ENTROPY MEASURES ON THE INDUCED COURSE LOAD MATRIX

by

Merit William Hudgins

Dissertation submitted to the Graduate Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

DOCTOR OF EDUCATION

in

Higher Education Administration

APPROVED:

Burton D. Friedman, Chairman

David J. Parks

Paul W. Hamelman

Robert J. Staaf

William A. Keim

Robert A. Huff

April 14, 1976
Blacksburg, Virginia

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The author wishes to convey his sincere appreciation to the following persons:

Dr. Burton Friedman, Chairman for his continuing support, valuable direction, expertise, and advice

Dr. Robert Huff, former chairman who originally inspired the project while a faculty member at Virginia Polytechnic Institute and State University. Indeed, it was Dr. Huff who first introduced the author to the various analytical tools of the National Center for Higher Education Management Systems. It was he who suggested that the concept of complementation be investigated for possible quantification

Dr. Paul Hamelman, committee member, who patiently helped the author explore a variety of ideas and methodologies of management science

Dr. William Keim, committee member, for his support and assistance during the early phase of the project while he was a faculty member at Virginia Polytechnic Institute and State University

Dr. David Parks, committee member, for his editorial comments and other helpful advice

Dr. Robert Staaf, committee member, who has given the author a valuable economic and choice-theoretic perspective on educational decision systems

Mr. James Farmer who was kind enough to listen to the author's initial version of this dissertation and who offered encouragement for its completion

Finally, special thanks go to Mr. Jack Lewis who provided much needed interaction and exchange of ideas which helped to refine and clarify the author's original version of this work. In addition, Mr. Lewis spent considerable time and effort in developing a computer programwhich, while not a part of this dissertation, provides the capability of implementation in actual institutions.

DEDICATION

The author would like to pause at this point to thank his wife, Barbara, to whom fell the hapless task of typing the entire manuscript. This was a difficult assignment in view of the numerous mathematical symbols which had to be kept straight.

Her main contributions, however, go far beyond clerical assistance. It was her constant encouragement and willingness to sacrifice, which were crucial in supplying the necessary motivation to overcome obstacles and actually complete the work. It is to her that this dissertation is dedicated.

LIST OF TABLES

Table

Chapter I

INTRODUCTION

Resource Allocation

The last decade has produced a flurry of activity related to resource allocation and utilization in higher education. The subject is receiving greater emphasis and is everywhere in evidence as enrollments decline, competition for funds increases, and the value of higher education itself is questioned.

A current objective in higher education is to achieve maximum efficiency for a given degree of effectiveness. Most educators agree with this objective but there are practical difficulties of measurement. The difficulties become apparent when the terms efficiency and effectiveness (Balderston, 1974) are examined. Suppose that efficiency is defined as the ratio of output to input, and effectiveness is defined as the extent to which goals are achieved. To utilize the terms it is necessary to define and measure input and output as well as to state goals and devise measures of degree of achievement. Inputs present no particular problem. Students, faculty, administration, equipment, supplies, support services, and real estate can all in some sense be considered input to the educational process. The conceptualization of output and the formulation of meaningful, quantifiable goals present much more difficulty.

If infallible measures of efficiency and effectiveness existed, it would make possible highly accurate comparisons of educational activities. For example, departments within a university could be compared as to efficiency and effectiveness. Statewide, it would be possible to compare separate institutions. However, such universally accepted measures have not yet arrived on the educational scene. Perhaps they will never be developed. Nevertheless, much work is being done in this area. On a less grand scale, other measures are being developed for specific applications (Miyataki, 1975).

Entropy

In this dissertation a measure based on the entropy concept of physics is proposed. This measure will quantify the extent to which academic disciplines and degree programs complement each other.

Of all the concepts of classical physics none is so elusive as that of entropy. The term was first used by the German physicist Clausius as a measure of the change in the ability of the universe to produce work in the future as a result of past or presently occurring processes. The first application of the entropy concept occurred in thermodynamics where only the macroscopic properties of aggregated matter were of interest. The underlying atomic structure was not considered. At the same time, an effort was undertaken by Gibbs in this country and others in Europe to provide a statistical correspondence between the atomic structure of any piece of matter and its macroscopic behavior. Gibbs called the resulting theory statistical mechanics. Within this theory, entropy plays a central role.

The formal system definitions of entropy are mathematical. Let us avoid this obstacle for the moment and take a descriptive approach. Consider a perfect gas isolated within a closed container and maintained at constant temperature, pressure and volume. At the atomic level, i. e., at the microscopic level, there are many different states of motion and energy which can produce the same macroscopic conditions of pressure, temperature, and volume. A knowledge of the macroscopic conditions does not necessarily permit a description of the microscopic states. Entropy is said to be directly proportional to the information describing the present microstate of the system: thus a system of high entropy is one in which a knowledge of the macroscopic state tells us very little as to which of the possible microstates the system is in.

In a very special way, therefore, entropy measures the orderliness of a system. Spontaneous processes tend toward equilibrium, which is the state of maximum disorder consistent with the constraints of the system. This state of maximum disorder is also the state of maximum entropy.

General systems theory, as it is being developed by von Bertalanffy (1968) and others is a discipline which attempts to formulate concepts useful to the understanding of all systems, physical and non-physical. The measure of entropy has been useful in the study of physical systems. May it not be of some value in the analysis of educational systems? After all, such systems also are subject to the effects of confusion, loss of energy, inefficiency and general disorder.

Complementation

In this dissertation the entropy concept is employed in a very special context. From it, a measure called complementation is drawn. The measure will provide a new way to look at the extent to which a complementary relationship exists between academic disciplines and degree programs.

Educational measurement, evaluation and planning, of course, are still in the experimental stage. Increased efficiency and effectiveness can only come about through careful monitoring and adjustment of the educational system. Input-output relations must be discerned empirically or theoretically developed. Measurement is crucial to this process and so are values. Certainly we cannot yet say that all significant interrelationships are known or sufficiently understood. Indeed, the most crucial relationships may be those for which measures have not yet been developed.

The measure developed here will relate academic disciplines and degree programs in terms of credit hours delivered and absorbed. By a degree program in a community college, liberal arts college, technical institute, or university is meant an educational package designed for a specific purpose; it draws on some combination of resources from various homogeneous units of knowledge referred to here as disciplines. Disciplines are aggregated on the basis of similarity and custom to form departments. Further aggregation can then be made if deemed desirable for organizational and analytical purposes.

As stated above, the purpose of this dissertation is to develop a quantitative measure of the extent to which academic departments and degree programs complement each other. By means of this measure it will be possible for the first time to compare departments, programs, and institutions with respect to the concept of complementation as defined below. This will provide one more way to measure a portion of the educational process and to permit comparisons to be made. As previously stated, which measures are really crucial is not yet known. Only future research will demonstrate whether complementation will be included in the category of crucial measures.

Monopoly and Equilibrium

The precise definition of complementation will be stated shortly. First, it is convenient to establish an intuitive grasp of the idea. Consider the ways in which the credit hours of a given department can be distributed to instructional programs. Two extremes are apparent. One program may absorb all the credit hours of the department; this is a condition herein termed monopoly. The other extreme occurs when all programs share equally in the output of the department, i. e., when no program absorbs more hours than any other. This is the condition termed equilibrium. The first case represents minimum complementation. Clearly, there are infinitely many more and less monopolistic or balanced possibilities between the two extremes. Non-monopolistic distributions are called competitive. A mathematical function C, called complementation, can be defined which will satisfy the following:

1. C will equal zero when the condition of monopoly occurs

2. C will equal 1 when the condition of equilibrium occurs

3. The value of C will lie strictly between zero and one when neither monopoly nor equilibrium holds

4. C is a continuous function

5. Of any two distributions, the one closer to monopoly will have a lower value of C, and the one closer to equilibrium will have a higher value of C.

The following example will illustrate what is meant by the phrases "closer to monopoly" and "closer to equilibrium".

Example 1. Discipline A contributes 50, 5, and 5 hours, respectively, to programs 1, 2, and 3; discipline B contributes 20, 19, and 21 hours, respectively.

Intuitively and by observation, one concludes that discipline A is closer than discipline B to the condition of monopoly and that discipline B is closer to the condition of equilibrium. It is the purpose of the complementation measure C to quantify these observations.

The mathematical formulation of the complementation function C was inspired by the entropy concept as utilized in statistical mechanics and, particularly, information theory. In statistical mechanics, entropy is viewed as a measure of disorder present in a physical system. In information theory, entropy represents the degree of uncertainty inherent in an experiment.

Information

Alternatively, the entropy of information theory can be regarded as the expected value of information.[1] In this case information is defined in probabilistic terms and carries a meaning somewhat different from that of conversational usage. The basic idea is as follows. A reliable message stating the occurrence of a highly probable event conveys very little information because it merely confirms the expected. On the other hand, if an event has a very low probability of occurrence, a reliable message stating that it did in fact occur would convey a great deal of information. Thus, probability and information content are inversely related.

---

[1]Expected value is used in the sense of elementary probability theory.

Information theorists sought to express the entwined probability--information concept as an additive function. That is, the probability information content of a message stating that events A and B have both occurred was required to be the sum of the probability information content of the message stating that A has occurred and that of the message stating that B has occurred. The information function proposed by Shannon (1949) to express the probability information relations was the logarithm applied to the reciprocal of the probability. The essential mathematical details are supplied in Appendix B for the interested reader.

Relative Entropy

Representing credit hours as proportional parts of the total hours for a department, the department's contribution to each degree program is expressed as a non-negative number between zero and one. This process will be referred to as normalization. These proportions are analogous to the probabilities of information theory, and the entropy function can be applied. In this case, student degree programs correspond to events with probabilities given by their proportional parts. The entropy corresponding to a given discipline is then simply the expected value of the information content of the message stating which event has occurred. It is shown in Appendix B that the maximum possible entropy occurs when the distribution satisfies the condition of equilibrium, and the value of this maximum entropy is a function only of the number of programs. When the actual entropy of a given distribution is divided by the maximum possible entropy a number between 0 and 1 is obtained and called relative entropy. Thus the condition of equilibrium provides a relative entropy value of 1 while that of monopoly is zero.

Viewed in this light, one may observe that a discipline's total output flows in different proportions to the various degree programs. The proportional parts may be regarded as prior probabilities and the programs may be regarded as events. What is the probability that a student credit hour selected at random from all those belonging to discipline D is being applied in program P? It is just the proportional part of discipline D's output which flows to program P. The event corresponding to program P is said to have occurred if the experiment of selecting the random student credit hour produces one which was in fact applied to program P. When all the proportional parts (probabilities) are equal, we have a situation of maximum uncertainty, i. e., maximum entropy. That is, the student credit hours are equally as likely to be in one program as another. This is also the condition of maximum competition. At the other extreme, when the discipline supplies only one program there is no uncertainty at all. Every student credit hour from that discipline must belong to that program. This is the condition of monopoly or zero entropy.

Complementation Rankings

It is shown in Appendix B that the entropy of a distribution approaches zero as the distribution approaches the condition of monopoly. Hence relative entropy also approaches zero. Due to these properties of the entropy function, complementation may be defined as relative entropy, for it should now be reasonably clear that relative entropy satisfies the five conditions listed above for the complementation function C.

These relationships may be examined from either of two viewpoints: (a) the determination of the extent to which the output of a department flows to the several programs or (b) the determination of the extent to which a given program draws upon the output of the several departments.

With respect to a given cluster of programs it thus becomes possible to rank the members of a given cluster of disciplines on a scale from 0 to 1 with respect to complementation. Alternatively, programs may be ranked by complementation down through disciplines. The complementation measure is a comparison device.

For example: discipline A has a complementation of .63 while discipline B has a value of .70; the distribution for discipline A is closer to monopoly than is that for discipline B. In simple cases such a comparison can be made by inspection, as was done in Example 1. It seemed clear that the distribution 50, 5 and 5 was by any reasonable standard closer to the condition of monopoly then the distribution 20, 19 and 21. In fact, the complementation for the former is only .5153 while for the latter it is .9992.[2] As the number of programs increases it becomes increasingly difficult to make judgments of this sort without access to a numeric measure. The mathematical formulation of the complementation measure is given in Chapter III, and various competing measures are discussed in Appendix A.

Global Measures

Credit hour distributions for each academic disicipline in an educational institution are most conveniently displayed in matrix form. In such a matrix, rows conventionally refer to academic disciplines and columns refer to degree programs. Thus, the element appearing in row i and column j represents the number of

---

[2] Rounded to four decimal places

credit hours contributed by the discipline i to program j. The rows (columns) of a credit hour matrix will be referred to as discipline (program) distributions. When the complementation function C is applied to a row of a credit hour matrix, a discipline complementation value is obtained; application to a column produces a program complementation. In Chapter III, entropy type measures are developed that apply to the entire matrix rather than only to a single row or column. Using these global measures credit hour matrices can be compared, e.g., to consider different points in time for the same educational institution or the same point in time for different institutions.

Application and Interpretation

The statistical or analytical relationship of discipline or program complementation to such items as unit costs, size of departments, nature of programs, etc., is not considered here, although such studies would be useful to determine how best to utilize the complementation measure. Any measure--whether it be unit cost, average number of FTE students, performance index, or complementation--is subject to possible misuse (Harris, 1973). For example, a complementation of zero probably is not desirable, but it does not necessarily follow that a value of 1 is ideal. For one thing, we would expect that the major program corresponding to a given discipline would absorb the major share of that discipline's output, thereby forcing the complementation to be less than 1. It is not the purpose here to state optimum values for particular departments or programs. What is suggested is that the complementation measure be used as a comparison device with respect to the conditions of monopoly and equilibrium.

The complementation concept is believed to furnish a new way of looking at the relationship between academic disciplines and instructional programs. For example, a discipline complementation can be regarded as a measure of the degree of integration of that discipline into all programs of the institution. On the other hand, a program complementation acts as a measure of interdisciplinarity for that program. In Chapter III a global complementation is proposed which in a certain sense can be regarded as a measure of total curriculum integration for the entire institution.

Summary

This dissertation makes use of the entropy concept of physics and of information theory to produce a mathematical measure of the extent to which academic disciplines and degree programs complement each other with respect to their production and consumption of credit hours. By a discipline distribution is meant a sequence of credit hours which flows from a given discipline to each degree program in a cluster of instructional programs. By a program distribution is meant a sequence of credit hours absorbed by a given program from each discipline in a cluster of academic disciplines. A discipline (program) distribution is said to satisfy the condition of monopoly if one program (discipline) absorbs (supplies) all the credit hours of the given discipline (program). A discipline (program) distribution satisfies the condition of equilibrium when all programs (disciplines) absorb (supply) the same number of credit hours. A discipline or program distribution satisfies the condition of competition whenever it does not satisfy the condition of monopoly. Equilibrium is a special case of competition and, in fact, represents maximum competition.

A distribution is normalized when its elements are replaced by proportional parts of its sum. Considering these proportional parts as probabilities with programs (disciplines) as events, a discipline (program) distribution can be considered to possess a value of entropy as given by the information theory of Shannon (1949). The complementation of a normalized distribution is then defined to be the ratio of its actual entropy to the maximum possible value. Complementation thus expresses the percentage of maximum entropy possessed by a given distribution. It is a measure of dividedness. The condition of monopoly produces a complementation of 0 while the condition of equilibrium gives a complementation of 1. In general, the complementation function ranks distributions on a scale from 0 to 1 with respect to where they stand in relation to monopoly and equilibrium.

Chapter II

REVIEW OF LITERATURE

Management Information Systems

As far as is known to the author, the application of entropy measures to credit hour matrices is new, and thus no previous articles on that application are available. This chapter therefore is restricted to a brief discussion of literature related to program planning, the credit hour matrix, and other applications of the entropy measure within the social sciences.

In any planning process information is vital. Friedman (1971a) defines a management information system (MIS) as

> ...an integrated, logically related set of policies, procedures and processes for collecting, ordering, storing, retrieving, analyzing, and reporting information required for effective discharge of management's responsibilities.

From this definition,the importance of management information systems to educational institutions is quite apparent. However, a number of difficulties have appeared. First, it is essential to match the MIS with the actual needs of management. Everyday operations require, in general, different levels of aggregation than does long range planning. The goal, perhaps unattainable, is for a "total information system." That is, one that will provide complete information for both the operational and planning aspects of management. In the case of public institutions there is the further subdivision of information into that useful for internal purposes and that required to be submitted to state and federal funding agencies. The ideal MIS would, presumably, provide both types of information.

Friedman (1971b) identifies twelve component processes of management which should be considered in the design of a management information system. They are the following:

1. Anticipating futures
2. Planning
3. Programming
4. Financing
5. Budgeting
6. Controlling
7. Organizing
8. Staffing
9. Administering
10. Evaluating
11. Relationship Building
12. Institutional Development

Clearly, as the above list implies, the development of a "total information system" will require a substantial effort.

Some authors have been quick to point out possible pitfalls in the use of these systems. For example, Ackoff (1967) has used the term "management misinformation systems" to describe many of them. A discussion of his provocative and controversial paper will not be given here but should be read by anyone seriously interested in the design of management information systems. Friedman (1966) has discussed the relationship between decision making and program planning budgeting systems (PPBS). He points out that PPBS, properly used, is a supplement to judgment and not its replacement. He further cautions against making hasty conclusions based on faulty cause-effect relationships.

