DESIGNS OF EXPERIMENTS IN INFORMATION RETRIEVAL

Donald W. King, Westat Research Analysts, Inc.

You have heard some of the inherent difficulties in the patent examining process. One of the major sources of difficulties is the novelty search which the patent examiner is charged to make. The R & D division of the U. S. Patent Office is committed to research, design and development of information storage and retrieval systems to provide the examiner with a workable search tool.

I am here with two purposes in mind. The first is to introduce the field of information storage and retrieval to members of the statistical profession as there has been a distinct lack of the use of statistical techniques in the research in this area. Secondly, I will discuss some experimental design techniques used in the research and development of such systems.

I will, first, briefly describe an information retrieval system. The system involves aggregating a set of books, publications, documents, etc., into a common library or file. An entire document can be put physically into this file or any portion of the document may be extracted and incorporated into a token or artificial file. The token file might include an abstract, classification, descriptions, facets, titles, authors or bibliographies from the document. A searcher may then look over a list of these artifices and request of the file all documents containing one or more artifices of interest to him or which he feels will most likely provide documents of interest to him. The problem then is to describe the documents in a token manner which will provide the examiner with a subset of documents which may be expected to satisfy his search needs. There are many combinations of possible systems. It is necessary, then, to seek out an "optimum" system from the standpoint of various trade-off parameters related to cost, reliability or system quality, and time. Time is distinguished from cost purposely.

The Patent Office is approaching the evaluation of examiner information storage and retrieval in three broad areas:

(1) Determination of the system user's needs.
(2) Determination of the system(s) which will "optimally" fulfill these needs.
(3) Determination of the effect of a new or modified information retrieval system on the entire examining process.

The research in the first area is currently limited to depth interviews and a questionaire survey in potential areas for examiner search aids. The second area involves advanced research into information storage and retrieval from the standpoints of computer software and hardware. Much of this research relates to advancement of the state of the art. In addition, much work is being done to determine the "optimum" systems under the present state of the art. Here, there are many problems to be solved such as defining meaningful file characteristics; establishing the most efficient general approach such as coordinate indices, classification, associative indices, etc.; and determining the best means of preparing the file.

One element of the third area has been discussed. This is the stochastic model describing the flow of applications through the system. A method of evaluating the quality of patent examining has been developed and is in use throughout the office. Some standard quality control practices are being employed in this effort. Another phase of this area involves investigation of the costing aspect of an information storage and retrieval system.* The developmental research for the information retrieved systems in the Patent Office involve a long dependent series of carefully conceived and designed experiments.

Today, I am going to discuss research involving the indexing phase of file preparation. Although research has begun in all of the areas mentioned above, it is further along in this area. The statistical techniques described are well known but are rarely used in research involving information retrieval systems.

Basically, this segment of research is devoted to measuring characteristics important to indexing. These are measures of:

(1) accuracy of indexing
(2) consistency of indexing
(3) time required to index a document.

The most important of these, and probably the most difficult to measure, is accuracy of indexing. Numerous attempts have been made by others to evaluate accuracy of indexing. The real point of departure of this research from other studies is the measurement of the effect of indexing errors in the retrieval of documents. This places the emphasis on the behavior of the system rather than on more philosophic grounds that errors are bad and therefore should be avoided.

*David A. Belsley, "The costing of information retrieval systems in the Patent Office through the application of a generalized costing structure," U. S. Department of Commerce, Patent Office, September 1962 (unpublished report).

