STUDY OF THE RELIABILITY OF CODING OF CENSUS RETURNS

Herman H. Fasteau, J. Jack Ingram and Ruth H. Mills
U. S. Bureau of the Census

General

In operations in which verbal descriptions are coded to alphabetic or numeric equivalents, the codes assigned may have varying degrees of reliability depending upon the quality of the verbal description, the coder's training and experience, and the coding instructions and materials he uses. Customarily, measures of the reliability of this type of coding have been in terms of clerical error rate, i.e., the proportion of cases coded in error.

The purpose of this paper is to outline a technique for evaluating the reliability of individual codes, to show the application of the technique to industry and occupation coding in the 1960 Census of Population, and by an analysis of selected codes to identify some areas in which improvements can be made in questionnaire design, training of respondents and interviewers, code structures, rules for coding, coding materials, and training of coders. Among the reasons for choosing industry and occupation coding for this analysis are: the operation is characterized in part by use of judgment on the part of the coder; there is a tremendous variety of verbal descriptions which must be fitted to a relatively small number of symbolic codes; the operation itself is relatively short term, being preceded by an intensive training program; and, the source materials and procedures are the results of years of research and development with improvements and changes being accumulated from Census to Census. The results contained in this report are partial and preliminary and will be included in a fuller report to be published later.

The Coding Operation[1/]

Information supplemental to basic Census questions was obtained in the Census from a sample consisting of members of every fourth household and from every fourth person not a member of a household. Industry and occupation questions were asked of each person in this sample who had worked since 1950, who was 14 years of age or over and who was not in the Armed Services at the time of the Census. The questions concerned the job held the previous week, or, if none, the last job held. They were:

1. "For whom did he work?
   (Name of company, business, organization or other employer.)"
2. "What kind of business or industry was this?
   (For example: County junior high school, auto assembly plant, TV and radio service, retail supermarket, road construction, farm.)"
3. "Is this primarily Manufacturing, Wholesale trade, Retail trade, or Other?"
4. "What kind of work was he doing?
   (For example: 8th grade English teacher, paint sprayer, repairs TV sets, grocery checker, civil engineer, farmer, farm hand.)"
5. Was this person an employee of (a) a private company, business or individual, for wages, salary, or commissions, (b) a Federal, State, county or local government, was he (c) self-employed in own business, professional practice or farm or was he (d) working without pay in a family business or farm?

About 82 percent of the information for persons in the 25 percent sample was obtained by self-enumeration; the respondent had these five questions before him and wrote out the answers himself. These answers were later transcribed by Census enumerators to the regular Census schedules. The third question did not appear on the Census schedule; however, in transcription, the enumerator was instructed to use this information in the answer to question number 2. Where any information was missing on the self-enumerated schedule, the enumerator was instructed to obtain the missing information from the respondent. In those cases directly enumerated the enumerator was instructed to obtain the answer to question number 3 as though it were included on the schedule and write the answer as a part of question number 2.

Based upon the answers to the above questions, the coder was required: to assign one of the 149 industry codes and one of the 296 occupation codes to the person, to verify that there was a code for class of worker on the schedule, and to determine that all three codes were in correct combination. Codes are stated in single alphabets or in groups of three digits, e.g., A "Agriculture" or 023 "Professional, technical, and kindred workers, clergymen."

The documents moved through the coding units a State at a time. Each coder was supplied with "Company Name Lists" for the counties and Standard Metropolitan Statistical Areas (SMSA's) in the State he was coding. The "Company Name

Lists" included the names and industry codes of all manufacturing companies known to have fifty or more employees and all other types of companies known to have 100 or more employees (obtained from the 1958 Economic Censuses). He was also provided with an Alphabetical Index of Occupations and Industries[2] containing about 30,000 different ways in which the 296 occupations could be described or combined with specific types of industries and somewhat fewer number of ways in which the industries could be described.[3]

The following is a rough summary of the series of steps taken by the coder in assigning the codes:

1. He searched the appropriate Company Name List for the name of the company given in question 1, above.

   a. If the company was found, he immediately assigned the designated industry code. If the company was not found,

   b. He matched the answers to questions 2 and 3 (the kind of business or industry) against the Alphabetical Industry Index. If the answers fitted a description in the Alphabetical Index, he assigned the designated code. If the description was not sufficient (after searching the words out of alphabetical sequence) to arrive at a code from the Industry Index, he used any relevant words in the company name or occupation entry.

   If an appropriate entry were found, he assigned the code given in the Index. If a highly similar entry could not be found, or there was contradictory information in the answers to the four questions, he referred the item to an expert coder.