The Ackoff article emphasizes the problems of excessive information. On the other hand, Davidson and Trueblood (1961) assert that both over- and under-information can impair the decision-making process. Excessive information can obscure the relevant; inadequate information may not contain it. The authors also point out that untimely information is another problem area. Information may come too late to be useful or may be supplied less frequently than necessary.

In spite of these limitations, it is generally thought that MIS offers great potential benefit to educational planning and administration.

Credit Hour Matrices

An important MIS output format is the credit hour matrix. The credit hour matrix concept was introduced to the academic community in an article by Suslow (1965). In this article, Suslow points out that an analysis based on the credit hour matrix was made as early as 1956 and developed into a projective technique programmed for computer application in 1964. Weathersby (1967) also used a form of credit hour matrix in his paper on university cost simulation.

Two basic credit hour matrices are employed in the simulation systems developed by NCHEMS at WICHE (Huff and Young, 1974).[1] They are the Induced Course Load Matrix (ICLM) and the Instructional Workload Matrix (IWLM). The ICLM for an educational institution is so constructed that each row gives the credit hours delivered by an academic discipline to an "average" student in each of the degree programs. Each column, therefore, contains the credit hours taken by an average student in a degree program from each of the academic disciplines. Usually, both disciplines and programs are divided into various levels. The levels normally considered are lower division, upper division, and graduate. The resulting configuration was called a coefficient

---

[1]NCHEMS is the National Center for Higher Education Management Systems.

WICHE is the Western Interstate Commission for Higher Education.

matrix or an input-output matrix by Suslow (1965) due to its resemblance to the input-output matrix of technological coefficients developed by Wassily Leontief of Harvard for economic analysis. However, the extensive use of this tool by the NCHEMS group at WICHE has resulted in the standard use of the terms ICLM and IWLM to refer to those credit hour matrices.

The ICLM shows the credit hours taken by the average student in each degree program from each of the academic disciplines. On the other hand, the IWLM gives the total number of credit hours in each academic discipline taken by the entire group of students enrolled in each degree program. Thus, for each historical ICLM, an associated IWLM can be obtained by the process of multiplying student enrollments in each degree program down through the corresponding column of the ICLM.

In practice, an historical IWLM is constructed from actual enrollment data. Five essential elements must be supplied by the student registration data. Assuming that aggregation will be at the departmental and degree program levels, the elements required are the following:

1. Department offering the course;

2. Course level,[2] i. e., lower division (freshman-sophomore), upper division (junior-senior), graduate, etc.;

3. Degree program, i. e., student major;

4. Program level,[2] i. e., lower division, etc.; and

5. Credit hours for the course.

The above items of information are obtained for each registrant in each course. Computer processing then permits the credit hours to be totaled for each department's contribution to each program by discipline level and program level. At this point, all the information for the historical IWLM is available. Traditionally, computer software packages have printed the entries in the IWLM/ICLM in disjointed sections. For example, a typical printing scheme would begin with, say, the lower division biology discipline and continue as before. This would constitute the contribution matrix. A separate print-out called the consumption-matrix would start with a program level and list its consumption

---

[2] "Course level" and "program level", respectively, refer to (a) the offering department's concept of the year in which the course normally will be appropriate for a student and (b) the actual status (i. e., freshman, sophomore, junior, etc.) of the students who enroll for this course. Departmental majors, of course, often take courses in the expected sequence, e. g., the biology major is likely to take third-year biology during his junior year. It is common, however, for seniors, for example, to enroll for first-, second- and third-year or masters-level courses, so that course level and program level are in fact separate factors.

of credit hours from all levels of each discipline.

A recent software package developed by Lewis (1976) permits the display of the IWLM/ICLM in full matrix format. Specifically, each page of the report prepared by the computer shows disciplines down the left margin and programs across the top. This permits portions of the credit hour matrix to be visually scanned in two dimensions rather than only one, as was the case earlier. Additional flexibility is provided with the inclusion of a "cluster" option. This option permits the user to select arbitrary subsets of disciplines or programs, i.e., "clusters", and to print a report only for the selected subset of the total matrix.

The columns of the historical IWLM are divided by the total number of full time equivalent (FTE) students in each program at each level to obtain the historical ICLM. From this historical ICLM, in turn, a projected IWLM can be built with projected enrollment figures by simple multiplication. Simplified examples of these matrices are given for a hypothetical institution in Tables I and II.

The major NCHEMS cost simulation package is known as the Resource Requirements Prediction Model 1.6 (Huff, 1973). It uses the institutional ICLM as required input along with program enrollment, faculty-student ratios, and other parameters. The

Table I

Key:
L=Lower Div.
U=Upper Div.
G=Grad. Div.

INDUCED COURSE LOAD MATRIX (Semester Hours)
MAJOR BY STUDENT LEVEL
(Sample)

		FTE ENROLLMENTS												
		143	186	52	121	94	45	85	61	17	180	206	204	
		History			Biology			Fine Arts			Business			
		L	U	G	L	U	G	L	U	G	L	U	G	Total
History	G	11.80			6.70	4.50		6.00	4.30		4.60	1.90		39.00
	U	4.20	11.30	4.50		3.90	2.11	2.29	7.61	7.41	2.30	6.10	4.70	56.42
	L			18.31										18.31
Biology	G	3.90	4.10		12.50			4.20	4.11	1.29	5.40			35.50
	U				5.80	13.70			2.00			4.00	2.80	28.30
	L					2.10	20.40							22.50
Fine Arts	G	3.70	2.70		2.70	1.80		10.89			0.60	3.50		25.89
	U		6.49	3.00		3.50	2.89	6.31	10.30			1.40	1.30	35.19
	L									19.29				19.29
Business	G	6.40	2.80		2.30	0.50		0.31	1.69		12.80	1.10		27.90
	U		2.60	4.19			4.60			2.00	4.30	10.30		27.99
	L											1.70	21.20	22.90
Total		30	30	30	30	30	30	30	30	30	30	30	30	360

Table II

Key:
L=Lower Div.
U=Upper Div.
G=Grad. Div.

INSTRUCTIONAL WORKLOAD MATRIX (Semester Hours)
MAJOR BY STUDENT LEVEL
(Sample)

		History			Biology			Fine Arts			Business			
FTE ENROLLMENTS		143	186	52	121	94	45	85	61	17	180	206	204	
		L	U	G	L	U	G	L	U	G	L	U	G	Total
History	G	1687			811	423		510	262		828	391		4912
	U	601	2102	234		367	95	195	464	126	414	1257	583	6438
	L			952										952
Biology	G	558	763		1512			357	251	22	972			4435
	U				702	1288			122			824	347	3283
	L					197	918							1115
Fine Arts	G	529	502		327	169		926			108	721		3283
	U		1208	156		329	130	536	628			288	161	3436
	L									328				328
Business	G	915	521		278	47		26	103		2304	227		4421
	U		484	218			207			34	774	2122		3839
	L											350	2629	2979
Total		4290	5580	1560	3630	2820	1350	2550	1830	510	5400	6180	3720	39420

output is a projected instructional program budget.

Hamelman (1970) makes use of the ICLM/IWLM concept in his description of inter-threshold resource planning in public higher educational systems. Huff and Clark (1973) and Haight and Martin (1975) give good discussions of ICLM preparation. In "Missions, Matrices and University Management," Hamelman (1970) points out the use of the credit hour matrix in the computation of program costs.

Aggregation

The terms "department" and "discipline" have been used almost interchangeably in the above. To see why, consider the question of aggregation. A credit hour matrix has been so defined that rows represent disciplines or levels of disciplines and columns represent programs or levels of programs. Clearly, the least discipline aggregation occurs when individual courses are listed.

Notice that two kinds of aggregation are employed. First, there is aggregation by discipline or program level: here, reference is to the hierarchy of lower division, upper division and graduate phases of study; that is, discipline and program level refer to the academic "levels" of course content and student programs, respectively. Second, there is greater or lesser aggregation of related courses into a hierarchy of sets and subsets: for example, the department of mathematics is at a higher level of aggregation than is a calculus course; similarly, the courses and disciplines of art, music, etc., may be further aggregated into "fine arts."

The level of aggregation is important in two significant ways. The first concerns the loss of information incurred when passing from a lower to a higher level of aggregation; the second concerns the question of the "stability" of the credit hour matrix as a predictive device. Within the credit hour matrix the information loss can occur in two dimensions. An example will make this clear. Consider the following normalized[3] credit hour matrix:

$$\begin{array}{ccccc}
\underline{\text{DISCIPLINES}} & & \underline{\text{PROGRAMS}} & & \\
 & P_1 & P_2 & P_3 \cdots & P_n \\
D_1 & a_{11} & \cdots & & a_{1n} \\
D_2 & \cdot & & & \cdot \\
D_3 & \cdot & & & \cdot \\
\cdot & \cdot & & & \cdot \\
\cdot & \cdot & & & \cdot \\
\cdot & \cdot & & & \cdot \\
D_m & a_{m1} & \cdots & & a_{mn}
\end{array}$$

---

[3] See Chapter I

Aggregation can occur among either disciplines or programs and in each case we can consider either the resulting discipline entropy or program entropy. Thus, there are the following four possibilities:

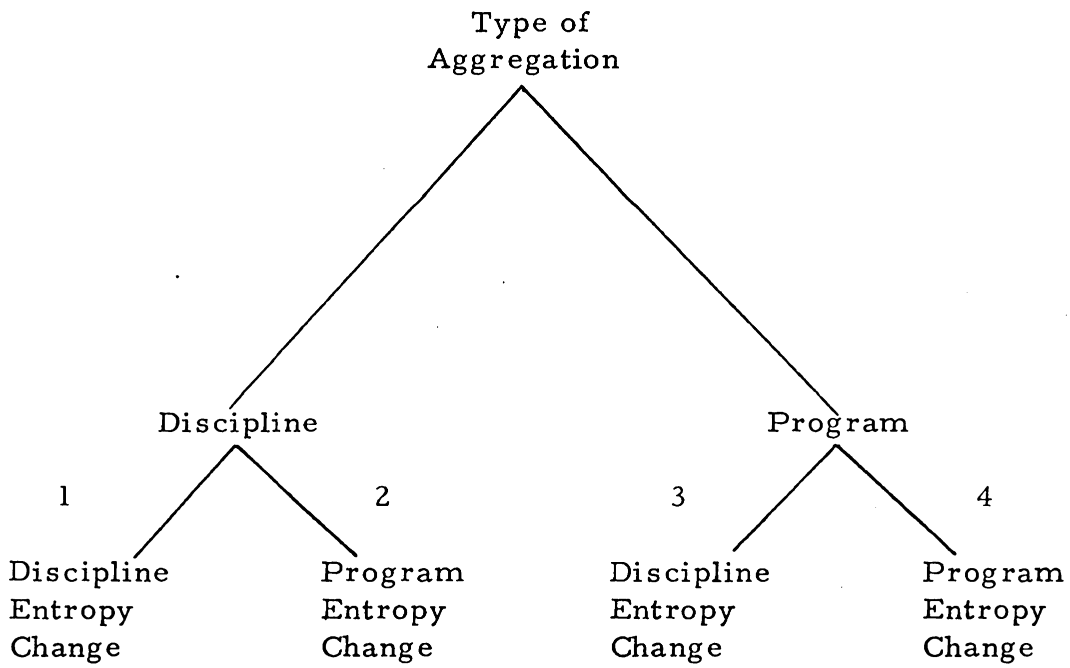
1. Discipline aggregation with the resulting change in discipline entropy.

2. Discipline aggregation with the resulting change in program entropy.

3. Program aggregation with the resulting change in discipline entropy.

4. Program aggregation with the resulting change in program entropy.

These possibilities are diagramatically illustrated as follows:



Suppose that for possibility 2 it is desired to aggregate D_1 and D_2. Before aggregation the program distribution is $P_1 = (a_{11}, a_{21}, \ldots, a_{m1})$. After aggregation the program distribution is $P_1' = (a_{11} + a_{21}, a_{31}, \ldots, a_{m1})$. Let S be the set of aggregated quantities, then $S = (a_{11}, a_{21})$. Let H be the entropy of P_1, H' the entropy of P_1', and H_s be the entropy of S. Then the change delta H in entropy due to aggregation is $H - H'$. Let $P_s = a_{11} + a_{21}$. With these definitions it is not difficult to show that delta $H = P_s H_s \geq 0$, with strict inequality holding if $a_{11} \neq 0 \neq a_{21}$. We are thus led to conclude that aggregation in this case has resulted in an information loss. Moreover, this loss is proportional to the portion aggregated (P_s) and to the "within group" entropy (H_s) of the portion aggregated.

Regarding the second question, that of stability, it would seem natural to find that stability increased with the degree of aggregation. That is, one would expect the variations in demand for individual courses to be greater than the variations in demand for an aggregated collection of courses offered by a department. This is analogous to the situation in thermodynamics discussed earlier. A gas maintained at constant temperature, pressure and volume may be thought of as representing great stability when examined in the aggregate. However, when the level of aggregation is reduced and molecules are examined, there is great instability.

Thus, in selecting a degree of aggregation, a trade off must be made between information loss and stability. Increase in aggregation produces an increase in stability but a decrease in information. Stated in another way, both stability and information loss are increasing functions of the degree of aggregation (Hudgins and Lewis, 1976.)

The employment of an entropy measure has been suggested as a means of comparison. If a credit hour matrix for one institution is to be compared with another, some very careful thought must be given to standardization. In this regard, the Higher Education General Information Survey (HEGIS) taxonomy codes become quite important. These codes (Huff, 1965) assign a four digit number to each of the standard disciplines taught at colleges and universities across the country today. The HEGIS codes form an integral part of the system of Information Exchange Procedures (IEP) that are intended to facilitate legitimate comparisons among educational institutions. In this system, the first digit of an eight-digit IEP Activity Structure Code identifies the highest level of aggregation within the structure. The second digit gives the next lower level of aggregation within the structure. For example, Instruction is labeled 1.0 and, under this 1.1 identifies General Academic Instruction. The next four digits (HEGIS) code will usually identify a specific discipline within one of the activity centers while the seventh and eighth digits define a course level within a discipline. In the HEGIS taxonomy codes, 20, 30 and 50 represent lower division, upper division, and graduate division, respectively. Thus, upper division general physics is coded 1.1.1902.30. Classification structures such as the HEGIS taxonomy codes become important if the complementation function C is to be used as a comparison device across institutional boundaries.

Other Applications of Entropy

Staaf and Kirschling (1974) have employed the entropy concept in their cell-counting approach to the understanding of choice in post-secondary education. Shannon (1949) was the first to give a mathematical formulation of information and entropy in probabilistic terms. Theil (1972) has made extensive use of Shannon's formulation within the social sciences. For example, he made use of the entropy function to measure the degree of (a) racial integration within school districts, (b) the distribution of votes among political candidates, and (c) industrial diversification. Zeleny (1974) has employed the concept of relative entropy to linear multiobjective programming. He has used entropy as a measure of importance in such a way as to provide a rational decision algorithm. Specifically, the algorithm selects an alternative from a finite set when weighted criteria are given.

An excellent bibliography relating to applications of entropy measure within the social sciences is given in the book by Theil (1972).

Chapter III

METHODOLOGY

Credit Hour Matrices

This dissertation is concerned with the problem of distilling the information provided by an ICLM and IWLM to show numerically certain aspects of the interrelationships existing between disciplines and programs. Recall that rows of the ICLM refer to discipline levels while columns refer to degree program levels. Thus, the entries down a column give the number of credit hours of each discipline level taken by an average student in the program level corresponding to the column selected. The entries across a given row are the numbers of credit hours of that discipline level taken by average students in each of the degree program levels. The IWLM is arranged in exactly the same way. The only difference is that all entries are aggregates rather than averages. Thus, the IWLM reflects total enrollment.

Whenever a degree program is based primarily on a specific discipline, the program is said to be associated with that discipline and vice versa. For example, the undergraduate degree program in mathematics is associated with the discipline of undergraduate level mathematics. In general, let m be the number of disciplines and n be the number of programs. Without loss of generality, the labeling can be arranged so that program P_k and discipline D_k are always associated up to some maximum index $t \geq 0$. That is, P_k and D_k are always associated for $k = 0, 1, 2, \ldots, t$. Thus, it may happen that $m \neq n$ and $t < \min(m, n)$. This would occur if there exist disciplines for which no degree program is offered or if a degree program is purely interdisciplinary in nature. It is assumed that each discipline is associated with at most one program and that each program is associated with at most one discipline. On this assumption it follows that $n > m$ implies that at least one program level is not associated with any discipline level and, analogously, that $m > n$ implies the existence of at least one discipline level not associated with any program level.

A credit hour matrix is called multidimensional if more than one discipline or program level is included. Multidimensional matrices are useful for some purposes but are inconvenient for the calculation of complementations. Thus, it is assumed here that each ICLM or IWLM contains only one discipline level and one program level. That is, the matrix under consideration is always an appropriate subset of the original multidimensional one. For example, it may be desired to examine the contributions of the lower division biology discipline to all upper division programs, etc.

Consider now the following illustrative questions. Questions 1 through 9 are phrased with respect to the biology discipline and program, but any discipline and program could obviously be used instead.

1. How are the ICLM/IWLM credit hours of the lower division biology discipline distributed across the various degree programs at each student (program) level?

2. How are the ICLM/IWLM credit hours of the upper division biology discipline distributed across the various degree programs at each student (program) level?

3. How are the ICLM/IWLM credit hours of the graduate division biology discipline distributed across the various degree programs at each student (program) level?

