule, where one works down to the correct description through a series of levels. But the term in the classification will be unambiguous by the time one has gone though its family tree, while the index term, if it is even present in the sought-after form, may turn out to be a homograph, homonym, or synonym of the concept desired. If speed is a factor, one may be tempted to use the index to the classification without checking the schedules, which alters the experiment to testing one index against another.[10]

In addition, comparison of a large system with a small one or the intellectual organization of information in a big subject with that in a little subject, on different scales, can be invidious. A universal classification, for instance, was never intended to do what a highly specialized subject classification or index does, and a big subject cannot be treated with the detail a small one commands except on a scale that so far has not been produced. Therefore, the first question to be asked before comparing any systems is "Are they comparable?" This leads to the second question, presuming the first is answered in the affirmative, "What was each system designed to do?" Testing the efficiency of one in relation to another should take the answer to this second question into account. There seems no sense in condemning a system for not doing what it was not designed to do. No burro has yet been ridden up the Jungfrau nor train taken to the bottom of the Grand Canyon. Yet information systems designed for experts have been roundly condemned because tyros cannot use them. The ordinary library

[9] The rudiments of the idea that evaluation should only be made with well-defined criteria are suggested in Swanson, *op. cit.*, p. 8. Alvin Goldwyn has discussed the problem in terms of how to determine what is being tested: "The Place of Indexing in the Design of Information Systems Tests," *Automation and Scientific Communication; Short Papers Contributed to the Theme Sessions of the 26th Annual Meeting . . . American Documentation Institute*, pt.2 (Chicago: 1963), pp. 321-22 (also as Western Reserve University Center for Documentation and Communication *Research Report* CSL:TR-3, August 1964); Allen Kent, "The Cleverdon-WRU Experiment: Purpose," *Information Retrieval in Action* (Cleveland: Press of Western Reserve University, 1963), pp. 75-82. Methodology for all evaluation, as distinct from criteria for experimental evaluation, has been ably discussed in Irving M. Klemper, "Methodology for the Comparative Analysis of Information Storage and Retrieval Systems: a Critical Review," *American Documentation*, XV (July 1964), 210-16.

[10] An experiment along these lines was made by Gerald Jahoda, "A Technique for Determining Index Requirements," *American Documentation*, XV (April 1964), 82-85.

dictionary card catalog is a case in point, being severely criticized because it does not have simple entry for collections running into the multiple millions, or for not having highly specific subject headings where its purpose is to provide initial introduction to a field, not a detailed analysis of it.

Comparisons which ignore the design of a system can be misleading when systems are similar as well as when they are different. In the Cranfield–Western Reserve University test, a specially designed faceted classification was compared with the semantic factoring method as to effectiveness in answering questions, mostly based on documents analyzed by both systems.[11] Explanations of the original interpretation of the results of this experiment considered practically everything *except* the purpose, design, and operation of each system in its natural habitat.[12] At this writing, there is considerable doubt as to what the correct interpretation of the results should be.[13] The two systems are not tremendously different in theory—both may be considered varieties of faceted classification—so that it may be that we have two systems enough alike to make comparison criteria easy to assign. A series of new comparison tests, based on sounder experimental procedure, might yield much more fruitful results than tests between unlike systems.

Another point that has come up as a result of comparison testing of systems is the matter of relationship between the amount of material recalled in answer to a question and the relevance of that material to the question asked or its pertinence to the needs (stated or unstated) of the user.[14] Attempts to put this on an objective basis have not been impressive. Again, analogy to the transportation complex illuminates one of the key points. How does one get an irrelevant document with a relevant indexing term? How does baggage checked through for a nonstop flight from New York to San

[11] Jean Aitchison and Cyril Cleverdon, *A Report on a Test of the Index of Metallurgical Literature of Western Reserve University* (Cranfield, Eng.: College of Aeronautics, 1963).

[12] Jessica Melton and William Buscher, "The Cleverdon-Western Reserve University Experiment: Search Strategies," *Information Retrieval in Action, op. cit.*, pp. 85-91; Alan Rees, "The Cleverdon-WRU Experiment: Search Results," *ibid.*, pp. 93-99; Robert A. Fairthorne, "Implications of Test Procedures," *ibid.*, pp. 109-13; Alan Rees, "The Aslib-Cranfield Test of the Western Reserve Indexing System for Metallurgical Literature; A Review of the Final Report," *American Documentation*, XVI (April 1965), 73-75.