2. Having assigned the industry code, he then matched the answer to question 4 (occupation description) to the occupation Index in a manner similar to 1(b), above. The occupation codes for a given occupation entry in the Index, however, often vary from one major industrial group or detailed industry code to another and/or by class of worker. If the industry code already decided upon and/or the class of worker on the schedule were not provided as allowable in combination with the occupation code, the item was to be referred to an expert coder.

3. Class of worker was then verified as to its consistency with the industry and occupation codes assigned.

Control of the Quality of Coding

The data used for this study are largely a by-product of the quality control scheme used in the 1960 Census. This scheme is discussed more fully in an earlier paper.[4] For the purposes of this discussion, it can be described briefly as follows:

For industry and occupation coding verification there was a selection of a 1 in 40 sample of households from the 25 percent Census sample. Persons in the experienced civilian labor force falling into the 1 in 40 sample households were coded independently by three different clerks having approximately the same training and coding experience. All three of these coders coded from the Census Schedule, but only the third and last coder entered his code on the Census Schedule. This coder is referred to below as the "Census Coder." The coded results were then matched against one another and placed into one of the four categories:

1. All three agreed on the code (category AAA).

2. Two agreed on a code and one disagreed (category $\overline{AAB}$).

3. Two agreed to refer the item and one coded it (category $\overline{RRA}$).

4. All three disagreed (category $\overline{ABC}$).

For quality control purposes, if two coders agreed and the third disagreed, a quality demerit was assigned to the disagreeing coder. The demerit was utilized as though it represented a defective item. In controlling the quality, categories RRA and ABC were excluded from the sample; however, in this analysis all four categories were included.

Methodology

The sample for this study includes one-fourth of the AAA cases and all cases of disagreement of any kind occurring in the 1 in 40 sample. A measure was developed for scaling the codes in terms of degree of consistency of application. It is called an Index of Consistency[5] and answers the question: "Given a code, how consistently was it independently applied by three different coders looking at the same description?"

The Index takes the form:

$$C = \frac{3\,N_{AAA} + N_{\overline{AAB}}}{3\,N_{AAA} + 2\,N_{\overline{AAB}} + (N_{\overline{ABB}} + N_{\overline{ABC}})}$$

where

N_{AAA} = the number of cases for which all three coders applied the code under consideration.

$N_{\overline{AAB}}$ = the number of cases for which any two of the coders applied the code under consideration and the third applied some other code.

$N_{\overline{ABB}}$ = the number of cases for which only one coder applied the code under consideration and the other two coders agreed upon some other code.

$N_{\overline{ABC}}$ = the number of cases in which all three coders disagreed but one of them applied the code under consideration.

The index of consistency is an average measure of agreement among coders in independently assigning a given code when looking at the description. It is not a direct measure of the quality of the codes in the Census, but serves as a useful device in pointing out those codes which may have low reliability because of confusion on the part of the coder as to what code to assign for a given description. A referral action is here treated as a bona fide code.

An auxiliary measure of consistency answers the question: "Considering the difference cases for a code, what are the associated codes and how frequently are they so associated?" Difference cases are all those cases not falling into the AAA category. An "associated" code is the code or codes that the other two coders entered when one coder entered the code under consideration. The purpose of this question is to point out broad paths which lead to identification of specific areas in which improvements can be made. The answer leads to further questions as to how and why the codes were assigned.

As a starting point, industry codes with low indices of consistency were selected for a two-phase analysis. The first phase of the analysis was to answer the above questions; the second phase was to determine the correct code for the difference case. One part of determining the correct code was to obtain the code which the coder should have assigned using all the information at his disposal; the other part was to obtain the one correct code which was arrived at through the use of detailed research materials coupled with the expert's wide experience. This distinction in "correct" codes is necessary because in some cases the descriptions obtained in the Census could lead to a wrong code, or in some cases the only action that the coder could correctly take was to refer.

The second phase of the analysis called for assigning a 50 percent sample of the difference cases to a panel of experts who assigned the correct code for each case. They received only the answers as obtained in the Census; they had no way of knowing what codes had been assigned by the three coders.

Once the correct code has been assigned, each of the above difference cases was analyzed in terms of:

1. The relation of the code under consideration to its associated code or codes and to the correct code. (It is possible that neither the code under consideration nor an associated code is correct.)
2. The correctness of the code used in the Census tabulations, whether or not the compared codes crossed major Census classifications, and whether or not they were "basket-type" codes such as "miscellaneous," or "not elsewhere classified," etc.
3. The relation of the Census description to entries in the coder's reference material according to ease and type of matching.
4. The presence or absence of "key" words in either the description obtained in the Census or that given in the reference material.
5. The apparent reason for the incorrect code, if any.[6] Major classifications of reasons are: (a) Coder's failure to follow instructions; (b) inadequate descriptions on the Census schedule; (c) inadequate reference materials or instructions; (d) inadequacies in both the description on the schedule and the reference materials or instructions; (e) clear-cut clerical errors such as the transposition of digits.