4. Which of the three levels of the biology discipline is most evenly distributed across the degree programs in the ICLM/IWLM?

5. How are the ICLM/IWLM credit hours of lower division biology majors distributed through the various disciplines at each level?

6. How are the ICLM/IWLM credit hours of upper division biology majors distributed through the various disciplines at each level?

7. How are the ICLM/IWLM credit hours of graduate biology majors distributed through the various disciplines at each level?

8. Which of the three levels of biology is most evenly distributed across the degree programs in the ICLM? In the IWLM?

9. Which of the three levels of students in biology is most evenly distributed through the various disciplines?

10. How do the above distributions vary over time for the same institution?

11. How do these distributions compare to those of similar institutions when the same disciplines and programs are considered both statically and dynamically?

12. How evenly is the program P_k distributed over the set of disciplines D_i which does not contain the associated discipline D_k, i.e., for $i \neq k$?

13. How evenly is the discipline D_k distributed over the set of programs P_j which does not contain the associated program P_k, i.e., for $j \neq k$?

14. How evenly is the Liberal Arts cluster of disciplines distributed across programs as compared to the non Liberal Arts cluster?

15. How do non Liberal Arts programs rank with respect to the evenness of their consumption of Liberal Arts disciplines?

16. How does the distribution of discipline output compare to the distribution of program consumption?

17. Can two ICLM/IWLM's compare with respect to an appropriate notion of curriculum integration?

It is asserted that the above and similar questions can be answered by the employment of the concept of complementation introduced in Chapter I. The power of the complementation approach is especially evident whenever distributions are to be compared on the basis of evenness as in the questions above.

Simulation models such as the NCHEMS Resource Requirements and Prediction Model (RRPM) require the IWLM as part of their data base. The projected IWLM is obtained from the historical ICLM by the simple process of multiplying anticipated program enrollments down through the columns of the ICLM. This process usually will produce row distributions very different from those of the ICLM. Nevertheless, the same kinds of questions can be addressed.

Example:

Given:

ICLM

	P_1	P_2	P_3	Sum
Enrollments	200	300	600	
D	30	20	10	60

The corresponding IWLM row would be as follows:

IWLM

	P_1	P_2	P_3	
D	6000	6000	6000	18000

Note that the ICLM distribution is uneven while the IWLM distribution for the same discipline D is perfectly even.

Since the ICLM is assumed to be independent of enrollments, it tends to reflect curriculum design. In contrast, the IWLM reflects the total workload induced by the given enrollments.

Hence, in the above example, unevenness is evident with respect to curriculum design but evenness occurs with respect to total workload considerations. That is, this particular enrollment pattern exactly cancelled out the unevenness inherent in the ICLM itself. For this reason, complementation measures should be applied to the ICLM for the purposes of curriculum description, analysis, and comparison and to the IWLM whenever budgetary considerations are of interest.

Complementation

Local Measures

In general, a credit hour matrix (ICLM or IWLM) can be represented in the following way:

	P_1	P_2	P_3	...	P_j	...	P_n	Row Sum	Prop. Parts
D_1	x_{11}	x_{12}	x_{13}	...	x_{1j}	...	x_{1n}	S_1	S_1/M
D_2	x_{21}	x_{22}	x_{23}	...	x_{2j}	...	x_{2n}	S_2	S_2/M
D_3	x_{31}	x_{32}	x_{33}	...	x_{3j}	...	x_{3n}	S_3	S_3/M
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
D_i	x_{i1}	x_{i2}	x_{i3}	...	x_{ij}	...	x_{im}	S_i	S_i/M
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$	$\vdots$
D_m	x_{m1}	x_{m2}	x_{m3}	...	x_{mj}	...	x_{mn}	S_m	S_m/M
Col. Sum	T_1	T_2	T_3	...	T_j	...	T_n	M = Matrix Sum	
Prop. Parts	$\frac{T_1}{M}$	$\frac{T_2}{M}$	$\frac{T_3}{M}$	...	$\frac{T_j}{M}$	...	$\frac{T_n}{M}$		

In this representation D_i stands for the ith discipline, P_j stands for the jth program, x_{ij} is the number of credit hours consumed from the ith discipline by the jth program, S_i is the total credit hours produced by the ith discipline, T_j is the total credit hours consumed by the jth program, and M is the sum of all entries in the matrix. Thus,

$$S_i = \sum_{j=1}^{n} x_{ij} ,$$

$$T_j = \sum_{i=1}^{m} x_{ij}, \text{ and}$$

$$M = \sum_{i=1}^{m} S_i = \sum_{j=1}^{n} T_j = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}$$

For ICLM's the x_{ij} are averages; for IWLM's the x_{ij} are aggregates.

The complementation measures to be presented here are based on the entropy concept of information theory (see Appendix B). Given a collection $(p_1, \ldots, p_n)$ of probabilities whose sum is 1, the entropy $H(p_1, \ldots, p_n)$ of the collection is given by

$$H(p_1, \ldots, p_n) \equiv \sum_{j=1}^{n} p_j \log \frac{1}{p_j}$$

where $\sum_{j=1}^{n} p_j = 1$. Since the p_j represents probabilities, the condition $0 \leq p_j \leq 1$ is also satisfied for each $j = 1, 2, \ldots, n$. For the case where for some j, $p_j = 0$ the expression $0 \log \frac{1}{0}$ is defined to be zero (see Appendix B). It is proven in Appendix B that the maximum value of $H(p_1, \ldots, p_n)$ occurs when and only when all p_j's are equal and this maximum entropy H_{max} is equal to log n. Relative entropy is then defined to be:

$$\frac{H(p_1, \ldots, p_n)}{H_{max}} = \frac{H(p_1, \ldots, p_n)}{\log n}$$

Thus, relative entropy is just that portion of the maximum entropy which is actually present. It is further demonstrated in Appendix B that relative entropy satisfies the five conditions required for a complementation measure (see Chapter II). Hence, complementation is defined as relative entropy.

Normalization. Complementation is then seen to be a function of n variables with n - 1 degrees of freedom. Each variable takes on values entirely contained within the unit interval. However, the x_{ij} entries in a credit hour matrix are not restricted in this fashion. How then is the complementation function to be applied? The answer lies in the process of normalization. The row normal form is obtained by replacing the x_{ij} entries of each row by the proportional parts of the row sum S_i. Column normalization is obtained in an analogous manner. Thus, the row $(x_{i1}, x_{i2}, \ldots, x_{in})$ becomes

$$\frac{x_{i1}}{S_i}, \frac{x_{i2}}{S_i}, \ldots, \frac{x_{in}}{S_i} \quad . \quad \text{Since } \sum_{j=1}^{n} \frac{x_{ij}}{S_i} = 1 \text{ and}$$

$0 \le \frac{x_{ij}}{S_i} \le 1$ for every i and j , the notation p_{ij} for $\frac{x_{ij}}{S_i}$ can be used to correspond with the original probabilistic definition of entropy in information theory (see Appendix B). With respect to the matrix (p_{ij}) define $p_{.j} \equiv \sum_{i=1}^{m} p_{ij}$. Then

$$\sum_{j=1}^{n} p_{.j} = \sum_{j=1}^{n} \sum_{i=1}^{m} p_{ij} = \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{x_{ij}}{S_i} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{x_{ij}}{S_i}$$

$$= \sum_{i=1}^{m} \frac{1}{S_i} \sum_{j=1}^{n} x_{ij} = \sum_{i=1}^{m} 1 = m$$

This is represented diagramatically as follows:

				Row Sums	Prop. Parts
	p_{11}	$\cdots$	p_{1n}	1	1/m
	$\vdots$	p_{ij}	$\vdots$	1	1/m
	p_{m1}	$\cdots$	p_{mn}	1	1/m
Col. Sums	$p_{.1}$	$\cdots$	$p_{.n}$	m = Matrix Sum	
Prop. Parts	$p_{.1}$	$\cdots$	$p_{.n}$		

Similarly, the column $(x_{ij}, x_{2j}, \ldots, x_{mj})$ becomes

$\left(\frac{x_{ij}}{T_j}, \frac{x_{2j}}{T_j}, \ldots, \frac{x_{mj}}{T_j}\right)$. If $\frac{x_{ij}}{T_j}$ is denoted by q_{ij}, it follows that the jth column can be represented as $(q_{1j}, q_{2j}, \ldots, q_{mj})$ where $0 \leq q_{ij} \leq 1$ and $\sum_{i=1}^{m} q_{ij} = 1$. With respect to the matrix (q_{ij}) define

$q_{i.} \equiv \sum_{j=1}^{n} q_{ij}$. Then $\sum_{i=1}^{m} q_{i.} = \sum_{i=1}^{m} \sum_{j=1}^{n} q_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{x_{ij}}{T_j}$

$= \sum_{j=1}^{n} \frac{1}{T_j} \sum_{i=1}^{m} x_{ij} = \sum_{j=1}^{n} 1 = n.$ The corresponding diagram is

				Row Sums	Prop. Parts
	q_{11}	$\cdots$	q_{1n}	$q_{1.}$	$q_{1.}/n$
	$\vdots$	q_{ij}	$\vdots$	$q_{i.}$	$q_{i.}/n$
	$q_{m.}$	$\cdots$	q_{mn}	$q_{m.}$	$q_{m.}/n$
Col. Sums	1	1	1	n = Matrix Sum	
Prop. Parts	1/n	1/n	1/n		1

Thus the entropy of the ith row is given by

$H_{i.} \equiv \sum_{j=1}^{n} p_{ij} \log \frac{1}{p_{ij}}$, while the entropy of the jth column is

$H_{.j} \equiv \sum_{i=1}^{m} q_{ij} \log \frac{1}{q_{ij}}$. Hence the complementation of the ith row is $C_{i.} \equiv \frac{H_{i.}}{\log n}$ and the complementation of the jth column is $C_{ij} \equiv \frac{H_{.j}}{\log m}$.

<u>Global Measures</u>

There are two natural ways to "average" the row complementations, viz., a simple average C_R and a weighted average $\bar{C}_R$. Define

$$C_R \equiv \frac{1}{m \log n} \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \log \frac{1}{p_{ij}} \quad \text{and}$$

$$\bar{C}_R \equiv \frac{1}{\log n \sum_{i=1}^{m} S_i} \sum_{i=1}^{m} \sum_{j=1}^{n} S_i \, p_{ij} \log \frac{1}{p_{ij}}$$

Similarly, for column complementation averages define

$$C_c \equiv \frac{1}{n \log m} \sum_{j=1}^{n} \sum_{i=1}^{m} q_{ij} \log \frac{1}{q_{ij}}$$

and $$\bar{C}_c \equiv \frac{1}{\log m \sum_{j=1}^{n} T_j} \sum_{j=1}^{n} \sum_{i=1}^{m} T_j \, q_{ij} \log \frac{1}{q_{ij}} \; .$$

Let $A \equiv (1, 2, \ldots, m)$

$B \equiv (1, 2, \ldots, n)$, and

$$M \equiv \sum_{i=1}^{m} S_i = \sum_{j=1}^{n} T_j = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}$$

Suppose that $D \subseteq A$ and $E \subseteq B$. Then a weighted complementation for the ith row with respect to the subset D can be defined by

$$\bar{C}_{i.}(D) \equiv \frac{S_i}{\sum_{k \in D} S_k} C_{i.} \; .$$

Similarly, for the jth column with respect to the subset E define

$$\bar{C}_{\cdot j}\ (E) \equiv \frac{T_j}{\sum_{k \in E} T_k}\ C_{\cdot j}\ .$$

If D = A, write $\bar{C}_{i.}$ for $\bar{C}_{i.}$ (D);

if E = B, write $\bar{C}_{\cdot j}$ for $\bar{C}_{\cdot j}(E)$.

Matrix Normalization. The (p_{ij}) matrix is obtained by row normalization and is convenient for the calculation of row complementation. The (q_{ij}) matrix is obtained by column normalization and is convenient for computing column complementations. If normalization is carried out with respect to the total sum of the (x_{ij}) matrix, a new matrix is obtained which is convenient for the calculation of both row and column complementation.

Define

$$r_{ij} \equiv \frac{x_{ij}}{M} \quad \text{where } M = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}\ .$$

Proceeding as before define

$$r_{i.} \equiv \sum_{j=1}^{n} r_{ij} \quad \text{and} \quad r_{\cdot j} \equiv \sum_{i=1}^{n} r_{ij}\ .$$

The diagram is

				Row Sums	Prop. Parts
	r_1	$\cdots$	r_{1n}	$r_{1.}$	$r_{1.}$
	$\vdots$	r_{ij}	$\vdots$	$\vdots$	
	$r_{m.}$	$\cdots$	r_{mn}	$r_{m.}$	$r_{m.}$
Col. Sums	$r_{.1}$	$\cdots$	$r_{.n}$	1	
Prop. Parts	$r_{.1}$	$\cdots$	$r_{.n}$		1 .

Note first that

$$\sum_{j=1}^{n} r_{.j} = \sum_{j=1}^{n}\sum_{i=1}^{m} r_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} \frac{x_{ij}}{M} = \frac{1}{M}\sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij} = 1$$

and $$\sum_{i=1}^{m} r_{i.} = \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{x_{ij}}{M} = \frac{1}{M}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = 1 \; .$$

Define <u>total complementation</u> C_T by

$$C_T \equiv \frac{1}{\log m\, n} \sum_{i=1}^{m}\sum_{j=1}^{n} r_{ij} \log \frac{1}{r_{ij}}$$

Even here a weighted average can be obtained by considering the product of the row and column sums corresponding to the element r_{ij} . That is, define

$$\bar{C}_T \equiv \frac{1}{\log m\,n} \; \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} S_i \, T_j \, r_{ij} \log \frac{1}{r_{ij}}}{\sum_{i=1}^{m} \sum_{j=1}^{n} S_i \, T_j} .$$

<u>Marginal Complementation.</u> Define row and column marginal probabilities by

$$r_i \equiv \sum_{j=1}^{n} r_{ij} \, , \; i = 1, 2, \ldots, m \quad \text{and} \quad r_{\cdot j} \equiv \sum_{i=1}^{m} r_{ij}, \; j = 1, 2, \ldots, n.$$

Then the marginal entropies are

$$H(D) \equiv \sum_{i=1}^{m} r_{i.} \; \log \frac{1}{r_{i.}} \; = \sum_{i=1}^{m} \frac{S_i}{M} \; \log \frac{M}{S_i} \quad \text{for disciplines, and}$$

$$H(P) \equiv \sum_{j=1}^{n} r_{i.} \; \log \frac{1}{r_{ij}} = \sum_{j=1}^{n} \frac{T_j}{M} \; \log \frac{M}{T_j} \quad \text{for programs}$$

The same definitions of row and column entropies can be applied to the column of row sums (S_i) and to the row (T_j) of column sums of the matrix (x_{ij}).

Define

$$H'(D) \equiv \sum_{i=1}^{m} \frac{S_i}{M} \; \log \frac{M}{S_i} \quad \text{and}$$

$$H'(P) \equiv \sum_{j=1}^{n} \frac{T_j}{M} \; \log \frac{M}{T_j} \; .$$

Then it is immediate that

$$H'(D) = H(D) = \sum_{i=1}^{m} r_{i.} \log \frac{1}{r_{i.}} \text{ and } H'(P) = H(P) = \sum_{j=1}^{n} \frac{T_j}{M} \log \frac{M}{T_j}.$$

This says that the marginal entropies are the same whether calculated directly from the (r_{ij}) matrix row and column sums or by using proportional parts of the row and column sums of the (x_{ij}) matrix.

Finally, the marginal complementations are given by

$$C(D) \equiv \frac{H(D)}{\log m} \text{ for disciplines, and } C(P) \equiv \frac{H(P)}{\log n} \text{ for programs.}$$

<u>Joint Complementation.</u> Define <u>joint entropy</u> by

$$H(D, P) \equiv \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} \log \frac{1}{r_{ij}} .$$ Then the joint complementation

is given by $$C(D, P) \equiv \frac{H(D,P)}{\log m + \log n}$$

<u>Conditional Complementations.</u> Fix a discipline D_i and allow the programs P_j to vary over $j = 1, 2, \ldots, n$. The conditional probabilities are then

$$\frac{r_{i1}}{r_{i.}}, \frac{r_{i2}}{r_{i.}}, \ldots, \frac{r_{in}}{r_{i.}} .$$ Thus, the conditional

program entropy with respect to discipline D_i is given by

$$H_i(P) \equiv \sum_{j=1}^{n} \frac{r_{ij}}{r_{i.}} \log \frac{r_{i.}}{r_{ij}} , \quad i = 1, 2, \ldots, m.$$

When r_{ij} is expressed in terms of x_{ij} it becomes evident that $H_i(P)$ is in fact equal to $H_{i.}$. Hence, $H_{i.}$ can be regarded as the conditional entropy of the programs with respect to discipline D_i. Similarly, the discipline's conditional entropy with respect to program P_j is

$$H_j(D) = \sum_{i=1}^{m} \frac{r_{ij}}{r_{.j}} \log \frac{r_{.j}}{r_{ij}} , \quad j = 1, 2, \ldots, n. \quad \text{Thus, } H_j(D) = H_{.j}.$$

The conditional complementations are therefore, $C_{i.}$ and $C_{.j}$.