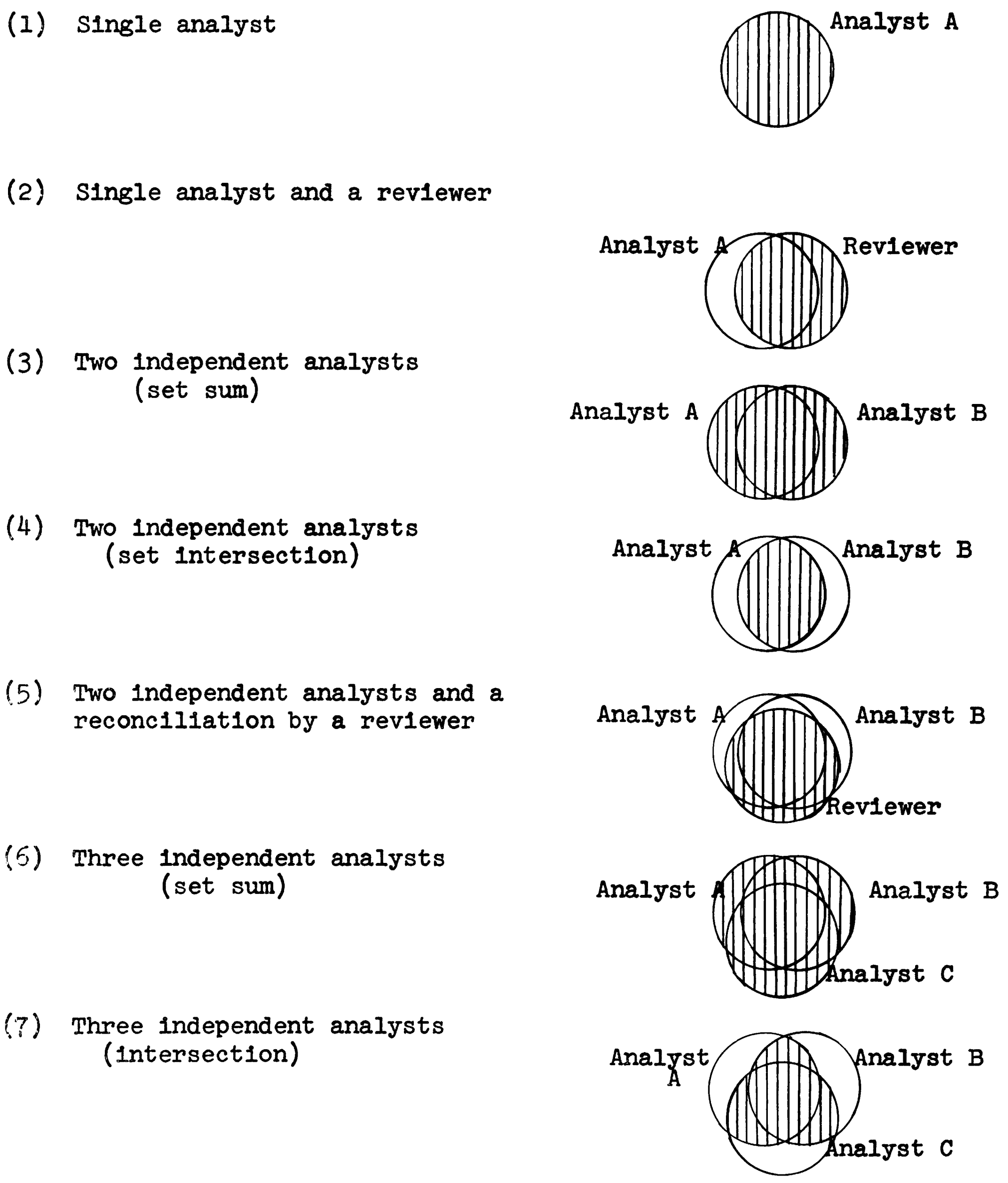

(1) Single analyst
Analyst A
(2) Single analyst and a reviewer
Analyst A
Reviewer
(3) Two independent analysts (set sum)
Analyst A
Analyst B
(4) Two independent analysts (set intersection)
Analyst A
Analyst B
(5) Two independent analysts and a reconciliation by a reviewer
Analyst A
Analyst B
Reviewer
(6) Three independent analysts (set sum)
Analyst A
Analyst B
Analyst C
(7) Three independent analysts (intersection)
Analyst A
Analyst B
Analyst C

Figure 1. Venn diagrams of codes selected using various analyst modes.

Analyst Experience

Reviewer Experience	Over Three Months in Organometallics Art	Less than Three Months in Organometallics Art
High	$\alpha^{1}_{1(1).23(2)}$ $\alpha^{2}_{2(1).13(2)}$ $\alpha^{3}_{3(1).12(2)}$ $\alpha^{7}_{1(2).23(1)}$ $\alpha^{8}_{2(2).13(1)}$ $\alpha^{9}_{3(2).12(1)}$	$\alpha^{4}_{4(1).56(2)}$ $\alpha^{5}_{5(1).46(2)}$ $\alpha^{6}_{6(1).45(2)}$ $\alpha^{10}_{4(2).56(1)}$ $\alpha^{11}_{5(2).46(1)}$ $\alpha^{12}_{6(2).45(1)}$
Low	$\alpha^{13}_{7(3).89(4)}$ $\alpha^{14}_{8(3).79(4)}$ $\alpha^{15}_{9(3).78(4)}$ $\alpha^{19}_{7(4).89(3)}$ $\alpha^{20}_{8(4).79(3)}$ $\alpha^{21}_{9(4).78(3)}$	$\alpha^{16}_{10(3).11\ 12(4)}$ $\alpha^{17}_{11(3).10\ 12(4)}$ $\alpha^{18}_{12(3).10\ 11(4)}$ $\alpha^{22}_{10(4).11\ 12(3)}$ $\alpha^{23}_{11(4).10\ 12(3)}$ $\alpha^{24}_{12(4).10\ 11(3)}$

$\alpha^{h}_{i(j).ii(j)}$ Experimental arrangement of documents; analysts, and reviewers.

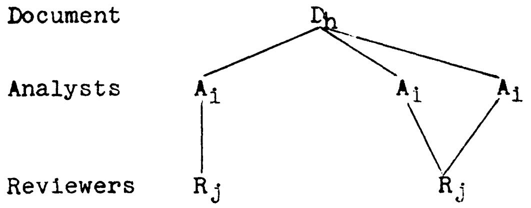
Figure 2. The experimental arrangement of the documents, analysts and reviewers for the organometallics intensive indexing experiment.

Indexing is defined as reducing information in the patent documents to a set of common identifiers and translating this set into unique codes for standard mechanized processing. There are two broad categories of indexing errors. Failure to select a code that should have been selected will be referred to as a Type I error, while selecting a code which should not have been selected is referred to as an error of Type II. Errors can also arise in the translation of information concepts into numerical codes. These, however, all result in errors of Type I or Type II.

We can assume the existence of conditional nrobabilities in a two-by-two contingency table with the actual selection of a code and whether the code should have been selected as the two bases for classification. A conditional probability that a code is not selected given that it should be corresponds to the Type I error and the conditional probability that a code is selected given that it should not be corresponds to the Type II error.

Some elementary mathematical models were derived to provide a means of estimating the correct retrieval and the false retrieval considering the conditional probabilities of indexing errors, the number of codes used in the searrh question and the logical relationship of the codes used in the search question.* The proportion of correct retrieval is highly sensitive to the conditional probability of selecting a code given that it should be selected, the number of codes used, and the use of search questions relating the codes conjunctively. Part of the purpose of the research reported here was to verify the simple model experimentally. The close correspondence between hypothetical and experimental data was quite revealing to the information retrieval staff. This was considered to be particularly important since many of the files in the Patent Office involve chemical compounds and coordinate index systems involving conjunctive descriptions of compound fragments.

The most important assumption of the models described was that the conditional probabilities of the codes used in a search question were independent. An experiment was designed to investigate this assumption to estimate the model parameters and other attributes of interest, and to determine the most efficient way of utilizing the indexing personnel. This experiment was conducted on 24 patent documents chosen randomly from the organometallic file containing a total of 3625 documents. The characteristics of interest were the conditional probabilities mentioned previously, measures of consistency of indexing, and estimates of the time required to index a patent document. The characteristics were observed for seven analyst modes. These are described pictorially in Figure 1 by Venn diagrams. The basic experimental arrangement is given in Figure 2. The order of the ith analyst and jth reviewer is given by their appearance in the diagram below:

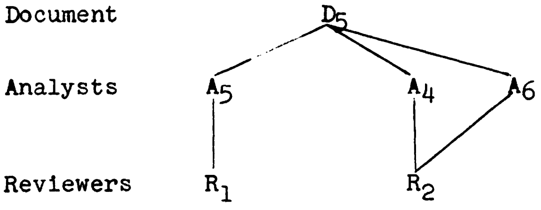


where: D_h, h=1,...,24, denotes documents.