Indices of Consistency

Table 1 gives the distributions of the 149 industry codes and 296 occupation codes by size of consistency index. A substantially larger proportion of occupation codes has high indices of consistency than do industry codes. Thirty-nine percent of the industry codes as compared with 48 percent of the occupation codes had indices between .90 and .99. For both types of coding, the .90 to .99 class accounted for about 74 percent of the experienced civilian labor force.

Table 1.--Number of Industry and Occupation Codes by Index of Consistency

Index of consistency	Industry codes			Occupation codes		
	Number of codes	Percent of codes	Estimated percent of Labor Force 4/	Number of codes	Percent of codes	Estimated percent of Labor Force 4/
.900-.999	59	39.3	73.6	142	48.0	74.4
.800-.899	56	37.3	18.8	93	31.4	22.5
.700-.799	23	15.3	6.6	41	13.9	2.6
.600-.699	6	4.1	0.8	11	3.7	0.2
.500-.599	2	1.3	0.1	4	1.3	0.3
Less than .500	4	2.7	0.1	5	1.7	3/
Total codes 1/	150	100.0	100.0	296 2/	100.0	100.0

1/ Includes the codes for "not reported."
2/ Excludes Code 000 "Accountants and auditors" because of programming error.
3/ Less than 0.5 percent.
4/ Estimates based on a sample of 420,000. These cases do not include those in which the Census coder referred the description to an expert.

Inspection of the distribution in Table 1 indicated that perhaps the greatest payoff from a preliminary analysis in depth would occur in analyzing the 20 percent or so codes having the lowest indices. The cutoff point for both industry and occupation codes is at Index .80.

The fact that the two above numbers seem to be complementary is only coincidental. Table 2 is a list of the 35 industry codes having Indices less than .80, and the most interesting thing about that list is that it includes every industry code for wholesale trade (Codes 606-629).

Table 2.--List of Industry Codes Having Indices of Consistency of Less Than .800

Code	Description	Number of cases in sample		Index
		AAA	Other	
	Manufacturing, Durable Goods			
208	Miscellaneous wood products, except furniture	76	379	.739
236	Miscellaneous nonmetallic mineral and stone products	129	469	.788
247	Fabricated structural metal products	307	1,200	.775
249	Not specified metal industries	-	38	.073
	Manufacturing, Nondurable Goods			
326	Not specified food industries	26	343	.568
367	Miscellaneous fabricated textile products	11	370	.462
389	Miscellaneous paper and pulp products	-	430	.283
419	Miscellaneous petroleum and coal products	28	138	.739
429	Miscellaneous plastic products	130	446	.799
459	Not specified manufacturing industries	-	94	.060
	Transportation, Communication and Other Public Utilities - Transportation			
516	Warehousing and storage	112	606	.724
519	Petroleum and gasoline pipelines	8	105	.564
526	Services incidental to transportation	52	350	.675
	Utilities and Sanitary Services			
568	Gas and steam supply systems	127	469	.788
569	Electric-gas utilities	103	438	.761
579	Other and not specified utilities	10	54	.714
	Wholesale and Retail Trade - Wholesale			
606	Motor vehicles and equipment	106	423	.775
607	Drugs, chemicals and allied products	87	340	.775
608	Dry goods and apparel	74	326	.760
609	Food and related products	479	1,625	.798
616	Farm products - raw materials	86	472	.725
617	Electrical goods, hardware, and plumbing equipment	208	705	.798
618	Machinery, equipment and supplies	137	1,049	.661
619	Petroleum products	111	654	.710
626	Miscellaneous wholesale trade	552	1,895	.795
629	Not specified wholesale trade	68	583	.642
	- Retail trade			
637	Dairy products stores and milk retailing	51	393	.660
676	Lumber and building materials retailing	364	1,400	.778
687	Fuel and ice dealers	105	475	.753
689	Miscellaneous retail stores	447	1,722	.779
696	Not specified retail trade	85	850	.607
	Business and Repair Services			
807	Miscellaneous business services	597	2,580	.760
809	Miscellaneous repair services	315	1,163	.786
	Professional and Related Services			
888	Nonprofit membership organizations	266	1,014	.781
898	Miscellaneous professional and related services	85	551	.683

Since "Wholesale trade" was the only industry group for which all the codes had low Indices of Consistency and since its codes alone comprised almost 30 percent of the highly confused industry codes, those ten codes were selected for this analysis.