Average Conditional Complementation. $H_i(P)$ measures the uncertainty of the programs P only under the condition $D = D_i$. But this condition is satisfied with probability $r_{i.}$. Therefore, the average conditional entropy of P over all disciplines D is given by

$$H_D(P) = \sum_{i=1}^{m} r_{i.} H_i(P) = \sum_{i=1}^{m} r_{i.} H_{i.} = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} \log \frac{r_{i.}}{r_{ij}} .$$

Thus, $H_D(P)$ measures the average uncertainty of the programs over the disciplines. Similarly,

$$H_P(D) = \sum_{j=1}^{n} r_{.j} H_j(D) = \sum_{j=1}^{n} r_{.j} H_{.j} = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} \log \frac{r_{.j}}{r_{ij}}$$

measures the average uncertainty of disciplines over the programs. But the maximum value of $H_D(P)$ occurs when $H_{i.} = \log n$. Then

$$H_D(P) = \sum_{i=1}^{m} r_{i.} H_{i.} \quad \sum_{i=1}^{m} r_{i.} \log n = \log n \sum_{i=1}^{m} r_{i.}$$

$$= \log n \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} = M \log n.$$

Hence, the average complementation of the programs over the disciplines can be defined as $C_D(P) \equiv \frac{H_D(P)}{M \log n}$. Similarly, the average complementation of the disciplines over the programs is $C_P(D) \equiv \frac{H_P(D)}{M \log m}$. It is easily shown that $C_D(P) = \frac{\bar{C}_R}{M}$ and $C_P(D) = \frac{\bar{C}_c}{M}$.

All previously defined complementations can be obtained from the (r_{ij}) matrix as follows:

(1) $$C_{i.} = \sum_{j=1}^{n} \frac{r_{ij}}{r_{i.}} \log \frac{r_{i.}}{r_{ij}}$$

(2) $$C_{.j} = \sum_{i=1}^{m} \frac{r_{ij}}{r_{.j}} \log \frac{r_{.j}}{r_{ij}}$$

(3) $$C_R = \frac{1}{m \log n} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{r_{ij}}{r_{i.}} \log \frac{r_{i.}}{r_{ij}}$$

(4) $$\bar{C}_R = \frac{1}{M \log n} \sum_{i=1}^{m} \sum_{j=1}^{n} M\, r_{ij} \log \frac{r_{i.}}{r_{ij}}$$

$$= \frac{1}{\log n} \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} \log \frac{r_{i.}}{r_{ij}} ,$$

since $p_{ij} = \frac{x_{ij}}{S_i} = \frac{r_{ij}}{r_{i.}}$, $S_i\ p_{ij} = S_i \cdot \frac{x_{ij}}{S_i} = x_{ij}$,

and $\frac{x_{ij}}{M} = r_{ij}$. Hence, $x_{ij} = M\ r_{ij}$

(5) $$C_c = \frac{1}{n \log m} \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{r_{ij}}{r_{.j}} \log \frac{r_{.j}}{r_{ij}}$$

(6) $$\bar{C}_c = \frac{1}{M \log m} \sum_{j=1}^{n} \sum_{i=1}^{m} M\ r_{ij} \log \frac{r_{.j}}{r_{ij}}$$

$$= \frac{1}{\log m} \sum_{j=1}^{n} \sum_{i=1}^{m} r_{ij} \log \frac{r_{.j}}{r_{ij}}$$

since $q_{ij} = \frac{r_{ij}}{r_{.j}}$, $T_j\ q_{ij} = x_{ij}$, and $x_{ij} = M\ r_{ij}$.

(7) $\bar{C}_i$ (D) is unchanged .

(8) $\bar{C}_{.j}$ (E) is unchanged .

<u>Mutual Complementation.</u> The <u>mutual information</u> $M(D_i, P_j)$ of the matrix (r_{ij}) is defined by $M(D_i, P_j) \equiv \log \frac{r_{ij}}{r_{i.}\ r_{.j}}$. The

expected mutual information $M(D,P)$ of the matrix (r_{ij}) is defined by

$$M(D,P) \equiv \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} \; M(D_i, P_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} \; \log \frac{r_{ij}}{r_{i.} \; r_{.j}} \; .$$

<u>The mutual complementation</u> $C_M(D,P)$ of the n x n matrix (r_{ij}) is defined by $C_M(D,P) \equiv \dfrac{M(D,P)}{Q}$ where Q is the maximum value of $M(D,P)$. It is shown in Appendix B that this maximum occurs only when (r_{ij}) is a diagonal matrix, i.e., when all off diagonal elements are zero. It is further shown that the maximum occurs when and only when these disgonal elements are equal, i.e., each diagonal element must equal M/n. Then, $M(D,P) = \log n = Q$ and $C_M(D,P) = \dfrac{M(D,P)}{\log n}$. Note that $M(D,P)$ is defined for an arbitrary m x n matrix (r_{ij}) while $C_M(D,P)$ is defined only for square matrices. $C_M(D,P)$ can be interpreted as follows:

(1) When $C_M(D,P)$ is near zero, the matrix (r_{ij}) is close to the condition of stochastic independence. This means that if the cell containing the element x_{ij} is selected with a probability r_{ij} (i.e., based on the proportion which x_{ij} bears to the matrix total M), then r_{ij} is also equal to the probability of selecting x_{ij} based first on selecting row i with a probability equal to the proportion which the row sum S_i bears to M and then selecting column j with a probability

equal to the proportion which the column sum bears to M, or vice versa.

(2) When $C_M(D, P)$ is near one, the matrix (r_{ij}) is close to the condition of global monopoly. This means that (r_{ij}) is close to being a diagonal matrix, i. e., close to the situation where each discipline supplies credit hours only to its own major program. Of course, the assumption is made here that (r_{ij}) is a square (n x n) matrix such that the kth discipline is associated with the kth program for each k = 1, 2, . . . , n.

<u>Example 1.</u>

	P_1	P_2	P_3	S_i	$H_{i.}$	$C_{i.}$
D_1	10	0	0	10	0	0
D_2	0	10	0	10	0	0
D_3	0	0	10	10	0	0
T_j	10	10	10	30=M		
$H_{.j}$	0	0	0			
$C_{.j}$	0	0	0			

$H(D) = \ln 3 = 1.0986 \qquad C(D) = 1$

$H(P) = \ln 3 = 1.0986 \qquad C(P) = 1$

$H(D, P) = \ln 3 = 1.0986 \qquad C(D, P) = .5$

$H_D(P) = \ln 3 = 1.0986 \qquad C_D(P) = .0333$

$H_P(D) = \ln 3 = 1.0986 \qquad C_P(D) = .0333$

$M(D, P) = \ln 3 = 1.0986 \qquad C_M(D, P) = 1$

Example 2.

	P_1	P_2	P_3	S_i	$H_{i.}$	$C_{i.}$
D_1	10	10	10	30	1.0986	1
D_2	10	10	10	30	1.0986	1
D_3	10	10	10	30	1.0986	1
T_j	30	30	10	90=M		
$H_{.j}$	1.0986	1.0986	1.0986			
C_{ij}	1	1	1			

$H(D)$	=	1.0986	$C(D)$	=	1
$H(P)$	=	1.0986	$C(P)$	=	1
$H(D, P)$	=	2.1972	$C(D, P)$	=	1
$H_D(P)$	=	1.0986	$C_D(P)$	=	.0111
$H_P(D)$	=	1.0986	$C_P(D)$	=	.0111
$M(D, P)$	=	0	$C_M(D, P)$	=	0

Example 3.

	P_1	P_2	P_3	S_i	$H_{i.}$	$C_{i.}$
D_1	2	3	3	9	1.0609	.9656
D_2	5	6	7	18	1.0893	.9915
T_j	7	9	11	27=M		
$H_{.j}$	.5983	.6365	.6555			
$C_{.j}$	.8631	.9183	.9457			

$H(D)$	=	.6365	$C(D)$	=	.9183
$H(P)$	=	1.0820	$C(P)$	=	.9849
$H(D, P)$	=	1.7163	$C(D, P)$	=	.9579
$H_D(P)$	=	1.0798	$C_D(P)$	=	.0364
$H_P(D)$	=	.6343	$C_P(D)$	=	.0339
$M(D, P)$	=	.0022	$C_M(D, P)$	=	...

Chapter IV

APPLICATION

Comparing Complementations

The complementation measure has been introduced and developed in Chapters I through III. This chapter will be devoted to its application and possible interpretation.

Complementation is intended as a comparison measure. Its numerical value for any single discipline or program and for any single period gives only a measure of the degree of monopoly involved. That isolated measure can be compared only to pure monopoly (C = 0) or perfect competition (C = 1). Given a set of numerical values for various disciplines and programs within an institution, however, the several disciplines and programs may be compared with each other on the basis of their relative complementation values, and on the basis of changes in those values from one time period to another. For example, Discipline A may have a complementation of 0.2 while Discipline B has complementation equal to 0.9. What conclusion can be drawn from these values? Discipline A is much closer to the condition described as that of monopoly. In fact, A is closer by the amount 0.7. If this is all the information available little more can be said. The 0.2 value for Discipline A may represent no cause for alarm and similarly the 0.9 value for Discipline B may be quite normal. These isolated values, alone, are not fully informative.

The comparisons available with the complementation measure become greater, however, when (a) data are available for the institution during two or more periods, so that past and present may be compared, (b) like elements of the institution may be compared legitimately, or (c) there is reason to believe that inter-institutional comparisons can be made legitimately.

Suppose now that considerable complementation statistics are available. For example, suppose that the national average of complementation values for Discipline A is 0.21 while that for Discipline B is .89. In this case then, based on complementation alone, there would be no cause for concern. On the other hand, suppose the figures to be as follows:

	Complementation	
Discipline	National Average	Our Institution
A	.8	0.2
B	.1	0.9

Can we now conclude that an imbalance exists in our case? Not necessarily. There may be very good reason for this departure from the national average. At least we should be motivated to look into the matter and make these reasons explicit. Romney (1972) states that the principal benefits of comparative analysis come from determining why differences exist.

Selection of Standards

Clearly, any standard used for this purpose should be carefully chosen. The above example employs "national average" as such a standard. This could, in fact, be very misleading. In the first place, a simple average does not necessarily imply that a "favored" value exists: if the complementation values for Discipline A across the country turn out to have no central tendency, their average value clearly then would be meaningless as a standard. Beyond this is the question of qualitative similarity. Should all institutions of higher education be included in the average or only those "similar" to the institution in question? It seems reasonable that a reference average be obtained using only institutions of "similar" character. Thus, we might expect a reference standard for a community college to differ from that of universities. Large urban community colleges might produce an average significantly different from that of small isolated community colleges, and both of these from liberal arts colleges located in rural areas. In addition to such physical characteristics as size and geographical environment it also may be necessary to consider goals, perspective, and even style. Balderston (1974) cites eight value dimensions useful for comparative analysis:

1. Long-term sustained vs. near-term relevant
2. Fundamental vs. applied
3. Advanced vs. basic start
4. Absolute achievement vs. value added
5. Idealistic-critical vs. utilitarian
6. Cosmopolitan vs. local
7. High-status style vs. common style
8. Selective vs. popular

Our most prestigious private universities may be seen as characterized by the first category in each of the above dimensions while most of the nation's community colleges would probably be viewed as representatives of the second category.

Thus, the selection of institutions to be included in the reference average is of prime importance. Of equal importance is the selection of programs (in the case of discipline complementation) or disciplines (in the case of program complementation). In comparing discipline complementations, for example, it is imperative that the same programs be used. Specifically, if mathematics complementation in institution A is based on twelve programs we should not try to compare this with that of mathematics in institution B where the complementation is based on fifteen programs. In fact, a comparison is not valid even in the case where the number of programs in institution B is also equal to twelve unless they are the same programs. For example, suppose that the complementation for mathematics in institution A is based on the English, history and art programs, while for institution B it is based on English, electrical engineering and chemistry. Then clearly, the two complementations are not comparable in the sense of contributing towards a reference average for the mathematics discipline. Even in this case, however, some information is gained. The comparison possible here is between the program triple (English, history, art) on the one hand and the triple (English, electrical engineering, chemistry) on the other. Implicit in this comparison is the assumption of similitude between the mathematics discipline at the two institutions. A way around this difficulty would be to obtain these complementations from the same institution. This could be done providing only that all five programs are offered.

In addition to the problem of selecting institutions appropriate for the accumulation of meaningful references averages, there is the question whether calculations should be made from the ICLM or from the IWLM. Under the assumption of stability, the ICLM is independent of enrollments and should, therefore, reflect the structure of curriculum design. Discipline complementations on the ICLM are, therefore, indicative of purely curriculum interrelationships. On the other hand, discipline complementations based on the IWLM reflect enrollment patterns as well. It is important to note that program complementations are independent of enrollments, hence are the same for the ICLM and the IWLM.

At the present time, complementation is presented as a descriptive rather than as an analytical device. This is necessarily the case since sufficient data have not yet been collected to establish the existence of useful reference standards. The collecting of such data by a single individual would be a formidable, if not impossible, task. However, the necessary information could be obtained with little difficulty if complementation data were included as part of the National Center for Higher Education Management Systems Information Exchange Procedures.

Curriculum Change

A simplified example suffices to illustrate the effect on complementation arising from curriculum change. Consider only two departments, Engineering (E) and Psychology (P), along with their associated major programs. Four cases are distinguished.

Case 1. The course in Industrial Psychology is transferred from the Psychology Department to Engineering.

	E	P
E	x_1	x_2
P	y_1+a	y_2

Before

	E	P
E	x_1+a	x_2
P	y_1	y_2

After

In this example, the number of credit hours for Industrial Psychology is equal to a. We assume each department always contributes the bulk of its hours to its own program. That is,

$y_2 > y_1+a$ and $x_1 > x_2$

Conclusion: The complementations of both E and P are lowered.

Case 2: Industrial Psychology is a new course offered only by Engineering to its own majors.

	E	P
E	x_1	x_2
P	y_1	y_2

	E	P
E	x_1+a	x_2
P	y_1	y_2

<u>Conclusion:</u> The complementation of E is lowered while that of P is unchanged.

<u>Case 3.</u> Industrial Psychology is a new course offered only by Psychology to Engineering majors.

	E	P
E	x_1	x_2
P	y_1	y_2

	E	P
E	x_1	x_2
P	y_1+a	y_2

<u>Conclusion:</u> The complementation of P is raised while that of E remains unchanged.

<u>Case 4.</u> The course in Industrial Psychology is transferred from Engineering to Psychology

	E	P
E	x_1+a	x_2
P	y_1	y_2

	E	P
E	x_1	x_2
P	y_1+a	y_2

<u>Conclusion:</u> The complementation of both E and P is raised.

We have assumed in this example that the course in Industrial Psychology is an applied course taken only by engineering majors.

Applications

We conclude this chapter with an example showing complementation measures based on realistic data. The institution selected is an actual community college of small size. The data have been aggregated to produce a 6 x 6 square matrix of associated disciplines and programs. Table III presents the ICLM. Table IV gives the ICLM Global Measures. Table V presents the IWLM. Table VI gives the IWLM Global Measures.

From Table VII it is immediately apparent that with respect to the ICLM, Life Sciences contributes more evenly to the programs than any other discipline. At the same time, as a program, Life Sciences draws more evenly from the disciplines than any other program.

Note that the ICLM and IWLM program complementation figures are nearly equal. This is not an accident. In theory these complementations are exactly equal since the column distributions differ only by a constant multiplier, viz., the enrollments. This, however, does not change the relative proportion of each distribution. The slight differences in the third and fourth decimal places are due to rounding error.

From Table VIII it is seen that with respect to the IWLM, Physical Science as a discipline contributes to programs more evenly than any other discipline. At the same time, as a program, Physical Science is the least uniform in its consumption of credit hours from disciplines.