[13] Swanson, *op. cit.*, 17-18.

[14] Harold Borko, *Evaluating the Effectiveness of Information Retrieval Systems* (Santa Monica, Calif.; System Development Corp., August 1962), SP-909/000/00; Harold Borko, *A Research Plan for Evaluating the Effectiveness of Various Indexing Systems* (Santa Monica, Calif.: System Development Corp., July 1961), FN 5649/000/01; Harry Bornstein, "A Paradigm for a Retrieval Effectiveness Experiment," *American Documentation*, XII (October 1961), 254-59; Alan Rees, "Relevancy and Pertinency in Indexing," *American Documentation*, XIII (January 1962), 93-94; Robert S. Taylor, "The Process of Asking Questions," *American Documentation*, XIII (October 1962), 391-96; Jessica Melton, "Machine Literature Searching at Western Reserve University," *Information Retrieval Today, op. cit.*, pp. 94-97; Cyril Cleverdon and J. Mills, "The Testing of Index Language Devices," *ASLIB Proceedings*, XV (April 1963), 106-30; John L. Melton, "Pertinency of Search Results to Computer Output," *Information Retrieval in Action, op cit.*, pp. 161-69; Alan Rees, "Semantic Factors, Role Indicators *et alia*," *ASLIB Proceedings*, XV (December 1963), 358-60; Arthur D. Little, Inc., *op. cit.*, 36-45; Carlos A. Cuadra, *On the Utility of the Relevance Concept* (Santa Monica, Calif.: System Development Corp., March 1964), SP-1595; F. W. Lancaster and J. Mills, "Testing Indexes and Index Language Devices: The ASLIB-Cranfield Project," *American Documentation*, XV (January 1964), 4-13; Gerard Salton, "The Evaluation of Automatic Procedures—Selective Test Results Using the SMART System," Harvard University Computation Laboratory, *Scientific Report ISR-8*, pt. 4 (1964), 36p.; Barbara R. F. Kyle, "Information Retrieval and Subject Indexing: Cranfield and After," *Journal of Documentation*, XX (June 1964), 55-69; Walter J. Johanningsmeier and Wilfred Lancaster, *Project SHARP . . . Information Storage and Retrieval System: Evaluation of Indexing Procedures and Retrieval Effectiveness* (Washington: U.S. Dept. of the Navy, Bureau of Ships, June 1964), NAVSHIPS 250-210-3; Harold Borko, "Measuring the Reliability of Subject Classification by Man and Machine," *American Documentation*, XV (October 1964), 268-73; Robert A. Fairthorne, "Basic Parameters of Retrieval Tests," *American Documentation Institute Proceedings*, I (1964), 343-45; Gordon C. Barhydt, "A Comparison of Relevance Assessment by Three Types of Evaluator," *ibid.*, 383-85; D. W. King and P. J. Terragno, "Some Techniques for Measuring System Performance," *ibid.*, 393-98; Jocelyn Brewer, "Quality Identification in a Technical Information System," *ibid.*, 247-54; Donald Hillman, "On Concept-formation and Relevance," *ibid.*, 23-39; Barbara A. Montague, "Testing, Comparison and Evaluation of Recall, Relevance and Cost of Coordinate Indexing with Links and Roles," *ibid.*, 357-67; Mortimer Taube, "A Note on the Pseudo-Mathematics of Relevance," *American Documentation*, XVI (April 1965), 69-72.

Francisco end up in Denver? The answer lies partly in the fact that multiple factors are involved, each of which can play more than one role. For example, one small but devastating factor in indexing is simply that terms with multiple meanings may not have been adequately differentiated and defined.

Another factor in the evaluation of systems is that of assessing the results of input overload.[15] Criteria here have a bearing on the initial motivation for design and development of systems. In the transportation field, the situation is that of providing for a steady climb in demand for air transportation between cities and also for seasonal "rushes" where unusual demands are placed on all forms of transportation. In information systems, demands have tended to be evolutionary and progressive in some areas, such as the social sciences and humanities, but explosive and revolutionary in other areas, such as sciences and technology. Contests, fads, special assignments, popular concern with a subject on a temporary basis due to publicity or some event can produce an overload in any subject area. With information systems, the temporary rushes are not predictable as with seasonal rushes in transportation. The reaction of an information system to stress may be just as significant in evaluation as relevance-pertinence. Both have much to do with the user.[16]

The user of either an information system or a transportation one is the ultimate authority. If he is taken to Rome via London when he wanted to go via Paris, he is less likely to rejoice that he has reached Rome than he is to fuss about the "miserable" trip he had getting there. In fact, so much is the user the ultimate authority that he can kill a system simply by refusing to use it. Edwin Castagna's threat of the dictionary catalog facing the fate of the railroads is not an idle one.[17] Both currently demand too much of the user. Also, a point overlooked in both cases, comparison with other systems does not answer problems arising from the weakness of *this* system. In each case, the faults are internal and only obliquely subject to revelation by comparison with other systems.