A_i, i=1,...,12, denotes analysts

R_j, j=1,..., 4, denotes reviewers.

For example, the arrangement



appears in Figure 2 as

$$\alpha^{5}_{5(1).46(2)}$$

The primary consideration motivating this design is that an analyst or reviewer cannot repeat his efforts on a given document because of the learning effect.

The design is quite flexible in that all seven analyst modes can be observed, although only five operations are performed on each document. This can be done since the individual analyst's work can be observed prior to review. The single-double-and triple- analyst modes can therefore be generated accordingly. The results of the indexing can be incorporated into the file if the experiment is conducted as a file is being prepared.

Note in the arrangement above that the design provides a means for evaluating analyst experience and reviewer experience. Experience was defined as at least three months in indexing organometallic

*For a more complete discussion of the model, see E. C. Bryant, D. W. King and P. J. Terragno, "Some technical notes on coding errors," WRA PO 7, July, 1963 (informal report to the Office of Research and Development, U. S. Patent Office) documents. All of the analysts experience analyzing chemical compounds in one file or another. The comparison could just as easily have been made with regard to experience indexing, educational background, age, or sex, depending on the hypotheses under investigation. When a file is first being indexed all of the analysts will be inexperienced in the particular art, therefore, this evaluation provides an indication as to what may be lost in the initial phases of indexing that file and what improvement may be expected as the analysts gain experience in the new art.

It is clear that a distinct experimental arrangement is suggested for analysis of each analyst mode. The experimental designs, analysis of variance with expected mean squares, and observed values are given in an informal report to the Office of Research and Development of the U. S. Patent Office.* One difficulty in statistical analysis was that a larger document-to-document variation than was anticipated reduced the sensitivity of the statistical tests. An improved experimental design resulted ** and is used in a similar experiment involving an electrical transistor file.

A summary of the estimates of the conditional probability that a code is selected, given that it should be; the conditional probability that a code is selected, given that it should not be; the total number of codes selected; and total time to index the documents is given in Tables 1 through 4.

An experiment was conducted:

(1) to test the parameters, assumptions and validity of error retrieval models described previously,

(2) to determine the effect of indexing errors on retrievals for three analyst modes, and

(3) to determine if synthetic search questions can be used in evaluating a file

This experiment involved preparing a mechanized token file from 201 organometallic patent documents chosen randomly. Essentially three files were prepared by three analyst modes; i.e., the single-analyst mode, double-analyst (set-sum) mode, and single-analyst-reviewed mode. These three files were searched simultaneously using search questions formulated by examiners using the system in the past. Synthetic questions formulated by analysts were also used. Table 5 gives a summary of the average proportion of correct retrieval for the three analyst modes. These values are plotted by triangles in Figure 3. The model estimates of the proportion of correct retrieval found from the 24 document intensive indexing experiment are plotted linearly on the semi-log scale. The observed average proportion of correct retrieval and model estimates yield similar results for both the double-analyst (set-sum) mode and the single-analyst reviewed mode.

The model assumes that the probability of indexing two or more codes incorrectly is independent for all codes, i.e. $P_3(T_jT_k) = P_3(T_j)\ P_3(T_k)$: This can best be explained by the fact that the terms represent a portion of an entire compound. A further investigation of the errors made by the single analysts reveals that a large number of the errors involve omission of the entire compound rather than merely indexing one fragment of the compound incorrectly.

Joint probabilities of $P_3(T_j,T_k,...,T_r)$ were estimated for one through four codes from the 24 document intensive indexing experiment. Observations of combinations of more than four codes were toorare to be used for this estimation. The scale of the "Average of original and repeated indexings" in Figure 3 gives the average observed proportion of correct retrieval (triangles), plotted least squares estimates of these values (linear plot), and the estimates of the joint probabilities for one through four terms mentioned above (crosses). It is seen that the joint probabilities are very close to those found using a least squares estimate of the observations. This demonstrates that it is quite important to test the assumptions of independence, particularly if there is reason to believe that the codes may be highly related, as they are when the indexing system involve fragments of compounds.

The estimates of the proportion of retrieval using the single-analyst-reviewed mode and the double-analyst (set-sum) mode are apparently representative of the actual observed retrievals. The assumption of independence becomes far less critical in these instances since a second person's review or independent analysis is involved and tends to cancel out the effect of the dependency.

Figure 4 also gives the proportion of correct retrieval over the range of the number of codes per search question for the original indexing and repeated indexing for the single-analyst mode. It is again noted that these plotted lines are least squares lines and are not derived from the model $Y=p_3^r$. The very small difference in these independent indexings

*E. C. Bryant, D. W. King, and P. J. Terragno, "Analysis of an indexing and retrieval experiment for the organometallic file of the U.S. Patent Office", WRA PO 10, August, 1963 (informal report to the Office of Research and Development, U. S. Patent Office)

**E. C. Bryant, D. W. King and P. J. Terragno, "Revised design for coding experiment, 307/88.5 file, "WRA PO 9, June 1963 (informal report to the Office of Research and Development, U.S.Patent Office.

Table 1. Estimates of p_3* for the various analyst modes with 95 per cent confidence limits (in parentheses)**

Analyst Mode	Experienced Analysts	Inexperienced Analysts	Combined Analysts
	—	—	
Single-analyst	(.86-.92)	(.78-.86)	.86
Double-analyst	—	—	
Set Sum	(.96-.99)	(.89-.93)	.95
Intersection	(.77-.83)	(.66-.73)	.75
Triple-analyst	—	—	
Set Sum	(.99-1.00)	(.95-.98)	.98
Intersection	(.88-.97)	(.75-.88)	.88
Single-analyst-reviewed	—	—	
Experienced reviewers	(.88-.97)	(.92-.99)	.94
Inexperienced reviewers	(.93-.99)	(.82-.93)	.93
Combined —			.94
Double-analyst-reviewed	—	—	
Experienced reviewers	(.94-.99)	(.99-1.00)	.98
Inexperienced reviewers	(.92-.98)	(.81-.91)	.91
Combined			.95

*p_3 - is the conditional probability that a code will be selected, given that it should be

** Standard errors, on which the confidence limits are based, contain variation due to differences in documents and analysts.