Table III

COMMUNITY COLLEGE ICLM

			(B) 1	(LA) 2	(LS) 3	(PS) 4	(SS) 5	(VT) 6	S_i	$H_{i.}$	$C_{i.}$
Enrollments			142	128	226	66	164	145			
Business	(B)	1	17.70	1.45	.17	.67	3.21	.47	23.67	.8735	.4875
Liberal Arts	(LA)	2	6.57	19.06	2.91	7.41	8.76	3.37	48.08	1.5933	.8892
Life Science	(LS)	3	1.92	3.69	7.61	1.41	7.16	7.72	29.51	1.6270	.9080
Physical Science	(PS)	4	6.50	6.14	4.62	24.36	4.91	4.59	51.12	1.5287	.8532
Social Science	(SS)	5	4.19	6.57	1.99	2.86	12.90	1.40	29.91	1.5191	.8478
Vocational Technical	(VT)	6	.02	.09	3.38	.47	.05	19.44	23.45	.5535	.3089
$T_{.j}$			36.90	37.00	20.68	37.18	36.99	36.99	M = 205.74		
$H_{.j}$			1.3704	1.3182	1.5394	1.0475	1.5155	1.3217			
$C_{.j}$			.7649	.7357	.8592	.5846	.8458	.7377			

Table IV

ICLM GLOBAL MEASURES

$H(D)$	=	1.7409
$H(P)$	=	1.7739
$H(D, P)$	=	3.1109
$C(D)$	=	.9716
$C(P)$	=	.9900
$C(D, P)$	=	.8681
$M(D, P)$	=	.4039
$C_M(D, P)$	=	.2254
$H_P(D)$	=	1.3370
$H_D(P)$	=	1.3700
$C_P(D)$	=	.7462
$C_D(P)$	=	.7646

Table V

COMMUNITY COLLEGE IWLM

			(B) 1	(LA) 2	(LS) 3	(PS) 4	(SS) 5	(VT) 6	S_i	$H_{i.}$	$C_{i.}$
Enrollments			142	128	226	66	164	145			
Business	(B)	1	2514	185	38	44	527	68	3376	.8543	.4768
Liberal Arts	(LA)	2	933	2440	658	489	1436	488	6444	1.6062	.8964
Life Science	(LS)	3	273	472	1719	93	1174	1120	4851	1.5139	.8449
Physical Science	(PS)	4	923	786	1045	1608	805	666	5833	1.7462	.9746
Social Science	(SS)	5	595	841	449	189	2116	203	4393	1.4497	.8091
Vocational Technical	(VT)	6	3	12	763	31	8	2819	3636	.6038	.3370
$T_{.j}$			5241	4736	4672	2454	6066	5364	M = 28533		
$H_{.j}$			1.3707	1.3183	1.5391	1.0472	1.5153	1.3215			
$C_{.j}$			.7650	.7358	.8590	.5845	.8457	.7376			

Table VI

IWLM GLOBAL MEASURES

$H(D)$	=	1.7650
$H(P)$	=	1.7600
$H(D, P)$	=	3.1433
$C(D)$	=	.9850
$C(P)$	=	.9823
$C(D, P)$	=	.8771
$M(D, P)$	=	.3817
$C_M(D, P)$	=	.2130
$H_P(D)$	=	1.3833
$H_D(P)$	=	1.3783
$C_P(D)$	=	.7720
$C_D(P)$	=	.7692

Table VII

COMMUNITY COLLEGE ICLM
COMPLEMENTATION RANKINGS

Disciplines	$C_{i.}$	Programs	$C_{.j}$
Vocational Technical	.3089	Physical Science	.5846
Business	.4875	Liberal Arts	.7357
Social Science	.8478	Vocational Technical	.7377
Physical Science	.8532	Business	.7649
Liberal Arts	.8892	Social Science	.8458
Life Science	.9080	Life Science	.8592

Table VIII

COMMUNITY COLLEGE IWLM
COMPLEMENTATION RANKINGS

Disciplines	$C_{i.}$	Programs	$C_{.j}$
Vocational Technical	.3370	Physical Science	.5845
Business	.4768	Liberal Arts	.7358
Social Science	.8091	Vocational Technical	.7376
Life Science	.8449	Business	.7650
Liberal Arts	.8964	Social Science	.8457
Physical Science	.9746	Life Science	.8590

Also Table VII in comparison with Table VIII shows that all disciplines except Physical Science and Life Science retain the same rank order in both the ICLM and IWLM. The two exceptions, Physical Science and Life Science, simply interchange positions.

Finally, the tables show the following:

1. Vocational Technical exhibits the lowest complementation in both the ICLM and the IWLM and by roughly the same order of magnitude.
2. Physical Science and Life Science seem to be struggling for the highest discipline complementation ranking, with Life Science winning in the ICLM and Physical Science in the IWLM.
3. Physical Science appears as the most independent program, having the lowest program complementation.
4. Life Science appears as the most interdisciplinary, having the highest program complementation.

Table IX indicates a remarkable agreement between the global complementations computed from the ICLM and from the IWLM. This result is not so surprising, however, when the enrollment mix is examined. Enrollments are seen to be relatively uniform across the six programs. How uniform? Again, the complementation measure can be invoked to answer this question; the complementation is .9688. Finally, notice that the program with the lowest enrollment is Physical Science with 66 while the most popular program is Life Science with an enrollment of 226.

Table IX

GLOBAL COMPLEMENTATIONS

ICLM		IWLM
0.97	C(D)	0.99
0.99	C(P)	0.98
0.87	C(D, P)	0.88
0.75	$C_P(D)$	0.77
0.76	$C_D(P)$	0.77
0.23	$C_M(D, P)$	0.21

Chapter V

CONCLUSIONS

Purpose and Limitations

It is important to reemphasize that the complementations and other measures described herein are not offered as measures of value or intrinsic worth of disciplines, departments, programs, or institutions. This point was made earlier but is repeated here for emphasis. Clearly, the lower division physics discipline, for example, cannot be faulted if its complementation across the institution is low compared to, say, lower division English. In fact, this is exactly what one would expect. Thus, the first purpose in comparing complementations is not to make judgments based on some normative standard of "value" or "worth." Rather, the purpose is to gain information concerning the actual interrelationships which exist and to extend our understanding of them. No experiment has been performed here. No hypothesis has been tested, and no statistical analysis has been performed. The sole purpose of this dissertation has been to introduce a meaningful way to make quantitative comparisons between academic disciplines and student programs, based only on information provided in the

ICLM and IWLM of an educational institution.

For the future, there exist possibilities both for legitimate use and for misuse of these measures. For example, suppose extensive statistical data were obtained involving complementation measures and that from the data certain norms were discovered. Specifically, suppose that the mean and standard deviation of complementation values for lower division biology were known for a population of small colleges, all of which possess a well defined set of characteristics that, in the aggregate, permit them to be designated as Type A colleges. Suppose further that one Type A college was found to have a specified complementation significantly at variance with the norm. We at least would want to ask why. It would be interesting to compare, within Type A colleges, those which have high unit costs to those with low unit costs. Is complementation involved as a factor?

Quantification of Complementation

The term complementation as it has appeared in the literature of higher education has always been used in a qualitative sense. The main purpose of this dissertation has been to propose quantitative definitions that would permit numerical comparisons to be made. Balderston (1974) referring to the Instructional

Workload Matrix, states:

> If all off-diagonal coefficients were zero, no interdependencies among academic areas would need to be considered by the institutional administration. Doubling enrollment in a major would affect only the number of classes, faculty, and other resources in that academic area, and eliminating that major entirely would eliminate the work load for that field but leave all others unaffected. Typically, of course, the historic enrollment distributions, and the curriculum requirements that are legislated by faculties and approved in an institution, do show some cross-relations between fields. Thus, the chief academic administrator of a campus, implicitly viewing these interdependencies as a fact, sees them as implying independence (zero coefficients) or complementation (positive coefficients) among programs. If it is suggested that a given academic area be dropped, and if it has a history of substantial cross-relations with other academic areas, the costs of reorganization, redesign of curricula in the remaining fields, and dislocation of student and faculty preferences are likely to be substantial. Such consolidations often require complex sequences of administration -- faculty study, debate, and negotiation, the costly process known as academic reform.

In the above quotation, Balderston has used complementation to describe the condition of positive coefficients, and the term independence to describe the condition of zero coefficients. In this dissertation the term "monopoly" has been used instead of "independence," and the condition of positive coefficients has been given a quantitative definition. Thus, under a purely qualitative definition of complementation, it is known only that the off diagonal elements are positive. This is somewhat unsatisfactory as it is not known "how positive." What if all off diagonal coefficients are positive except for one, which is zero. This would represent a condition very close to complementation but fails to satisfy the definition. Suppose all the coefficients were positive but many were extremely close to zero. Technically, the qualitative definition would hold but the matrix would be very close to non complementation. Using the quantitative definition presented in Chapter I, the degree of complementation existing can be stated precisely and still be consistent with the notion as used by Balderston (1974).

Some Uses of Complementation

What then are the uses of complementation? This question has been partly answered earlier but is considered again here. A counter question is "useful to whom?" The academic vice president and the various college deans will require that data be presented at a higher level of aggregation than will department heads and institutional researchers, but the same types of comparison can be made. Within an institution, discipline and program complementations can be compared. For example, when disciplines and programs are rank ordered, administrators and planners can tell at a glance the relative degrees of monopoly possessed by disciplines and the relative degree of interdisciplinarity possessed by programs. The final use to which this information will be put is, of course, a matter for the administrator to determine. One way in which it can be used has already been suggested by Balderston above, viz., to gauge the effects of proposed changes in curricula. During a period of financial difficulty, when "retrenchment" must be considered, the complementation measure could be of use in anticipating the consequences of different reductions in program. For example, a discipline possessing both high complementation and a large credit hour contribution across the institution can be expected to act as a critical element with respect to curriculum changes. On the other hand, a discipline with low complementation would have little effect even if its total credit hour count was high. Similarly, a program with high complementation and high credit hour count would have significant impact while a low complementation would imply a low impact.

In terms of the usefulness question, the complementation measure shares some similarity with the notion of unit cost. Of what use is unit cost? Is unit cost information useful in that the sole and continuing objective is to reduce unit cost? Of course not, for obviously quality could be shamelessly abandoned in the effort. Similarly, can the objective be only to achieve high complementation? Again, the question must be answered in the negative, for this result could be achieved by introducing unnecessary courses in each curriculum. Clearly, this is not desirable.

However, it can be said that lower unit cost is preferable to higher unit cost if quality is not lost in the process. Similarly, it is probably reasonable to prefer large complementations to smaller, provided that quality and appropriateness of the curriculum are not lost in the process. The alternative is monopoly or, to use Balderston's term, complete independence, i. e, each discipline supporting only its own major program.

The complementation measure is advanced as a new means for examining relationships (and changes in relationships) between and among academic disciplines and student programs. Final recipes are not offered regarding just how to "use it." Suggested uses are offered, however, and it is the hope of the author that the usefulness of the complementation measure hereafter will be extended, as increasingly reliable statistics are accumulated over time and as sophisticated users develop experience with its use and interpretation. The C measure is advanced as being potentially useful in both inter- and intra-institutional studies. In both cases, it is further advanced, sub-division into static and dynamic measurements predictably will be useful.

The following paragraphs contain suggestions regarding lines of inquiry that the C measure may facilitate. The enumeration is not intended to be exhaustive.

1. Breadth, variety, and integration of educational experience have long been valued within higher education. C can be used to measure how interdisciplinary each degree program is, hence, may become useful for illuminating the examination of that question and for providing a "scorekeeping" device, however rudimentary, for the purpose. Hence, the C measure is potentially useful for describing breadth and variety, if not the integration of educational experience.

2. The consumption of a department's credit hour production by students in degree programs other than its own is often considered to constitute the department's "service" role within the institution. C can be used as a measurement or index of each department's contribution of such service.

3. The volume and variety of courses and disciplines required for certain degree programs are pertinent for determining an institution's staffing requirements. The measures here advanced, in conjunction with the matrices and other quantitative devices, could be useful in deciding -- e.g., at a small, private liberal arts college -- whether a proposed new degree program is feasible: that is, could it be offered essentially by making further use of existing courses, manpower and facilities, or would it require extensive additions?

Suggestions for Further Research

1. Is there -- and should there be -- such a thing as a typical complementation value for a given discipline over a selected set of programs?
2. Is there -- and should there be -- a typical rank order of disciplines among institutions of similar type?
3. How do complementations vary over time at any one institution? How should they?
4. How do complementations vary from institution to institution?
5. How is complementation related to other organizational parameters?
6. What is the significance of the relationship between the complementation of a discipline and that of its associated major program?
7. Other things being equal, is a discipline with high complementation more secure financially during periods of scarcity?
8. What other characteristics are shared by institutions possessing high curriculum integration?

BIBLIOGRAPHY

BIBLIOGRAPHY

Ackoff, R. L. "Management Misinformation Systems." Management Science, XIV #4:147-156, December, 1967.

Angell, G. W. "Campus Master Plan, 1973-80." State University of New York College of Arts and Science, Plattsburgh, New York.

Arnheim, R. "Entropy and Art." An Essay on Disorder and Order. University of California Press, Berkeley, 1971.

Balderston, F. E. "Managing Today's University." San Francisco, Jossey-Bass, Inc., 1974.

Balzhiser, R. E., Samuels, M. R., and Eliassen, J. D. "Chemical Engineering Thermodynamics." Englewood Cliffs, N. J., Prentice Hall, 1972.

Boyd, J. D., Browne, A. D., Harcleroad, F. F., and Lawrence, B. "Comprehensive Information Systems for Statewide Planning in Higher Education." American College Testing Special Report Two, 1971.

Cope, R. G. (ed.). "Tomorrow's Imperatives Today." Proceedings of the 13th Annual Forum. Vancouver, British Columbia. The Association for Institutional Research, 1973.

Davidson, H. J. and Trueblood, R. M. "Accounting for Decision-Making." The Accounting Review XXXVI:577-582, 1961.

DeLuca, A., and Termini, S. "A Definition of a Nonprobabilistic Entropy in the Setting of Fuzzy Sets Theory. Information and Control, 20:301-312, 1972.

Dresch, S. P. "An Economic Perspective on the Evaluation of Graduate Education." Washington, D. C., National Academy of Sciences Technical Report #1, March 1974.

Drucker, P.F. "Long-Range Planning: Challenge to Management Science." Management Science, V, 1959.

Education Commission of the States. "Comprehensive Planning for Postsecondary Education." Denver, Colorado, Report of the Task Force on Statewide Comprehensive Planning for Postsecondary Education, Report #37, May, 1973.

Etherington, E.D. "Systems and Simulation; New Technology Goes to Work on Decision-Making." College and University Business, 46:55-61, March, 1969.

Farmer, J. "An Approach to Planning and Management Systems Implementation." California State Colleges, Los Angeles, 1971.

Fast, J.D. "Entropy." New York, Gordon and Breach Science Publishers, 1968.

Fincher, C. "Planning Models and Paradynamics in Higher Education." Journal of Higher Education, 43:754-767, 1972.

Friedman, B.D. "Program-Oriented Information." Maryland State Department of Education, Baltimore, 1966.

Friedman, B.D. "State Government and Education: Management in the State Agency." Chicago, Illinois, Public Administration Service, 1971 a.

Friedman, B.D. "Information Systems and School District Management." Chicago, Illinois, Public Administration Service, 1971 b.

Georgescu-Roegen, N. "The Entropy Law and the Economic Process." Cambridge, Harvard University Press, 1971.

Haight, M. and Martin, R. "NCHEMS Costing and Data Management Systems Student Data Module Reference Manual." National Center for Higher Education Management Systems at WICHE, Technical Report #60, January, 1975.

Hakaka, D. R. "Resource Allocation Problems of Reduced Growth in College Enrollment." Journal of the College and University Personnel Association, 25:40-43, January, 1974.

Hamelman, P. W. "Mission, Matrices and University Management." Academy of Management Journal, March, 1970.

Hamelman, P. W. "Inter-Threshold Resource Planning for Public Higher Education Systems." Socio-Economics Planning Science, 4:469-485, 1970.

Harris, A.H. "The Comparability Question! Potential Uses and Misuses of Data, Tomorrow's Imperatives Today." Cope, R. G., ed. Proceedings of the 13th Annual Forum, Vancouver, British Columbia, The Association for Institutional Research, 1973.

Hoenack, S. A. "Incentives and Resource Allocation in Universities." Journal of Higher Education, 45:21-37, January 1974.

Hudgins, M. W. and Lewis, J. M. "Stability of Credit Hour Matrices." Unpublished Monograph, New River Community College, 1976.

Huff, R. A. "Focus on MIS." A Report on the WICHE-ACE Higher Education Management Information Systems Seminar. Washington, D. C., April 24-26, 1969. Western Interstate Commission for Higher Education, 1969.

Huff, R. A. "Overview of the Cost Estimation Model." Boulder, National Council on Higher Education Management Systems at WICHE, April, 1971.

Huff, R. A. "Higher Education Planning and Management Systems, A Brief Explanation." Boulder, National Council on Higher Education Management Systems at WICHE, May, 1972.

Huff, R. A. "Instructional Program Budgeting in Higher Education." Boulder, National Council on Higher Education Management Systems at WICHE, September, 1972.

Huff, R. A. "Introduction to the Resource Requirements Prediction Model 1.6." Boulder, National Council on Higher Education Management Systems at WICHE, Technical Report #31A, 1973.

Huff, R.A. and Young, M.E. "A Blueprint for RRPM 1.6 Application." Boulder, National Council on Higher Education Management Systems at WICHE, April, 1973.

Huff, R.A. and Young, M.E. "Profiles of Management Information Uses." Boulder, National Council on Higher Education Management Systems at WICHE, May, 1974.

Huff, R.A. "Information Exchange Procedures, Cost Study Procedures Manual." Boulder, National Council on Higher Education Management Systems at WICHE, Technical Report #65, June, 1965.

Khoury, E.N. and Nelson, H.W. "Simulation in Financial Planning." Management Sciences, II, 1965.

Lewis, J.M. "The Development of a Computerized Costing and Budgeting Simulation Model for Higher Education." Unpublished Dissertation, Virginia Polytechnic Institute and State University, 1976.

Lukitsh, R.A. "Community College Built Itself a Cost-Analysis System Based on Experience." College and University Business, 55:26-27, December, 1973.

MacKay, D.M. "Information, Mechanism and Meaning." Cambridge, M.I.T. Press, 1969.

Management Division Academy for Educational Development. Let's End the Confusion About Simulation Models!" Society for College and University Planning, 1973 Spring Conference 2.

Meeth, L.R. "Quality Education for Less Money." San Francisco, Jossey-Bass Series in Higher Education, 1974.

Minihan, J.P. "Administrative Cost Accounting: Whose Cost and Whose Accounting?" Journal of Higher Education, 45:38-47, January, 1974.

Miyataki, Glenn, K. "Academic Unit Planning Manual, Field Review Edition." Boulder, National Council on Higher Education Management Systems at WICHE, Technical Report #72, January, 1975.

Morowitz, H. J. "Entropy for Biologists; An Introduction to Thermodynamics." New York, Academic Press, 1970.