Does this mean that comparison of systems should be undertaken primarily from the user's point of view? Of course not, but the user should be considered. In dealing with information systems he should be led to take a much more active part in indicating his exact destination and the route by which he wishes to reach it. He should be guided into revealing the question he really wants answered, rather than asking the question he thinks the system can answer. He should be diverted from "helpfully" posing his question the way he thinks the system will answer it.[18] With some systems it should be remembered that the user is the middleman who operates it, not the ultimate recipient of its products.

There are all kinds of users, at all levels of sophistication, and using information systems devised at many levels

[15] Richard L. Meier, "Information Input Overload: Features of Growth in Communications—Oriented Institutions," *Libri*, XIII (1963), 1-44.

[16] The literature on user needs and practices is tremendous. The *Proceedings of the International Conference on Scientific Information, Washington, 1958* (Washington: National Academy of Sciences, National Research Council, 1959), is a good starting source. Readers should also refer to the "Literature Notes" section of *American Documentation* for recent and current bibliographical sources.

[17] Edwin Castagna. Speech to first session of the Legislative Workshop, American Library Association, Midwinter Meeting, Washington, D.C., January 28, 1965.

[18] This attitude is taken in Arthur D. Little, Inc., *op. cit.*, p. 24; a refreshingly opposite tack is taken by Lauren B. Doyle: "perhaps the author has as much right to be served as the searcher, *i.e.*, in order that his articles should be retrieved by relevant readers" ("Is relevance an adequate criteria in retrieval system evaluation?" *Automation and Scientific Communication*, v.2, pp. 199-200); Doyle also expresses concern about the inability of users to state their true needs "in simple form." For user problems, see Donald J. Hillman, "The Notion of Relevance (I)," *American Documentation*, XV (January 1964), 28-29; Charles Bernier, "Correlative Indexes VI: Serendipity, Suggestiveness and Display," *American Documentation*, XI (October 1960), 277-87; Alvin J. Goldwyn, "The Semantic Code: Predetermined Relevance," *Information Retrieval in Action*, *op. cit.*, pp. 171-82; John A. Swets, *op. cit.*, pp. 245-50.

of complexity. Since the user is the final evaluator of any system, one had better be sure, in considering the user, that he and the system are matched.[19] Perhaps some of the most vocal dissatisfaction with systems for the intellectual organization of information in recent years has been due to the fact that users and systems were mismatched. Sometimes this has been true because the users, particularly in the scientific community, are impatient to get back to their laboratories and either do not have or do not take the time to learn how to use the systems available to them. Sometimes it is because the makers of the systems assume the user understands the system much better than he actually does. At other times, the problem might be resolved by removing the system from public consultation and making it available through a middleman, a solution, incidentally, that is a foregone conclusion with some automated or computer-oriented information retrieval systems.

As with transportation systems, there is no accounting for the tastes, needs, desires, limitations, adaptibility, and just plain cussedness of the individual user. Rather than trying to achieve the impossibility of being all things to all men, which is again giving at least lip service to the principle of "one best way," it is more prudent to have several systems with differing objectives, different levels of complexity or detail, as well as varying design, composition and operation, and to use them in conjunction with each other, both directly in some cases and indirectly in others—all in all a shotgun rather than a rifle approach.[20]

To compare systems for the intellectual organization of information for future retrieval without explicitly stating the criteria of the comparison is to beg the question. Variant systems should not be "run against" each other, but tested for efficiency according to well-stated conditions, for complementarity, mutual support, and for success or failure in achieving the purposes for which they were designed. It may be possible to carry over ideas and goals from one system to another, but each system is an entity and judgment as to the degree of success in retrieval should not be placed primarily on carryover factors between systems.