Table 2. Estimates of the probability that a code will be selected, given that it should not be, p_2, for the various analyst modes.

Analyst Mode	Experienced Analysts	Inexperienced Analysts	Combined Analyst
Single-analyst	(.0002-.0007)	(.0024-.0036)	.0014
Double-analyst	—	—	
Set Sum	(.0008-.0013)	(.0011-.0016)	.0033
Intersection	(.0000-.0001)	(.0001-.0003)	.0001
Triple-analyst	—	—	
Set Sum	(.0008-.0037)	(.0067-.0113)	.0051
Intersection	(.0000-.0001)	(.0003-.0007)	.0002
Single-analyst-reviewed		—	
Experienced reviewers	(.0000-.00005)	(.0000-.0004)	.0001
Inexperienced reviewers	(.0000-.0004)	(.0007-.0029)	.0005
Combined			.0002
Double-analyst-reviewed	—		
Experienced reviewers	(.0003-.0022)	(.0011-.0041)	.0016
Inexperienced reviewers	(.0000-.0004)	(.0013-.0045)	.0008
Combined			.0012

Table 3. Estimates of the total number of codes selected for the various analyst modes.

Analyst Mode	Experienced Analysts	Inexperienced Analysts	Combined Analysts
—		—	
Single-analyst	(132.4-136.9)	(136.7-141.2)	136.8
Double-analyst	—	—	—
Set Sum	(146.1-152.4)	(159.8-166.0)	159.0
Intersection	(111.1-117.5)	(106.2-112.6)	111.8
Triple-analyst	—	—	
Set Sum	(94.3-227.7)	(120.6-254.0)	174.2
Intersectien	(77.9-191.9)	(80.7-194.7)	136.3
Single-analyst-reviewed	—	—	
Experienced reviewers	(61.7-248.3)	(24.5-211.1)	136.4
Inexperienced reviewers	(41.7-228.3)	(74.2-260.8)	151.2
Combined –			143.8
Double-analyst-reviewed	-	—	
Experienced reviewers	(56.1-250.9)	(35.5-230.4)	143.2
Inexperienced reviewers	(41.3-236.1)	(71.4-266.2)	153.8
Combined			148.5

Table 4. Estimates of total time (in minutes) required to index the documents for various analyst modes.

Analyst Mode	Experienced Analysts	Inexperienced Analysts	Combined Analysts
—	—	--	
Single-analyst	(40.6-54.1)	(75.0-88.5)	64.6
Double-analyst	—	—	
Set Sum	(87.1-100.9)	(156.1-169.9)	128.6
Intersection			
Triple-analyst	-	—	
Set Sum	(43.4-238.9)	(147.6-343.1)	193.2
Intersectien			
Single-analyst-reviewed	—	—	
Experienced reviewers	(0-201.9)	(0-215.7)	96.3
Inexperienced reviewers	(0-190.7)	(62.9-287.7)	126.8
Combined –			111.6
Double-analyst-reviewed	—	—	
Experienced reviewers	(0-330.9)	(0-365.0)	156.2
Inexperienced reviewers	(0-325.4)	(96.1-479.5)	210.8
Combined			183.5

Table 5 A summary of the average proportion of correct retrieval in the searches of the 201 sample file for three-analyst modes.

No. of codes per search question	No. of searches	Ave. proportion of correct retrieval			
		single-analyst mode	single-analyst-rev. mode	double-analyst mode	single-analyst (joint probabilities*)
1	0	-	-	-	.90
2	3	.80	1.00	1.00	.89
3	12	.76	.82	.95	.85
4	31	.84	.85	.96	.80
5	29	.71	.79	.90	
6	36	.68	.78	.96	
7	54	.66	.79	.88	
8	30	.59	.79	.90	
9	46	.60	.76	.89	
10	35	.50	.59	1.00	
11	44	.52	.54	.79	
12	36	.47	.29	.83	

*The joint probabilities were estimated from the proportion of missed correct codes observed jointly two, three and four codes at a time.