Morse, P.M. "Thermal Physics." New York, W.A. Benjamin, Inc., 1969.

Raisbeck, G. "Information Theory. An Introduction for Scientists and Engineers." Cambridge, M.I.T., 1949, 1972.

Romney, L. C. "Information Exchange Procedures. A Synopsis." Boulder, National Council on Higher Education Management Systems at WICHE Technical Report #28, September, 1972.

Sandin, R.T. (ed.). "Educational Comment. The University as a System." Ohio, The University of Toledo, 1974.

Shannon, C.E. and Weaver, W. "The Mathematical Theory of Communication." Chicago, University of Illinois Press, 1949.

Shoemaker, W.A. "Systems Models and Programs for Higher Education. A Catalogue." College Management, 9:12-14, and 35, April, 1973.

Sivazlian, B.D. and Stanfel, L.E. "Analysis of Systems in Operations Research." Englewood Cliffs, New Jersey, Prentice Hall, 1975.

Staaf, R. "Understanding Choice in Postsecondary Education: A Cell-Counting Approach." Boulder, National Council on Higher Education Management Systems at WICHE, August, 1974.

Stephens, H. (ed.). "Conversations, Faculty Productivity." ARA Services, Inc., 1973.

Suslow, S. "Design and Methodology in Institutional Research." C.H. Bagley (ed.). Proceedings of Fifth Annual National Institutional Research Forum. New York, State University of New York at Stony Brook. May 3-4, 1965.

Systems Research Group, Inc. "Campus VII, College and University Resource Planning, Washington, D.C., 1964.

Theil, H. "Statistical Decomposition Analysis." New York, North-Holland Publishing Co., 1972.

Topping, J. R. "Program Measures." Boulder, National Council on Higher Education Management Systems at WICHE Technical Report #TR-35, February, 1973.

von Bertalanffy, L. "General System Theory." New York, George Braziller, 1968.

Wartgow, J. F. "Computerized Institutional Planning Models: An Objective Analysis." Park Forest South, Illinois, Governors State University, March, 1973.

Watanabe, S. "Knowing and Guessing. A Formal and Quantitative Study." New York, John Wiley and Sons, 1969.

Weathersby, G. "The Development and Application of a University Cost Simulation Model." Berkeley Graduate School of Business Administration and Office of Analytical Studies, University of California, June, 1967.

Zeleny, M. "Lecture Notes in Economics and Mathematical Systems." M. Beckmann, Providence and H. P. Kunzi, Zurich, (Managing eds.). New York, Springer-Verlag, 1974.

APPENDICES

APPENDIX A

NON-ENTROPIC MEASURES

As previously defined, C is calculated in terms of the logarithm function. At first sight this may seem somewhat artificial. What is the rationale behind this?

First, some straightforward attempts will be made to accomplish the purpose without resorting to anything much more complicated than addition, subtraction, multiplication, or division. What properties are desired in a complementation measure? For simplicity, assume that only one discipline D is involved along with three programs, P_1, P_2, and P_3. Intuitively, C should be zero if one program takes all of the credit hours of D. Moreover, C should be equal to 1 if P_1, P_2, and P_3 share D equally. Then for the normalized row (1, 0, 0) C should be defined so that in this case C = 0 while for the row (1/3, 1/3, 1/3) the same definition of C should yield C = 1. In a certain sense it is desired to measure "how near" the elements of the row are to being equal. Maximum nearness to equality (i. e., the state of actually being equal) should produce a measure of 1. Minimum nearness of being equal (i. e., one element equal to 1 and the rest zero) should produce a measure of 0.

First Attempt

Since the concern is with how near to equality all the elements of a row actually are, try the following:

1. Find the set S of all possible pairwise differences. Average the numbers in S, and call it $\bar{s}$. Define $C = \bar{s}$. Recall that the ICLM has been normalized. Therefore, C defined in this way will be a number between 0 and 1.

Example 1.

	P_1	P_2	P_3
D	1/3	1/3	1/3

The differences are $1/3 - 1/3 = 0$, $1/3 - 1/3 = 0$ and $1/3 - 1/3 = 0$. The average $\bar{s}$ is $\frac{0+0+0}{3}$, i.e., $\bar{s} = 0$ and not 1 as desired. What happens if the definition of C is changed to $C \equiv 1 - \bar{s}$? Then, $C = 1 - 0 = 1$ and this is the desired result. Now try this definition on another example.

Example 2.

	P_1	P_2	P_3
D	1	0	0

The differences are $1 - 0 = 1$, $1 - 0 = 1$, and $0 - 0 = 0$

Hence, $\bar{s} = \frac{1+1+0}{3} = 2/3$ and $C \equiv 1 - \bar{s} = 1 - 2/3 = 1/3$ and not zero as desired.

Second Attempt

Since, in a certain sense, C should measure a kind of "variance" of the numbers in a row, we try the usual "variance" of statistics. Again using Example 1, the mean = $\bar{x}$ = 1/3 while

$$\text{variance} = \sigma^2 = \frac{(1/3 - 1/3)^2 + (1/3 - 1/3)^2 + (1/3 - 1/3)^2}{3 - 1} = 0.$$

Suppose C is defined as $C = 1 - \sigma^2$. Then $C = 1 - 0 = 1$ as desired. Applying this definition to Example 2, $\bar{x} = \frac{1 + 0 + 0}{3} = 1/3$

$$\text{while} \qquad \sigma^2 = \frac{(1 - 1/3)^2 + (0 - 1/3)^2 + (0 - 1/3)^2}{3 - 1}$$

$$= \frac{(2/3)^2 + (1/3)^2 + (1/3)^2}{2} = 1/3$$

Then, $C = 1 - \sigma^2 = 1 - 1/3 = 2/3$ and not zero as desired.

Third Attempt

First, find the maximum difference M between elements and form a new row matrix R' as follows:

$$R' = \left(\frac{x_1 - M}{M}, \frac{x_2 - M}{M}, \frac{x_3 - M}{M} \right)$$

where $R = (x_1, x_2, x_3)$ was the original normalized row of the ICLM. Then, define C to be the average of the elements of R', i.e.,

$$C \equiv \frac{x_1 - M + x_2 - M + x_3 - M}{3M} = \frac{x_1 + x_2 + x_3 - 3M}{3M}$$

Note that already a difficulty appears in trying to form R' from $R = (1/3, 1/3, 1/3)$, since $M = 0$ and division by zero is undefined.

Fourth Attempt

Let F_i be the set of all possible differences of the elements of the ith row of the ICLM. Let $L_i = \max_{f \in F_i} f$

$$\text{Define } C_i \equiv 1 - \frac{L_i}{\sum_{j=1}^{n} x_{ij}} = 1 - \frac{L_i}{S_i}$$

where the diagram is

$$\begin{array}{cccccc} & P_1 & P_2 \cdots & P_j \cdots & P_n & \\ D_i & x_{i1} & x_{i2} \cdots & x_{ij} \cdots & x_{in} & . \end{array}$$

Example 3.

	P_1	P_2	P_3	Sum
D_1	30	0	0	30

$$\text{Then } F_1 = \{30 - 0,\ 30 - 0,\ 0 - 0,\ 0 - 30\}$$
$$= \{\pm 30,\ 0\} \quad ,$$

$$L_1 = 30, \quad \text{and} \quad \sum_j x_{1j} = 30$$

Hence, $C_1 = 1 - \frac{30}{30} = 0$ as required. Working with the normalized row instead, given

	P_1	P_2	P_3
D_1	1	0	0

. The definition of C_1 would be $C_1 = 1 - \max\limits_{f \in F_1} |f|$.

Hence, $C_1 = 1 - 1 = 0$ as required.

<u>Example 4.</u>

	P_1	P_2	P_3	Sum
D_2	30	30	30	90

Then, $F_2 = (30 - 30,\ 30 - 30,\ 30 - 30) = (0)$

$$\max_{f \in F_2} |f| = 0, \text{ and } \quad x_{ij} = 90$$

$$\therefore \quad C_2 = 1 - \frac{0}{90} = 1 \text{ as required.}$$

Again, with the normalized form

	P_1	P_2	P_3
D_1	1/3	1/3	1/3

, it follows that

$$C_1 = 1 - \max_{f \in F_1} |f| = 1 - 0 = 1 \qquad \text{as required.}$$

Consequently, this measure works for the two extreme cases. How does it perform for intermediate situations?

Example 5.

	P_1	P_2	P_3	Sum
D_1	90	10	0	100

$$F_1 = \{\pm (90 - 10), \pm (90 - 0), \pm (10 - 0)\}$$
$$= \{\pm 80, \pm 90, \pm 10\}.$$

Then $\max_{f \in F_1} |f| = 90$ and $x_{ij} = 100$.

$$\therefore C_1 = 1 - \frac{90}{100} = 1 - .9 = .1$$

For this row the entropy measure is .3.

Example 6.

	P_1	P_2	P_3	Sum
D_2	40	35	25	100

$$F_2 = \pm (40 - 35), \pm (40 - 25), \pm (35 - 25)$$
$$= \pm 5, \pm 15, \pm 10$$

Then $\max_{f \in F_2} |f| = 15$, $\sum_j x_{ij} = 100$,

and $C_1 = 1 - \frac{15}{100} = .85$. The entropy measure is .98.

Example 7.

	P_1	P_2	P_3	Sum
D_1	10	10	20	40
D_2	15	10	5	30
D_3	5	10	5	20

Maximum Difference Measure

$$C_1 = 1 - \frac{10}{40} = .75$$

$$C_2 = 1 - \frac{10}{30} = .66^+$$

$$C_3 = 1 - \frac{5}{20} = .72^+$$

Average = C = $.72^+$

Complementation (based on entropy)

$$C_1 = .95$$

$$C_2 = .92$$

$$C_3 = .95$$

Average = .94

Note: From the way in which the maximum difference measure was defined it would appear to require the following steps:

1. Calculate and store all possible differences between elements of the row. The number of these differences, assuming n columns, will be $\frac{n(n-1)}{2}$.

2. Search for the largest difference L_i.

3. Sum the row to get S_i.

4. Calculate $1 - \frac{L}{S_i}$ to get C_i.

The difficulty with this procedure is that if, for example, a university had 80 programs and 15 departments and assuming 3 levels, the ICLM would be a (15 x 3) x (80 x 3) matrix, i.e., it would contain 10,800 entries. Moreover, for each level the set of all possible differences could contain $\frac{80\ (80\ -\ 1)}{2} = 3160$ elements!

A Way Out

Instead of computing each of the $\frac{n(n\ -\ 1)}{2}$ differences, simply perform the following steps:

1. Find the largest entry A in the row.

2. Find the smallest entry B in the row.

3. Find the sum S_i of the row.

The the largest difference L_i is simply $L_i = A - B$. Hence,

$C_i = 1 - \frac{L_i}{S_i} = 1 - \frac{A - B}{S_i}$. When the normalized matrix p_{ij} is used the expression becomes $C_{i.} = 1 + \min_j p_{ij} - \max_j p_{ij}$

To distinguish this maximum difference measure from the relative entropy measure of entropy based complementation the former will be referred to as simply the difference measure C^d and the latter as the entropy measure C. This leads to the following definition: For the p_{ij} matrix define:

$C_{i\cdot}^{d} \equiv 1 + \min_{j} p_{ij} - \max_{j} p_{ij}$. For the q_{ij} matrix define

$C_{\cdot j}^{d} \equiv 1 + \min_{j} q_{ij} - \max_{j} q_{ij}$. For the r_{ij} matrix define

$C^{d} \equiv 1 + \min_{i,j} r_{ij} - \max_{i,j} r_{ij}$.

This is perhaps the simplest of all approaches. What are the disadvantages? How does it compare to the entropy measure? Clearly, the C^d measure has the advantage of being very easy to calculate. Equally clear is that its main disadvantage lies in the fact that it depends only on the largest and the smallest elements of the row, completely disregarding the variation which may occur in the remaining elements. For example, the row (12, 3, 6, 7, 8, 9, 10, 11) would have the same C^d value as the row (12, 3, 3, 3, 3, 3, 3, 3). Another disadvantage is that C^d is not a function of n as is the case with the entropy measure. For example, the rows (1/2, 1/2, 0), (1/2, 1/2, 0, 0), (1/2, 1/2, 0, 0, 0), etc., all have a C^d value of .5, but the C values would all be different. For the above case with n - 50, the entropy measure would yield C = .1772. Thus, the entropy measure takes all the elements of the row into account as well as the total number of elements whereas the difference measure only considers the largest and the smallest.

What are some other approaches? The ICLM matrix with maximum complementation is suggestive of chi-square analysis of observed and expected frequencies. Suppose the means of a row are used as the expected frequencies and chi-square is calculated as a measure of complementation. How would the result compare with the entropy measure? At least, chi-square is dependent on every entry of the row as well as the total number of entries in contrast to C^d.

Example 8.

expected	E_j	20	20	20	20	20	20
actual	O_j	16	19	27	17	23	18
normalized		$.133^+$	.158	.225	.142	.192	.150

Total = 120 = N

$C^d = 1 + .133 - .225 = .908$

C = .9902

Then,

$$\text{chi-square} = \sum_{j=1}^{n} \frac{(0_j - E_j)^2}{E_j} = \left(\sum_{j=1}^{n} \frac{0_j^{\,2}}{E_j} \right) - N = 4.40 \; .$$

Relative chi-square $= \dfrac{4.40}{600} = .0073^+$ where 600 is the maximum possible value of chi-square representing the worst malagreement. Since a relative chi-square value of zero represents perfect agreement of observed and expected values and a relative chi-square of 1 represents worst malagreement, one can obtain a measure of complementation by subtracting the relative value of chi-square from 1. Thus, define

$$C\ (\text{chi-square}) = 1 - (\text{relative chi-square}),$$

$$\text{i.e., } C\ (\text{chi-square}) = 1 - \frac{(\text{chi-square})}{\text{Max chi-square}}$$

In terms of our previous notation, $N = S_i$. Then

$$\begin{array}{lcccccccc} & P_1 & P_2 & \cdots & P_j & \cdots & P_n & S_i \\ D_i & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{in} & \sum_{j=1}^{n} x_{ij} \end{array}$$

and $\quad N = S_i = \sum_{j=1}^{n} x_{ij}$. So the row mean is

$$\bar{x}_{ij} = \frac{N}{n} = \frac{S_i}{n} = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$$

The worst malagreement occurs when for some $j = k$ it happens that $x_{ik} = N$ and $x_{ij} = 0$ for each $j \neq k$. Without loss of generality, suppose that $k = 1$. Then the row appears as

$$\begin{array}{lcccccccc} & P_1 & P_2 & \cdots & P_j & \cdots & P_n & S_i \\ D_i & N & 0 & \cdots & 0 & \cdots & 0 & N \end{array}$$

and $\quad \bar{x}_{ij} = \frac{N}{n} = \frac{x_{i1}}{n}$

Thus,

	P_1	P_2	P_3	...	P_j	...	P_n
0_j	N	0	0	...	0	...	0
E_j	$\frac{N}{n}$	$\frac{N}{n}$	$\frac{N}{n}$	...	$\frac{N}{n}$	...	$\frac{N}{n}$

$$\text{Hence chi-square} = \left(\sum_{j=1}^{n} \frac{0_j^{\,2}}{E_j}\right) - N$$

$$= \frac{N^2}{\frac{N}{n}} - N = nN - N = N\,(n - 1)$$

$$\text{i.e., maximum chi-square} = N\,(n - 1)$$

$$= S_i\,(n - 1)$$

$$\text{Thus, } C\,(\text{chi-square}) = 1 - \frac{\left(\sum_{j=1}^{n} \frac{x_{ij}^2}{S_i/n}\right) - S_i}{S_i\,(n - 1)}$$

Finally,

$$C(\text{chi-square}) = 1 - \frac{n \frac{\sum_j x_{ij}^2}{\sum_j x_{ij}} - \sum_j x_{ij}}{(n - 1) \sum_j x_{ij}} =$$

$$\frac{n}{n - 1} \cdot \frac{\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \sum_{j=1}^{n} x_{ij}^2}{\left(\sum_{j=1}^{n} x_{ij}\right)^2}$$

Note: In terms of the p_{ij}, $\text{C(chi-square)} = \frac{n}{n-1}\left(1 - \sum_{j=1}^{n} p^2_{ij}\right)$.

Applied to the data of Example 8, this formula yields C (chi-square) = .993. Using the unnormalized data a value of $.0073^+$ was obtained for the relative chi-square. Hence, C(chi-square) = $1 - .0073 = .9927 \cong .993$. In this same example, the entropy measure $C = \frac{.7705}{\log 6}$. Summarizing this information, it is seen that for the row (16, 19, 27, 17, 23, 18)

$$C = \frac{.7705}{\log 6} = .990,$$

$$C^d = .908, \quad \text{and}$$

$$\text{C(chi-square)} = .9937$$

Sum of Squared Differences. Ignoring the row, write x_j for x_{ij} and S for S_i.

Define

$$\sum \text{diff}^2 \equiv \sum_{j=1}^{n}\left(\frac{x_j}{S} - \frac{1}{n}\right)^2$$

$$= \sum_j \left(\frac{x_j^2}{S^2} - \frac{2}{n}\frac{x_j}{S} + \frac{1}{n^2}\right)$$

$$= \frac{1}{S^2}\left(\sum_j x_j^2\right) - \frac{2}{nS}\sum_j x_j + \sum_j \frac{1}{n^2}$$

$$= \frac{1}{S^2}\left(\sum_j x_j^2\right) - \frac{2}{n} + \frac{1}{n} = \frac{1}{S^2}\left(\sum_j x_j^2\right) - \frac{1}{n}$$

(since $\sum_j x_j = S$ and $\sum_{j=1}^{n} \frac{1}{n^2} = n \cdot \frac{1}{n^2} = \frac{1}{n}$).