A good beginning towards achieving more valid comparisons has been made by Pauline Atherton in devising a table of criteria to standardize reporting of results of evaluation experiments.[21] Adopting standardized criteria for reporting will affect the *methods* of testing, since one cannot meet these standards with inadequate working procedures. A statement of exactly what features are being compared and how is certainly another "must." One may, for example, be considering the effect of precision in terminology on the speed of accurate retrieval, or structured versus unstructured vocabularies as factors in pertinence of retrieved results to stated needs of users, or ease of entry into a system for the inexperienced user, or relative applicability in a predetermined, limited situation of the results achieved by asking a certain question of two or more systems, or relative suitability of various systems to scattered or single queries as searching techniques for a given subject, or the

(Continued on page 36)

[19] *Cf.* use of G. Mannoury's term "intersubjectivity" to describe properties pertaining to this matching process in W. Goffman, J. Herhoeff, and Jack Belzer, "Use of Meta-Language in Information Retrieval Systems," *American Documentation*, XV (January 1964), 14-15.

[20] The Library of Congress has used a dual approach for decades (classification plus related subject headings). A sequential approach has been suggested by Robert A. Fairthorne, "Similarity and Stability of Textual Interests," *Information Retrieval in Action*, *op. cit.*, p. 190. Fairthorne also makes the point that "the single Master Method is as false a target as was the Philosopher's Stone." *Ibid.*, p. 197.

[21] Pauline Atherton, "A Proposed Standard Description for Reporting Evaluation Tests of Retrieval Systems," *Proceedings of the Seventh Institute on Information Storage and Retrieval, American University, February 1-4, 1965.* In press.

TABLE 2

Type of Institution	High	Median	Low
University (U)	4,703,876	607,206	17,025
Liberal arts college (LA)	1,077,422	73,937	16,701
Teachers college (TEA)	128,060	64,493	11,199
Technological school (TEC)	896,513*	45,572	1,754
Theological school (THE)	112,856	33,735	14,509
School of fine arts (FA)	39,024	8,835	2,350
Other professional school (OTH)	74,835	36,560	9,276
Junior college (JC)	125,051	11,010	1,000
Technical institute (TI)	8,003	4,353	703
Semiprofessional school (SP)	2,800	2,229	1,658

* Of the ten LC libraries reported, this one is much higher in volume count than any of the others. The second highest number of volumes is 139,186.

SYSTEMS EVALUATION . . .

(Continued from page 30)

relative efficiency of systems to expand searching in chain or array, or the degree to which systems permit self-verification, self-referral or self-correction, and so on. Possible criteria are limited only by the imagination of the experimentalist. Up to the present, very little has been done in experimental situations to alter only one variable at a time, so that much experimentation suffers from the presence of too many uncontrolled variables. The Cranfield project had so many variables going at once that one is seriously justified in asking whether the results mean anything at all. Publicly verifiable results have been remarkably rare in many recent experiments.[22] Improvement in methodology, leading to repeatability, is urgently needed in many areas of documentation.

What of the future? Currently systems evaluation by comparison testing is essentially a negative operation. Baldly comparing *what is fundamentally incomparable unless the criteria which form the basis for comparison are clearly stated* is neither objective nor valid.[23] Comparison testing does have merit and especially collateral values, provided its limitations are honestly accepted. Refinement in comparison testing technique is called for, particularly any technique which has to do with possible value judgments. It is said that all roads lead to Rome. Let us not judge them all by their approximation to an ideal Appian Way. When the roads are systems for the intellectual organization of information for storage and future retrieval, let us judge each one on its own merits, letting them complement each other, and aiming always to keep in mind the *variety* of needs of users, who are the ultimate authority in the success or failure of any system. ■■

[22] Compare Christine Montgomery and Don R. Swanson, "Machine-Like Indexing by People," *American Documentation*, XIII (October 1962), 359-66, with John O'Connor, "Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems," *American Documentation*, XV (April 1964), 96-104; and A. Resnick and T. R. Savage, "The Consistency of Human Judgments of Relevance," *American Documentation*, XV (April 1964), 93-95, with A. DeLucia, "Index-Abstract Evaluation and Design," *American Documentation*, XV (April 1964), 121-25. The DeLucia article covers work based entirely on index terms, while the Resnick-Savage work included other types. Bornstein's comments on this kind of thing are particularly apt. *Cf.* Harry Bornstein, "A Paradigm...," *op cit.*, p. 254.

[23] For an excellent discussion in some detail, see Alan Rees, "The Evaluation of Retrieval Systems," *Proceedings of the Second Annual Conference on Technical Information Center Administration, Drexel Institute of Technology, Philadelphia, June 14-17, 1965.* In press.