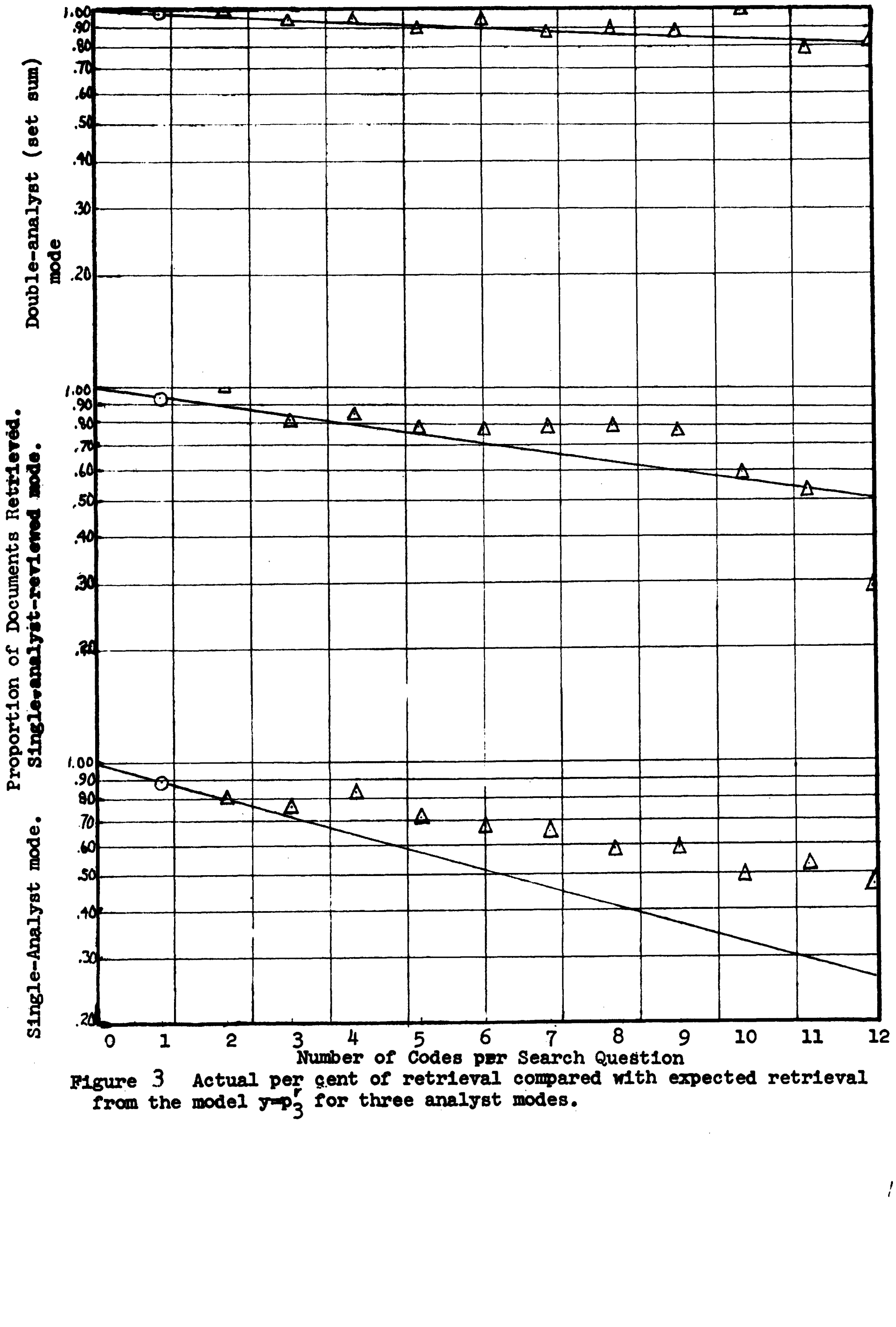


Figure 3 Actual per cent of retrieval compared with expected retrieval from the model $y=p_3^r$ for three analyst modes.

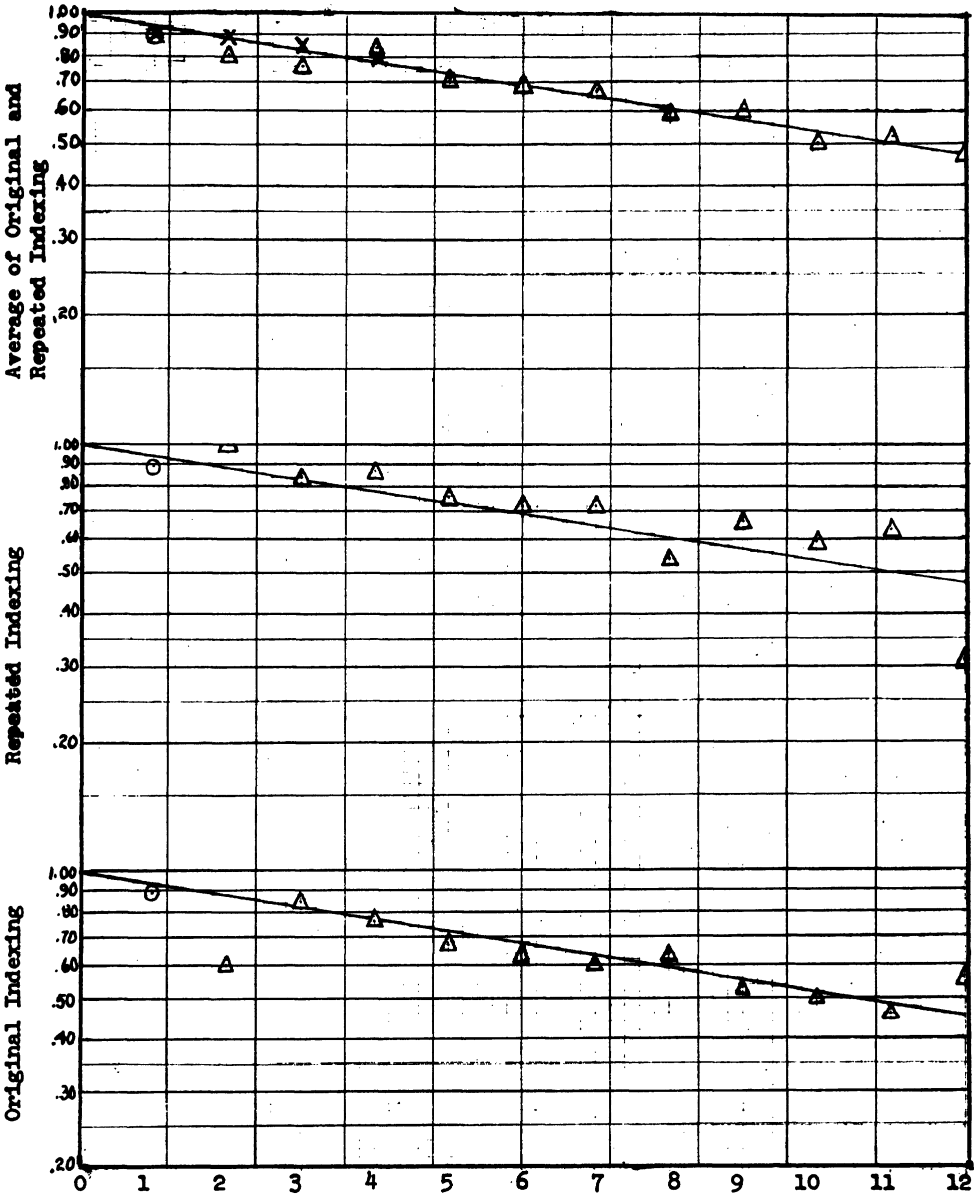


Figure 4 Actual per cent of retrieval and fitted least squares estimate for the original indexing, repeated indexing, and average indexing for the single-analyst mode.

indicates that indexing under experimental conditions (as was done in the repeated indexing) did not significantly affect the experimental findings. The behavior of the data from the intensive indexing experiment also supports the contention that the analysts operated normally even though they were aware that they were being observed.