Thus $$\sum \text{diff}^2 = \frac{1}{S^2}\left(\sum_{j=1}^{n} x_j^2\right) - \frac{1}{n}$$

But $$\frac{x_j}{S} = p_j \text{, hence, } x_j^2 = S^2 p_j^2$$

and $$\therefore \sum_j x_j^2 = S^2 \sum_j p_j^2$$

Hence, $$\sum \text{diff}^2 = \frac{1}{S^2} \cdot \left(S^2 \sum_j x_j^2\right) - \frac{1}{n}$$

$$= \left(\sum_j p_j^2\right) - \frac{1}{n}$$

Now max $\sum \text{diff}^2$ occurs when the diagram is

$$\begin{array}{cccccc} O_j & 1 & 0 & 0 & \ldots & 0 \\ E_j & \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{array} .$$

Then $$\max \sum \text{diff}^2 = \left(1 - \frac{1}{n}\right)^2 + \left(\frac{1}{n}\right)^2 \cdot (n - 1)$$

$$= 1 - \frac{2}{n} + \frac{1}{n^2} + \frac{n-1}{n^2} = \frac{n^2 - 2n + 1 + n - 1}{n^2}$$

$$= \frac{n^2 - n}{n^2} = \frac{n(n-1)}{n^2} = \frac{n-1}{n}$$

Hence, $(\text{Relative} \sum \text{diff}^2) = \dfrac{\sum \text{diff}^2}{\max \sum \text{diff}^2}$

$$= \frac{\left(\sum_j p_j^2\right) - \frac{1}{n}}{\frac{n-1}{n}} = \frac{n}{n-1}\left[\left(\sum_j p_j^2\right) - \frac{1}{n}\right]$$

Thus, $(\text{Relative} \sum \text{diff}^2) = \dfrac{n}{n-1}\left[\left(\sum_j p_j^2\right) - \dfrac{1}{n}\right]$

Define L by

$$L \equiv 1 - (\text{Relative} \sum \text{diff}^2) . \qquad \text{Then}$$

$$L = 1 - \frac{n}{n-1}\left[\left(\sum_j p_j^2\right) - \frac{1}{n}\right]$$

$$= 1 - \frac{n}{n-1} \sum_j p_j^2 + \frac{1}{n-1}$$

$$= \frac{n - 1 - n\sum p_j^2 + 1}{n-1}$$

$$\equiv \frac{n}{n-1}\left(1 - \sum_j p_j^2\right) = C(\text{chi-square}) \cdot$$

Thus L is equivalent to the chi-square measure. Note that the equality occurs since Relative $\sum \text{diff}^2$ = Relative chi-square. Note also that $\sum \text{diff}^2 \neq$ chi-square. In fact $\sum \text{diff}^2 = \left(\sum_j p_j^2\right) - \frac{1}{n}$

and $\max \sum \text{diff}^2 = \dfrac{n-1}{1}$ while chi-square $= S\left[\left(n\sum_j p_j^2\right) - 1\right]$ and max chi-square $= S(n-1)$

Let $y = \left(n \sum_j p_j^2\right) - 1$. Then

chi-square $= Sy$ and $\sum \text{diff}^2 = \frac{y}{n}$, while

max chi-square $= S(n-1)$ and max $\sum \text{diff}^2 = \frac{n-1}{n}$.

Hence:

Relative chi-square

$$= \frac{x^2}{\max\ ^2}$$

$$= \frac{Sy}{S(n-1)}$$

$$= \frac{y}{n-1}$$

Relative $\sum \text{diff}^2 =$

$$\frac{\sum \text{diff}^2}{\max \sum \text{diff}^2}$$

$$\frac{\frac{y}{n}}{\frac{n-1}{n}}$$

$$= \frac{y}{n-1}$$

<u>Sum of Absolute Deviations</u>

Define $A \equiv \sum_{j=1}^{n} \left| p_j - \frac{1}{n} \right|$

Now max A occurs when the diagram is

$$\begin{array}{llllll} p_j : & 1 & 0 & 0 & \cdots & 0 \\ \bar{p}_j : & \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{array}$$.

Thus, $\max A = 1 - \frac{1}{n} + \frac{1}{n}(n-1)$. Since $n \geqslant 1$,

$$\max A = \frac{n-1}{n} + \frac{n-1}{n} = 2\,\frac{n-1}{n}$$

Hence, $$\text{Relative } A = \frac{n}{2(n-1)} \sum_j \left| p_j - \frac{1}{n} \right|$$

Define C(A) = 1 - Relative A

Then $$C(A) = 1 - \frac{n}{2(n-1)} \sum_{j=1}^{n} \left| p_j - \frac{1}{n} \right| = 1 - \frac{n}{2S(n-1)} \sum_{j=1}^{n} \left| x_j - \frac{S}{n} \right|$$

and C(A) is a candidate for a complementation measure.

Let $U = \left\{ j : p_j - \frac{1}{n} \geqslant 0 \right\}$

and $B = \left\{ j : p_j - \frac{1}{n} < 0 \right\}$

Then $$\sum_{j=1}^{n} \left| p_j - \frac{1}{n} \right| = \sum_{j \in U} \left| p_j - \frac{1}{n} \right| + \sum_{j \in B} \left| p_j - \frac{1}{n} \right|$$

But $$\sum_{j \in U} \left| p_j - \frac{1}{n} \right| = \sum_{j \in B} \left| p_j - \frac{1}{n} \right|$$

since the sum of the deviations from the mean is always zero.

Hence, $$\sum_{j=1}^{n} \left| p_j - \frac{1}{n} \right| = 2 \sum_{j \in U} \left| p_j - \frac{1}{n} \right| = 2 \sum_{j \in B} \left| p_j - \frac{1}{n} \right|$$

$$= 2 \sum_{j \in U} \left(p_j - \frac{1}{n} \right) = 2 \sum_{j \in B} \left(\frac{1}{n} - p_j \right);$$ therefore,

$$C(A) = 1 - \frac{n}{n-1} \sum_{j \in U} (p_j - \frac{1}{n})$$

$$= 1 - \frac{n}{n-1} \sum_{j \in B} (\frac{1}{n} - p_j) = 1 + \frac{n}{n-1} \sum_{j \in B} (p_j - \frac{1}{n}) .$$

It may be supposed, without loss of generality, that there exists a maximum integer $k \geq 1$ such that for $j = 1, 2, \ldots, k$,

$p_j - \frac{1}{n} \geq 0.$ Then $\sum_{j \in U} (p_j - \frac{1}{n}) = \left(\sum_{j=1}^{k} p_j\right) - \sum_{j=1}^{k} \frac{1}{n}$

$= \left(\sum_{j=1}^{k} p_j\right) - \frac{k}{n} = \dfrac{n\left(\sum_{j=1}^{k} p_j\right) - k}{n}$. Hence $C(A) =$

$$1 - \frac{n}{n-1} \; \frac{n\left(\sum_{j=1}^{k} p_j\right) - k}{n} = 1 - \frac{n\left(\sum_{j=1}^{k} p_j\right) - k}{n-1} .$$

That is, $C(A) = \dfrac{n + k - 1 - n \sum_{j=1}^{k} p_j}{n-1}$.

But also, $C(A) = 1 - \frac{n}{n-1} \sum_{j \in B} (\frac{1}{n} - p_j)$.

Now $\sum_{j \in B} (\frac{1}{n} - p_j) = \left(\sum_{k+1}^{n} \frac{1}{n}\right) - \sum_{k+1}^{n} p_j$

$$= \frac{n-k}{n} - \sum_{k+1}^{n} p_j = \frac{n - k - n\sum_{k+1}^{n} p_j}{n}$$

Hence $C(A) = 1 - \frac{n}{n-1} \cdot \dfrac{n - k - n \sum_{k+1}^{n} p_j}{n}$

$$= 1 - \frac{n - k - n \sum_{k+1}^{n} p_j}{n-1} \qquad = \frac{n - 1 - n + k + n \sum_{k+1}^{n} p_j}{n-1}$$

That is, $C(A) = \dfrac{k - 1 + n \sum_{k+1}^{n} p_j}{n-1}$. To obtain C(A) in terms of the x_j, note that

$$C(A) = 1 \quad \frac{n}{2\,(n-1)} \sum_{j=1}^{n} \left| p_j - \frac{1}{n} \right|$$

$= 1 - \dfrac{n}{2\,(n-1)\,S} \displaystyle\sum_{j=1}^{n} \left| x_j - \frac{S}{n} \right|$ where $S = \sum_{j=1}^{n} x_j$. Also

$$C(A) = \frac{n + k - 1 - n \sum_{j=1}^{k} p_j}{n-1}$$

$= \dfrac{(n + k + 1)\,S - n \sum_{j=1}^{k} x_j}{S\,(n-1)}$. Alternatively,

$$C(A) = \frac{k - 1 + n \sum_{k+1}^{n} p_j}{(n - 1)}$$

$$= \frac{k - 1 + \frac{n}{S} \sum_{k+1}^{n} x_j}{n - 1} = \frac{(k - 1)\, S + n \sum_{k+1}^{n} x_j}{S(n - 1)}$$

The measure $C(A) = 1 - \frac{n}{2\, S(n-1)} \sum_{j=1}^{n} \left| x_j - \frac{S}{n} \right|$

suffers from a serious disadvantage which is illustrated in the following example.

Example 9.

	$n = 10$	Sum
D_1	(19, 9, 9, ..., 9)	100
D_2	(19, 1, 10, ..., 10)	100

Denote C(A) by A.

Then for D_1 we have $A_1 = 1 - \frac{10}{200\,(9)} \left[(19 - 10) + (10 - 9)\,(9)\right]$

$$= 1 - \frac{10}{200\,(9)}\,(9 + 9) = 1 - \frac{10}{200\,(9)}\,(18).$$ For D_2 ,

$$A_2 = 1 - \frac{10}{200\,(9)} \left[(19 - 10) + (10 + 1)\right] = 1 - \frac{10}{200\,(9)}\,(9 + 9)$$

$$= 1 - \frac{10}{200\,(9)}\,(18)$$ Hence, $A_1 = A_2$

Note, however, that $C_1 \neq C_2$, in fact $C_1 = .9841$ and $C_2 = .9520$ so that $C_2 < C_1$ as we should expect.

Thus, it has been shown that if C_1 and C_2 are the complementations for distributions D_1 and D_2, respectively, then $C_1 < C_2$ does not imply that $A_1 < A_2$ where A_1 and A_2 are the respective absolute difference measures described above. On the other hand, $A_1 < A_2$ does imply that $C_1 < C_2$. This will be proven here for distribution pairs of the form:

$$D_1(n) = \left(\frac{1}{n} - x, \frac{1}{n}, \ldots, \frac{1}{n}, \frac{1}{n} + x\right)$$

$$D_2(n) = \left(\frac{1}{n} - y, \frac{1}{n}, \ldots, \frac{1}{n}, \frac{1}{n} + y\right)$$

where $y < x < \frac{1}{n}$, $n > 2$.

The assertion is that $A_1 < A_2$ implies that $C_1 < C_2$.

First, it is shown that $A_1 < A_2$.

Now $A_1 = 1 - \frac{n}{n-1} \sum_{j \in B} \left(\frac{1}{n} - p_j\right)$

But here $B = \{1\}$, and $p_1 = \frac{1}{n} - x$.

Hence, $\frac{1}{n} - p_1 = \frac{1}{n} - \left(\frac{1}{n} - x\right) = x$. Then $A_1 = 1 - \frac{n}{n-1}\, x$.

Similarly, $A_2 = 1 - \frac{n}{n-y}$, But $y < x$. Therefore $-y > -x$, and

$$1 - \frac{n}{n - y} \cdot y > 1 - \frac{n}{n - 1} \quad x.$$

That is, $A_2 > A_1$ and it has been established that for the distribution pair (D_1, D_2), $A_1 < A_2$. To show that $C_1 < C_2$, it will suffice to show that

$$(\frac{1}{n} - x) \log \frac{1}{\frac{1}{n} = x} + (\frac{1}{n} + x) \log \frac{1}{\frac{1}{n} + x}$$

$$< (\frac{1}{n} - y) \log \frac{1}{\frac{1}{n} - y} + (\frac{1}{n} + y) \log \frac{1}{\frac{1}{n} + y} ,$$

i.e., that $\frac{1 - nx}{n} \log \frac{n}{1 - nx} + \frac{1 + nx}{n} \log \frac{n}{1 + nx}$

$$< \frac{1 - ny}{n} \log \frac{n}{1 - ny} + \frac{1 + ny}{n} \log \frac{n}{1 + ny} .$$

Denoting the left hand side and right hand side of the above inequality by LHS and RHS, respectively, it follows that

$$\text{LHS} = (\frac{1}{n} - x) \log \frac{n}{1 - nx} + (\frac{1}{n} + x) \log \frac{n}{1 + nx}$$

$$= \frac{1}{n} \left[\log n - \log (1 - nx)\right] - x \left[\log n - \log (1 - nx)\right]$$

$$+ \frac{1}{n} \left[\log n - \log (1 + nx)\right] + x \left[\log n - \log (1 + nx)\right]$$

$$= \frac{1}{n} \log n - \frac{1}{n} \left[\log (1 - nx)\right] - (x \log n) + x \log (1 - nx)$$

$$+ \frac{1}{n} (\log n) - \frac{1}{n} \log (1 + nx) + x \log n$$

$$- x \log (1 + nx) = \frac{2}{n} \log n - \frac{1}{n} \log (1 - nx)$$

$$+ x \log (1 - nx) - \frac{1}{n} \log (1 + nx) - x \log (1 + nx),$$

and

$$\text{RHS} = \frac{2}{n} \log n - \frac{1}{n} \log (1 - ny) + y \log (1 - ny)$$

$$- \frac{1}{n} \log (1 + ny) - y \log (1 + ny)$$

Thus, it suffices to show that

$$- \frac{1}{n} \log (1 - nx) + x \log (1 - nx) - \frac{1}{n} \log (1 + nx)$$

$$- x \log (1 + nx) < - \frac{1}{n} \log (1 - n) + y \log (1 + ny)$$

$$- \frac{1}{n} \log (1 + ny) - y \log (1 + ny) .$$

That is, it suffices to show that

$$(x - \frac{1}{n}) [\log (1 - nx)] - (x - \frac{1}{n}) \log (1 + nx)$$

$$< (y - \frac{1}{n}) [\log (1 - ny)] - (y + \frac{1}{n}) \log (1 + ny).$$

This is equivalent to showing that

$$(1 - nx) \log (1 - nx) + (1 + nx) \log (1 + nx)$$

$$> (1 - ny) \log (1 - ny) + (1 + ny) \log (1 + ny).$$

By hypothesis, $y < x < \frac{1}{n}$ and therefore

$ny < nx < 1$. Let $ny = a$ and $nx = b$. Then $a < 1$ and $b < 1$. Thus, it must be shown that

$$(1 - a) \log (1 - a) + (1 + a) \log (1 + a)$$

$$< (1 - b) \log (1 - b) + (1 + b) \log (1 + b),$$

i.e., that $\left[\log (1 - a)^{1 - a}\right] + \log (1 + a)^{1 + a}$

$< \left[\log (1 - b)^{1 - b}\right] + \log (1 + b)^{1 + b}$, i.e. that

$$\log \left[(1 - a)^{1 - a} (1 + a)^{1 + a}\right] < \log \left[(1 - b)^{1 - b} (1 + b)^{1 + b}\right].$$

But this will follow if it can be shown that

$$(1 - a)^{1 - a} (1 + a)^{1 + a} < (1 - b)^{1 - b} (1 + b)^{1 + b},$$

since the logarithm is a strictly increasing function. The problem now becomes the following: Show that

$(1 + a)^{1 + a} (1 - a)^{1 - a} \quad (1 + b)^{1 + b} (1 - b)^{1 - b}$ if $0 < a < b < \frac{1}{n}$ and $n \geq 2$. This can be proven as follows:

Let $f(a) = (1 + a)^{1 + a} (1 - a)^{1 - a}$. Then it suffices to show that f is an increasing function of a. Hence, it suffices to show that

$f' > 0$. Now $f(a) = (1 - a^2) \left(\dfrac{1 + a}{1 - a}\right)^a$

so that $f' = (1 - a^2) \dfrac{d}{da} \left(\dfrac{1 + a}{1 - a}\right)^a + \left(\dfrac{1 + a}{1 - a}\right)^a \dfrac{d}{da} (1 - a^2)$.

Let $y = \left(\frac{1 + a}{1 - a}\right)^a$ and $z = \frac{1 + a}{1 - a}$

Then $y = z^a$ and so $\ln y = a \ln z$ and $\frac{1}{y}\frac{dy}{da} = a\frac{1}{z} \cdot \frac{dz}{da} + \ln z.$

$$\text{That is, } \frac{dy}{da} = \left(\frac{1 + a}{1 - a}\right)^a \left[a \; \frac{1 - a}{1 + a} \cdot \frac{2}{(1 - a)^2} + \ln \frac{1 + a}{1 - a}\right]$$

$$= \left(\frac{1 + a}{1 - a}\right)^a \left(\frac{2a}{1 - a^2} + \ln \frac{1 + a}{1 - a}\right) = \frac{d}{da} \left(\frac{1 + a}{1 - a}\right)^a .$$

$$\text{Hence, } f' = (1 - a^2)\left(\frac{1 + a}{1 - a}\right)^a \left(\frac{2a}{1 - a^2} + \ln \frac{1 + a}{1 - a}\right)$$

$$+ \left(\frac{1 + a}{1 - a}\right)^a (-\,2a)$$

$$= \left(\frac{1 + a}{1 - a}\right)^2 \left[(1 - a^2)\left(\frac{2a}{1 - a^2} + \ln \frac{1 + a}{1 - a}\right) - 2a\right]$$

$$= \left(\frac{1 + a}{1 - a}\right)^a \left[2a + (1 - a^2)\left(\ln \frac{1 + a}{1 - a}\right) - 2a\right]$$

$$= \left(\frac{1 + a}{1 - a}\right)^a (1 - a^2) \ln \frac{1 + a}{1 - a} .$$

Finally, it must be shown that $f' > 0$ for $a \in (0, 1)$. Now $1 + a > 0$ since $a > 0$, and $1 - a > 0$ since $a < 1$. Hence, $\frac{1 + a}{1 - a} > 0$. Thus, $\left(\frac{1 + a}{1 - a}\right)^a > 0$.

Similarly, $a^2 < a < 1$ and therefore, $1 - a^2 > 0$. Thus, it suffices to show that $\ln \frac{1 + a}{1 - a} > 0$. But this is true if $\frac{1 + a}{1 - a} > 1$. Clearly, this is the case and the assertation is proved.

Appendix B

SOME RESULTS FROM INFORMATION THEORY

This appendix has been attached to give the interested reader a logical foundation for some of the information theory results which have been stated earlier without proof. No attempt is made at an exposition of the theory. For that purpose the reader is invited to consult one of the many excellent books on the subject, some of which are listed in the bibliography.

Definition 1

A probability distribution is a finite[1] collection $(p_1, p_2, \ldots, p_n)$ of numbers such that

(i) $0 \leq p_j \leq 1$, for each $j = 1, 2, \ldots, n$.

(ii) $\sum_{j=1}^{n} p_j = 1$

Definition 2

An experiment is a process whereby an element is selected from a given set.

---

[1] The general theory treats the case of infinite collections.

Definition 3

A message is a statement declaring the outcome of an experiment. All messages are assumed true.

Definition 4

The information content of a message is defined as follows: Let the elements of a set be numbered 1 through n. Suppose that in an experiment the probability of selecting the jth element is denoted by p_j . Then the information content I of a message stating that in an experiment the jth element was chosen is given by

$$I \equiv \log \frac{1}{p_j} .$$

Definition 5

Let E be a collection of elements $E_1, E_2, \ldots, E_n$ on which an experiment is performed. Let p_j be the probability that E_j is selected. Then $(p_1, p_j, \ldots, p_n) = (p_j)$ is a probability distribution. The entropy $H(p_j)$ of the probability distribution is defined to be the expected value of the message which states the outcome of the experiment. The information content of the message stating that the outcome is E_j is just $\log \frac{1}{p_j}$. But this occurs with probability p_j. Hence, the expected value, i.e., the entropy, is given by

$$H(p_j) = H(p_1, p_2, \ldots, p_n) \equiv \sum_{j=1}^{n} p_j \log \frac{1}{p_j} .$$

Theorem 1

As p approaches zero, the expression of $p \log \frac{1}{p}$ approaches zero.

Proof

Assume, without loss of generality, that the logarithm is Napierian. Since $p \log \frac{1}{p} = -p \log p$, it suffices to show that $\frac{\log p}{(\frac{1}{p})}$ approaches zero as p approaches zero. By L'Hopital's Rule, it is only necessary to establish this limit for the expression $\frac{p^{-1}}{-p^{-2}} = -p$. Clearly, this is the case and the theorem is proved.

Theorem 2

The maximum value of $H(p_1, \ldots, p_n)$ is equal to log n, and this maximum occurs when and only when $p_j = \frac{1}{n}$ for each $j=1, 2, \ldots, n$.

Proof

It is equivalent to minimize the function

$-H = \sum_{j=1}^{n} p_j \log p_j$ subject to the constraints

$\sum_{j=1}^{n} p_j = 1$ and $p_j \geqslant 0$ for $i=1, 2, \ldots, n$.

The technique of Lagrange multipliers requires the expression

$$\sum_{j=1}^{n} p_j \log p_j - u\left(\sum_{j=1}^{n} p_j - 1\right)$$

to be differentiated with respect to p_j and the result set equal to zero. The quantity u is, of course, a Lagrange multiplier. Without loss of generality, the logarithm may be considered Napierian and the result is

$$1 = \log p_j - u = 0 \quad , \text{ or } \quad \log p_j = u - 1 \quad .$$

This must be true for each j and hence p_j is independent of j. Thus all p_j's are equal and so it must be the case that

$$p_j = \frac{1}{n} \quad \text{for each } j = 1, 2, \ldots, n \ .$$

From this, it follows that

$$\sum_{j=1}^{n} p_j \log \frac{1}{p_j} = \log n \quad .$$

Since the domain of the entropy function H is a closed set and H is bounded, the existence of both a maximum and a minimum is assured. A check reveals that points on the boundary cannot be minimums. Finally, since at all points within the domain the partial derivatives exist and are continuous, the proof is complete.

The Entropy Decomposition Theorem

In Chapter II it was stated that discipline aggregation resulted in a loss of program information and that this loss is proportional to the portion aggregated and the "within group" entropy of the portion aggregated. This fact is just a special case of the Entropy Decomposition Theorem.

Let $E = (E_1, E_2, \ldots, E_n)$ be a set of events with probability distribution $(p_1, p_2, \ldots, p_n)$. Let $S = (S_1, S_2, \ldots, S_k)$ be a partition of E where $1 \leq k \leq n$. Let J_i be the set of all j such that p_j is contained in S_i.

Define $P_i \equiv \sum_{j \in J_i} p_j$.

Then, the set of all $\frac{p_j}{P_i}$ such that $j \in J_i$ is a probability distribution for each $i = 1, 2, \ldots, k$.

Define $H(S_i) \equiv \sum_{j \in J_i} \frac{p_j}{P_i} \log \frac{p_j}{P_i}$ for each $i = 1, 2, \ldots, k$.

Then $(P_1, P_2, \ldots, P_k)$ is a probability distribution and

$$H(P_1, P_2, \ldots, P_k) = - \sum_{i=1}^{k} P_i \log P_i \quad .$$

<u>Theorem 3 (Entropy Decomposition)</u>

$$H(p_1, \ldots, p_n) = H(P_1, P_2, \ldots, P_k) + \sum_{i=1}^{k} P_i H(S_i)$$

Proof

$$= -\sum_{j=1}^{n} p_j \log p_j = -\sum_{i=1}^{k} \sum_{j \in J_i} p_j \log p_j$$

$$= \sum_{i=1}^{k} P_i \sum_{j \in J_i} \frac{p_j}{P_i} \left(\log \frac{1}{P_i} + \log \frac{P_i}{p_j} \right)$$

$$= \sum_{i=1}^{k} P_i \left(\sum_{j \in J_i} \frac{p_j}{P_i} \log \frac{1}{P_i} + \sum_{j \in J_i} \frac{p_j}{P_i} \log \frac{P_i}{p_j} \right)$$

$$= \sum_{i=1}^{k} \left(P_i \sum_{j \in J_i} \frac{p_j}{P_i} \log \frac{1}{P_i} + P_i \sum_{j \in J_i} \frac{p_j}{P_i} \log \frac{P_i}{p_j} \right)$$

$$= \sum_{i=1}^{k} P_i \log \frac{1}{P_i} \left(\sum_{j \in J_i} \frac{p_j}{P_i} \right) + \sum_{i=1}^{k} P_i \sum_{j \in J_i} \frac{p_j}{P_i} \log \frac{P_i}{p_j} .$$

But $\sum_{j \in J_i} \frac{p_j}{P_i} = \frac{1}{P_i} \sum_{j \in J_i} p_j = 1$. Hence $-\sum_{j=1}^{n} p_j \log p_j$

$$= \sum_{i=1}^{k} P_i \log \frac{1}{P_i} + \sum_{i=1}^{k} P_i \sum_{j \in J_i} \frac{p_j}{P_i} \log \frac{P_i}{p_j}$$

$$= H(P_1, P_2, \ldots, P_k) + \sum_{i=1}^{k} P_i H(S_i) .$$

That is,

$$H(p_1, p_2, \ldots, p_n) = H(P_1, P_2, \ldots, P_k) + \sum_{i=1}^{k} P_i H(S_i)$$

which was to be proved. The entropy decomposition theorem thus states that the entropy "before aggregation" is the sum of the "between group" entropy and the average "within group" entropy.

Note that the change in entropy due to aggregation can be expressed as

$$H(p_1, p_2, \ldots, p_n) - H(P_1, P_2, \ldots, P_k)$$

$$= \sum_{i=1}^{k} P_i \; H(S_i) \quad . \quad P_i \gtrless 0 \text{ and } H(S_i) \gtrless 0 \text{ for each}$$

$i = 1, 2, \ldots, k$. It follows that the entropy after aggregation is never greater than that of the original distribution. In fact, the existence of even one i such that $H(S_i) > 0$ will force the entropy change to be positive, and aggregation will result in an information loss.

Theorem 4

Let (p_i) be given such that $p_i \geq 0$ for each i, $\sum_{i=1}^{n} p_i = 1$, and $i = 1, 2, \ldots, n$. Define $C \equiv \dfrac{1}{\log n} \sum_{i=1}^{n} p_i \log \dfrac{1}{p_i}$. Then the following statements hold:

(i) If for exactly one i, $p_i = 1$, then $C = 0$.

(ii) If for each i, $p_i = 1/n$, then $C = 1$.

(iii) If $p_i \neq 1$ for any i but $p_i \neq 1/n$ for some i, then $0 < C < 1$.

(iv) C is a continuous function.

Proof

(i) Without loss of generality suppose that $p_i = 1$. Then $p_i = 0$ for each $i \neq 1$ and for such i, $p_i \log \frac{1}{p_i}$ has been defined to be zero. Hence, $C = \frac{1}{\log n} (1 \log \frac{1}{1}) = 0$.

(ii) Suppose $p_i = 1/n$ for each $i = 1, 2, \ldots, n$. Then

$$C = \frac{1}{\log n} \left(\frac{1}{n} \log n + \ldots + \frac{1}{n} \log n \right)$$ where the expression in paranthesis contains n terms. Thus,

$$C = \frac{1}{\log n} (\log n) = 1.$$

(iii) Since each term of C is non-negative, $C = 0$ only if each term is zero. Now $p_i \log \frac{1}{p_i} = 0$ only if $p_i = 0$ or $p_i = 1$. But $p_i \neq 1$ for any i. Hence, $p_i > 0$ for some i, say, for $i = k$. Then $p_k \log \frac{1}{p_k}$ 0. By Theorem 2 the maximum value of $\sum_{i=1}^{n} p_i \log \frac{1}{p_i}$ is $\log n$ and this occurs only when each $p_i = 1/n$. But $p_i \neq 1/n$ for some i, hence, $C < 1$. Thus $0 < C < 1$.

(iv) C is a continuous function, being a linear combination of continuous functions.

Theorem 5

$$M(D, P) = H(D) + H(P) - H(D, P)$$

Proof

$$M(D,P) \equiv \sum_i \sum_j r_{ij} \log \frac{r_{ij}}{r_{i.}\, r_{.j}}$$

$$= \sum_i \sum_j r_{ij} \log r_{ij} - \sum_i \sum_j r_{ij} \log r_{i.} - \sum_j \sum_i r_{ij} \log r_{.j}$$

$$= \sum_i \sum_j r_{ij} \log r_{ij} - \sum_i r_{i.} \log r_{i.} - \sum_j r_{.j} \log r_{.j}$$

$$= -H(D,P) + H(D) + H(P)$$

Theorem 6

The mutual entropy $M_A(D,P)$ of any matrix of the form

$$A = \begin{bmatrix} d_1 & x \\ y & d_2 \end{bmatrix}$$

is less than the mutual entropy $M_B(D,P)$ of

$$B = \begin{bmatrix} d_1 + x & 0 \\ 0 & d_2 + y \end{bmatrix}$$ where

d_1, d_2, x, $y > 0$.

Proof

Let $M = d_1 + d_2 + x + y$. By Theorem 5, $M_A(D,P)$ $= H_A(D) + H_A(P) - H_A(D,P)$. Thus,

$$M_A(D,P) = \frac{d_1 + x}{M} \log \frac{M}{d_1 + x} + \frac{d_2 + y}{M} \log \frac{M}{d_2 + y} +$$

$$\frac{d_1+y}{M} \log \frac{M}{d_1+y} + \frac{d_2+x}{M} \log \frac{M}{d_2+x}$$

$$- \frac{d_1}{M} \log \frac{M}{d_1} - \frac{d_2}{M} \log \frac{M}{d_2} - \frac{x}{M} \log \frac{M}{x} - \frac{y}{M} \log \frac{M}{y}$$

$$= \log M - \frac{d_1+x}{M} \log (d_1+x) - \frac{d_2+y}{M} \log (d_2+y)$$

$$- \frac{d_1+y}{M} \log (d_1+y) - \frac{d_2+x}{M} \log (d_2+x)$$

$$+ \frac{d_1}{M} \log d_1 + \frac{d_2}{M} \log d_2 + \frac{x}{M} \log x + \frac{y}{M} \log y.$$

Replacing d_1 with d_1+x, x with 0, y with 0, and d_2 with d_2+y yields,

$$M_B(D, P) = \log M - \frac{d_1+x}{M} \log (d_1+x) - \frac{d_2+y}{M} \log (d_2+y)$$

Hence, $M_B(D, P) - M_A(D, P)$

$$= \frac{d_1+y}{M} \log (d_1+y) + \frac{d_2+x}{M} \log (d_2+x)$$

$$- \left(\frac{d_1}{M} \log d_1 + \frac{d_2}{M} \log d_2 + \frac{x}{M} \log x + \frac{y}{M} \log y\right)$$

$$= \left[\frac{d_1}{M} \log (d_1+y) - \frac{d_1}{M} \log d_1\right] + \left[\frac{y}{M} \log (d_1+y) - \frac{y}{M} \log y\right] +$$

$$\left[\frac{d_2}{M} \log (d_2+x) - \frac{d_2}{M} \log d_2\right] + \left[\frac{x}{M} \log (d_2+x) - \frac{x}{M} \log x.\right]$$

But each term in brackets is positive since the logarithm is an increasing function. Therefore, $M_B(D, P) > M_A(D, P)$.

Theorem 7

The mutual entropy of any non diagonal matrix with positive diagonal elements is less than the mutual entropy of some comparable diagonal matrix with the same sum.

Proof

Apply the methodology of Theorem 6 to obtain the result.

Theorem 8

The mutual entropy of any diagonal matrix A is not greater than a comparable diagonal matrix B having all diagonal elements equal.

Proof

Let M be the sum of the elements of A. Then,

$$M_A(D, P) = 2\,H\left(\frac{d_1}{M}, \ldots, \frac{d_n}{M}\right)$$

$$= H\left(\frac{d_1}{M}, \ldots, \frac{d_n}{M}\right)$$

$$= H\left(\frac{d_1}{M}, \ldots, \frac{d_n}{M}\right) \log n$$

where n is the number of rows of the square matrix A. Also, $M_B(D, P) = H\left(\frac{d}{M}, \ldots, \frac{d}{M}\right) = \log n$ where d is an element of the diagonal of B. Hence, $M_A(D, P) \leq M_B(D, P)$.

The vita has been removed from
the scanned document

ENTROPY MEASURES ON THE INDUCED

COURSE LOAD MATRIX

by

Merit William Hudgins

(ABSTRACT)

Matrices have been developed by the National Center for Higher Education Management Systems at Western Interstate Commission for Higher Education that illustrate and quantify the relationships between and among credit hours produced (in courses, departments, disciplines, and other aggregations) and the degree programs (lower division, upper division, and graduate students in various majors) that consume the credits. A matter of interest is the extent to which course offerings (herein referred to as disciplines) complement and are complemented by the various student programs.

A numerical measure of complementation among disciplines and programs has been developed. This measure, based on the entropy concept of information theory, is always a number between zero and one. Applied to a discipline, the measure expresses the extent to which that discipline's credit hour output is either

monopolized by a subset of programs or consumed evenly by all programs. Similarly, applied to a program the measure expresses the extent to which it is interdisciplinary in content. Hence, the measure provides an expression of monopoly or equilibrium on a scale from zero to one, and an expression of interdisciplinarity, also on a scale of zero to one. This permits both disciplines and programs (at whatever levels of aggregation) to be quantified and ranked, for purposes of description, analysis, projection, goal-setting, programming, budgeting, and other processes of institutional management.

In addition, certain global measures are developed which express the complementation between all disciplines, between all programs, and between all disciplines and programs. These are measures of curriculum integration and can be applied to a single institution or portion thereof at a single point in time; they can be applied through time. Under certain conditions and subject to various precautions, they can be applied to produce interinsitutional comparisons.

The development of these measures is presented at length, and their application is illustrated at length. Potential uses and appropriate lines of related inquiry are enumerated.