The average number of correct retrievals, the average number of missed documents and the false drops observed in the searches conducted on the sample file of 201 documents is summarized in Table 6.

The experimentation reported here has evaluated various indexing modes in terms of indexing time, errors in the search file, and errors in retrieval. We have, then, bases for making rational decisions concerning the utilization of indexes so as to accomplish specific objectives. For three indexing modes used in the retrieval experiment, the "tradeoff parameters" are estimated as follows:

	Single-analyst	Double-analyst	Single-analyst-reviewed
Indexing time (Min)	64.3	128.6	111.6
Missed docs. per search	9.6	2.5	6.7
False retr. per search	4.5	9.2	5.6

With a given number of documents to be indexed, a given number of searches to be conducted per year, and some rough indication of the relative importance of missed documents and false retrievals one can determine the analyst mode which fits his particular needs.

Under the current Patent Office examining system the number of missed documents is so important as to lead one to choose the double-analyst (set sum) mode. Furthermore, this permits an easy evaluation of consistency of indexing, a parameter which can easily be controlled by usual methods of industrial quality control.

The importance of accurate coding in preventing missed documents has been demonstrated. Unfortunately it is difficult to assess accuracy of coding because of differences in opinion concerning what should be coded. Consistency of coding is relatively easy to measure, however, and, intuitively, it seems that consistency and accuracy should be closely related. It is clear that high accuracy implies a high degree of consistency between two coders, but that a high degree of consistency will only imply high accuracy if biases are not present.

Within the restriction that the so-called "correct" coding actually may not be correct, one can estimate the conditional probability, p_3, (that a term which should be coded will be coded).

Let $a_{ijk} = 1$ if, in the i^{th} document, term T_j was coded correctly by the k^{th} analyst. Let $a_{ijk} = 0$ otherwise. Let $b_{ij} = 1$ if the j^{th} term should have been coded in the i^{th} document, and let $b_{ij} =$ 0 otherwise. Then,

$$\hat{p}_3^{ik} = \sum_j a_{ijk} \Big/ \sum_j b_{ij} \qquad (1)$$

is an estimate of p_3 averaged over all terms in the i^{th} document and for the k^{th} analyst. Similarly, one can define

$$\hat{p}_3^{i} = \sum_{jk} a_{ijk} \Big/ k\sum_j b_{ij} \qquad (2)$$

where k stands for the number of analysts, as well as an index designator for them. This provides an estimate for the given document. For a given term one can construct the estimate

$$\hat{p}_3^{j} = \sum_{ik} a_{ijk} \Big/ k\sum_i b_{ij} \qquad (3)$$

and so on. An overall estimate is provided by

$$\hat{p}_3 = \sum_{ijk} a_{ijk} \Big/ k\sum_{ij} b_{ij} \qquad (4)$$

We are concerned here primarily with estimates similar to (3) and (4) above.

There are various approaches to estimates of consistency. Consider the following notation for two analysts:

C_{00}^{ij} = 1 if neither analyst coded j^{th} term in the i^{th} document.

=0otherwise

C_{10}^{ij} = 1 if the j^{th} term in the i^{th} document was coded by the first analyst, but not by the second.

= 0 otherwise

C_{01}^{ij} = 1 if the j^{th} term in the i^{th} document was coded by the second analyst, but not by the first

= 0 otherwise

C_{11}^{ij} = 1 if both analysts coded the j^{th} term in the i^{th} document

= 0 otherwise

Table 6 Average number of correct retrievals, missed documents and false drops per search for the 201 sample file for three analyst modes.

No. of codes asked	No. of Searches	Ave. no. of correct retrieval per search	Ave. no. of missed docs. per search			Ave. no. of false drops per Search		
			Single-analyst mode	Double-analyst mode	Single-analyst-rev. mode	Single-analyst mode	Double-analyst mode	Single-analyst-rev. mode.
2	3	1.67	.33	0	0	0	0	0
3	12	2.25	.54	.11	.42	.41	.83	.68
4	31	2.74	.45	.11	.42	.34	.65	.42
5	29	1.96	.57	.20	.41	.24	.62	.31
6	36	1.89	.61	.08	.42	.44	.81	.55
7	54	1.67	.57	.20	.35	.34	.68	.41
8	30	2.10	.87	.21	.43	.32	.70	.36
9	46	1.48	.60	.16	.35	.35	.61	.39
10	35	.63	.31	0	.26	.12	.17	.14
11	44	.93	.44	.19	.43	.08	.25	.11
12	36	.61	.40	.10	.30	.03	.06	.03
Ave. 7.9	32.4	1.54	.53	.14	.37	.25	.51	.31
95% Confidence Limits			.46-.61	.09-.19	.31-.43	.20-.31	.45-.57	.25-.37

This is shown as follows:

		1st Coder	
		No	Yes
2nd	No	c_{00}^{ij}	c_{10}^{ij}
Coder	Yes	c_{01}^{ij}	c_{11}^{ij}

Then, some measures of consistency which have been suggested are:*

$$cc^{i} = \sum_{j} c_{11}^{ij} \Big/ \sum_{j} \left[c_{11}^{ij} + c_{01}^{ij} + c_{10}^{ij}\right] \quad (5)$$

$$cc^{j} = \sum_{j} c_{11}^{ij} \Big/ \sum_{i} \left[c_{11}^{ij} + c_{01}^{ij} + c_{10}^{ij}\right] \quad (6)$$

$$cc = \sum_{ij} c_{11}^{ij} \Big/ \sum_{ij} \left[c_{11}^{ij} + c_{01}^{ij} + c_{10}^{ij}\right] \quad (7)$$

Some work done by the Census Bureau** on response differences is related to this problem. Here the emphasis is on an "index of inconsistency" which, for our purposes, we may define as follows:

$$II^{j} = \frac{n^{j}\, g^{j}}{c_{1}^{j}\,(n^{j} - c_{1}^{j}) + c_{2}^{j}\,(n^{j} - c_{2}^{j})} \quad (8)$$

where n_j = number of documents over which consistency of the j^{th} term is being observed.

$$g^{j} = \sum_{i} (c_{01}^{ij} + c_{10}^{ij}) \quad (9)$$

$$c_{1}^{j} = \sum_{i} (c_{10}^{ij} + c_{11}^{ij}) \quad (10)$$

$$c_{2}^{j} = \sum_{i} (c_{01}^{ij} + c_{11}^{ij}) \quad (11)$$

By summation over j one can obtain an inconsistency index over all terms. Since we are primarily interested in consistency, rather than inconsistency, we use 1 - II to compare with CC, above, and with estimates of p_3.

In the sample of 24 documents there were four independent codings of each document; the original coding and three codings in the experiment. Thus there are $\binom{4}{2}$ = 6 pairs of codings to be considered. While it is possible to define consistency in terms of the number of agreements among 3 coders (or 4) there is some merit in reducing all such measures to a two-coder basis. This was done by dividing the average (over 6 pairs) of the number of documents in which the given term was coded by two analysts by the average number of times it was coded by either. Thus, for the j^{th} term

$$cc^{j} = \sum_{\alpha}\sum_{j} c_{11\alpha}^{ij} \Big/ \sum_{\alpha}\sum_{j} \left[c_{11\alpha}^{ij} + c_{01\alpha}^{ij} + c_{10\alpha}^{ij}\right]$$

where indicates any of the six combinations of two analysts. A similar averaging process was used for the index of inconsistency.

As can be seen the number of times a term was coded by both members of a pair of analysts serves as the numerator for a coefficient of consistency. As a computational convenience, this variable is averaged over all possible pairs of analysts. An interesting statistical question arises concerning the variance of such an average.

Suppose that two independent coders encode a sample of n documents. We focus our attention on a particular term, T_j, and record the following:

m_{00} = the number of times coded by neither coder
m_{11} = the number of times coded by both coders
m_{10} = the number of times coded by coder 1, but not coder 2
m_{01} = the number of times coded by coder 2, but not coder 1

To obtain the variance of m_{11} we let y_2 = 1, if, in a single document, both coders select term T_j and y_2 = 0, otherwise. Let p = the probability that term T_j will be coded on a single trial (assumed to be the same for both coders).

Then

$$P(y_2 = 0) = 1-p^2$$

$$P(y_2 = 1) = p^2$$

$$Ey_2 = p^2$$

$$Ey_2{}^2 = p^2$$

$$\mathrm{Var}\ y_2 = p^2(1-p^2)$$

*See, for example, J. Jacoby and V. Slamecka, "Indexer consistency under minimal conditions", Documentation Incorporated, 7900 Norfolk Avenue, Bethesda, Maryland, November, 1962.

**See Morris H. Hansen, William N. Hurwitz and Leon Pritzker, "The estimation and interpretation of gross differences and the simple response variance", U.S. Department of Commerce, Bureau of the Census, February 15, 1963.

Thus, in a sample of n independent document,

$$m_{11} = \sum_{i=1} y_{2i}, \text{ and}$$

$$\text{Var } m_{11} = n\, p^2(1-p^2)$$

Now, consider k independent coders, each of whom encodes the sample of n documents. We wish to obtain the variance of

$$\bar{n}_k = \frac{n_{12} + n_{13} + \cdot \cdot \cdot + n_{k-1,k}}{\binom{k}{2}} \qquad (13)$$

That is, $\bar{n}_k$ is an average of the number of terms encoded in common over all possible pairs of coders. Let

y_k = 0 if no pair of coders has coded T_j

= 1 if exactly one pair of coders has coded T_j

= 2 if exactly two pairs of coders have coded T_j

etc.

Then,

$$p(y_k = 0) = (1-p)^k + k\, p\, (1-p)^{k-1}$$

$$p(y_k = 1) = \binom{k}{2} p^2\, (1-p)^{k-2}$$

$$p(y_k = 2) = 0$$

$$p(y_k = 3) = \binom{k}{3} p^3\, (1-p)^{k-3}$$

etc.

In general,

$$p(y_k = 0) = (1-p)^k + kp\, (1-p)^{k-1}$$

$$p\left(y_k = \binom{r}{2}\right) = \binom{k}{r} p^r\, (1-p)^{k-r}, \; (r = 2, 3, \ldots, k)$$

$$= 0, \text{ otherwise.}$$

Hence,

$$E(y_k) = \sum_{r=2}^{k} \binom{r}{2}\binom{k}{r} p^r\, (1-p)^{k-r}$$

$$= \binom{k}{2} p^2 \sum_{r=2}^{k} \left\{ \frac{(k-2)!}{(r-2)!\,(k-r)!}\, p^{r-2}\, (1-p)^{k-r} \right\}$$

$$= \binom{k}{2}\, p^2 \qquad (14)$$

$$\text{Var } y_k = E(y_k^{\,2}) - (Ey_k)^2$$

$$= \sum_{r=2}^{k} \left\{ \binom{r}{2}^2 \binom{k}{r} p^r\, (1-p)^{k-r} - \binom{k}{2}^2 p^4 \right\} \qquad (15)$$

The first term of (15) can be recognized as the expectation of $1/4\; r^2\, (r-1)^2$ which reduces to

$$1/4\, [k(k-1)(k-2)(k-3)p^4 + 4k(k-1)(k-2)p^3 + 2k(k-1)p^2]$$

Hence

$$\text{Var } y_k = \left\{ \binom{k}{2} p^2\, (1-p^2) + 2(k-2)\binom{k}{2} p^3(1-p) \right\} \qquad (16)$$

Also,

$$\bar{n}_k = \sum_{i=1}^{n} y_{ki} / \binom{k}{2} \qquad (17)$$

where the sum is over all n documents in the sample. Since the documents are independent (by assumption)

$$\text{Var } \bar{n}_k = \frac{n}{\binom{k}{2}} [p^2(1-p^2) + 2(k-2)p^3(1-p)] \qquad (18)$$

If instead of $\binom{k}{2}$ pairs of related coders we had $\binom{k}{2}$ independent pairs of coders the variance would be just the first term of the above expression. Hence we lose sensitivity by taking all possible pairs among a group of coders.

It is more important, however, to think in terms of the amount of information supplied per coder. The number of coders required for $\binom{k}{2}$ independent pairs is $2\binom{k}{2}$ and for $\binom{k}{2}$ related pairs is k. Therefore, the relative efficiency (per coder) of the related-pairs estimate to the independent-pairs estimate is

$$\text{R. E.} = \frac{\dfrac{2\binom{k}{2} n\, [p^2(1-p^2)]}{\binom{k}{2}}}{\dfrac{kn[p^2(1-p^2) + 2(k-2)p^3(1-p)]}{\binom{k}{2}}}$$

$$= \frac{2p^2(1-p^2)\binom{k}{2}}{k[p^2(1-p^2) + 2(k-2)p^3(1-p)]}$$

$$= \frac{(k-1)\, p^2(1-p^2)}{[p^2(1-p^2) + 2(k-2)p^3(1-p)]} \qquad (19)$$

The relative efficiency for $k = 6$ and $k = 12$ has been computed for various values of p as follows:

	p								
k	.1	.2	.3	.4	.5	.6	.7	.8	.9
6	2.9	2.2	1.8	1.5	1.4	1.3	1.2	1.1	1.0
12	3.9	2.5	2.0	1.6	1.4	1.3	1.2	1.1	1.1

Since, it is easy, computationally, to handle pairs as though they were independent, it is important to have a correction factor for the adjustment of such variances, to put them on an "independent-pairs" basis. Such an adjustment factor has been computed for various values of p and k and is shown in Table 7. The correction is made by multiplying the dependent-pairs variance by the correction term.

It will be recalled that the coefficient of consistency is, in effect, the set intersection of documents in which a given term is coded divided by the set union. That is,

$$CC = \frac{\bar{n}_k}{n - \bar{n}_k'} \qquad (20)$$

where $\bar{n}_k'$ is the average number of times in which the given term was not coded by both members of a pair, averaged over all possible pairs.

Table 7. Adjustment factors* for variances computed from related pairs of matched codes.

	k					
p	3	4	5	6	7	8
0.1	1.18	1.36	1.55	1.73	1.91	2.09
0.2	1.33	1.67	2.00	2.33	2.67	3.00
0.3	1.46	1.92	2.39	2.85	3.31	3.77
0.4	1.57	2.14	2.71	3.29	3.86	4.43
0.5	1.67	2.33	3.00	3.67	4.33	5.00
0.6	1.75	2.50	3.25	4.00	4.75	5.50
0.7	1.82	2.65	3.47	4.29	5.12	5.94
0.8	1.89	2.78	3.67	4.56	5.44	6.33
0.9	1.95	2.90	3.84	4.79	5.74	6.68

* $\dfrac{p^2(1-p^2) + 2(k-2)p^3(1-p)}{p^2(1-p^2)}$

It is clear that the above means and variances for $\bar{n}_k$ will hold for $n - \bar{n}_k'$ if we replace p by $1 - p = q$, that is, the probability that a given term will not be coded.

We consider now the variance of the ratio given by equation (20). Consider random variables u and v whose true values are U and V, respectively, and consider the variable $W = u/v$ as an estimator of $w = U/V$. Then

$$\text{Var } w^2 = \left[W^2 \frac{\text{Var } u}{U^2} + \frac{\text{Var } v}{V^2} - 2 \frac{\text{Cov } uv}{UV} \right] \qquad (21)$$

where Cov uv is the covariance between u and v and the other quantities are as defined previously. In our case, we identify u with $\bar{n}_k$ and v with $n - \bar{n}_k'$, and we must find Cov uv.

Consider k coders (analysts) and the following variables:

$y_k = 0$ if no pair of coders has coded

T_k

= 1 if exactly one pair of coders has coded T_k

= 3 if exactly three pairs of coders have coded T_k

etc.

z_k = 0 if no pair of coders has not coded T_k

= 1 if exactly one pair of coders has not coded T_k

= 3 if exactly three pairs of coders have not coded T_k

etc.

In general,

$$p(y_k = 0) = (1-p)^k + kp\ (1-p)^{k-1}$$

$$p(y_k = \binom{r}{2}) = \binom{k}{r} p^r\ (1-p)^{k-r},\ (\ r = 2, 3, \ldots, k)$$

As before, and

$$p(z_k = 0) = (1-q)^k + k\ q\ (1-q)^{k-1}$$

$$p(z_k = \binom{r}{2}\ ) = \binom{k}{r}\ q^r\ (1-q)^{k-r}\quad (r = 2, 3, \ldots, k)$$

It can be shown that

$$\text{Cov } y_k z_k = \binom{k}{2}\left\{\left[\binom{k-2}{2} - \binom{k}{2}\right] p^2 (1-p)^2\right\} \qquad (22)$$

This covariance summed over n documents and averaged over all possible pairs of coders would be Cov $\bar{n}_k$, $\bar{n}'_k$. Hence

$$\text{Cov } \bar{n}_k\ (n-\bar{n}'_k) = \frac{n\left[\binom{k-2}{2} - \binom{k}{2}\right]\ p^2(1-p)^2}{\binom{k}{2}} \qquad (23)$$

where $\binom{k-2}{2}$ is defined to be zero for k=3.

Thus, to obtain an estimate of the variance of the coefficient of consistency, (equation 20), we replace, in equation (21),

W by $\bar{n}_k/(n - \bar{n}'_k)$

U by $\bar{n}_k$

V by $n - \bar{n}'_k$

Var u by eq. (18)

Var v by eq. (18) with p replaced by (1-p)

Cov uv by eq. (23)

and insert an estimate of p from the sample of n documents.