Associated Codes

Often just knowing the codes (and their descriptions) that have a high degree of association with low-consistency codes can simplify the identification of problems and provide clues to the reasons for inconsistency. This is well demonstrated by the wholesale trade codes.

For seven of the ten "Wholesale trade" codes the associated code was "Referral" more frequently than any other code. The codes not included in the seven are "Motor vehicles and Equipment," "Drugs, chemicals and allied products" and "Dry goods and apparel," Next to "Referral" the code most frequently associated with the code under consideration was one specifying the same type of product but in retail trade or manufacturing. Table 3 shows the distribution of general classes of associated codes for each of the Wholesale codes.

Table 3.--Types of Codes Associated With the Wholesale Trade Difference Cases

Wholesale codes	Percent distribution of types of associated codes				Total cases (100%)**
	Referrals	Same type product but retail or Mfg.*	Other wholesale codes	Other codes	
Motor vehicles and equipment	22	49	10	19	(415)
Drugs, chemicals and allied products	18	42	20	20	(328)
Dry goods and apparel	14	45	19	22	(319)
Food and related products	19	57	8	16	(1,587)
Farm products-raw materials	21	7	20	52	(462)
Electrical goods, hardware and plumbing equipment	27	24	14	35	(697)
Machinery, equipment and supplies	34	21	14	31	(1,018)
Petroleum products	30	44	13	13	(634)
Miscellaneous wholesale trade	18	--	18	64	(1,729)
Not specified wholesale trade	45	--	30	25	(566)

*For example, "Manufacturing-motor vehicles and motor vehicle equipment" and "Motor vehicles and accessories retailing" for the first Wholesale trade code listed. See Appendix B for a list of the ways in which the specific codes were grouped.

**The number of cases here is less than the number of difference "Other" cases presented in Table 2 because the cases where the associated code was a blank or an impossible code have been excluded from this table.

It can be assumed that a large proportion of the cases falling into the class "Referrals" and "Same type product but retail or manufacturing" was due to a lack of sufficient information supplied to the coder. To a large extent this is true, as will be shown below. However, "Referral cases are not confined to the "Same type product" class. Appendix C shows how Wholesale codes are associated in coding with the various industry groups. A referred case could have been coded ultimately into any of these industry groups.

Preliminary Analysis of Inconsistent Wholesale Codes

One of the more interesting results of the analysis of these codes is that, if the coder had followed his instructions precisely, he would have referred 75 percent of the difference cases because the descriptions and/or instructions were not adequate for proper coding. In spite of this, in 45 percent of the difference cases the Census coder managed to arrive at the correct code, and in an additional 19 percent of the cases he did refer; so that the presumption is that 64 percent of the difference cases were correctly coded in the Census (assuming that the experts handling the referral cases assigned the correct code). Perhaps one of the reasons for not referring a case on the part of the Census coders was that the rules for referring were very rigid and the coder often felt he knew the correct code.

On an overall basis, for the Wholesale codes, the inconsistent code under analysis was the ultimately correct code for the given case in slightly more than half of the cases. This is not always true, however, for each of the ten codes. At the extremes: when "Wholesale trade, not specified" (629) was used, it was correct in 32 percent of the cases; on the other hand, when "Wholesale trade, electrical goods, hardware and plumbing equipment" (617) was used, it was correct 72 percent of the time.

Table 4.--Codes Under Consideration (Wholesale trade) in Difference Cases, Classified by Correctness and Effect Upon Census

Classification	Code under consideration										
	Total	606	607	608	609	616	617	618	619	626	629
Code under consideration correct:	55%	44%	40%	52%	53%	56%	72%	57%	64%	56%	32%
Assigned by Census coder	32	20	31	40	35	38	36	33	33	32	17
Assigned by another coder but Census coder assigned a different code	12	14	3	6	11	9	18	16	9	12	0
Assigned by another coder but Census coder referred	11	10	6	6	7	9	18	8	22	12	15
Code under consideration incorrect:	45	56	60	48	47	44	28	43	36	44	68
Assigned by Census coder	19	30	28	9	22	27	9	12	13	18	37
Assigned by another coder but Census coder assigned a different code	18	19	26	27	16	12	17	25	13	18	12
Assigned by another coder but Census coder referred	8	7	6	12	9	5	2	6	10	8	19
Total number of cases (100%)	(918)	(59)	(32)	(33)	(221)	(56)	(78)	(141)	(86)	(171)	(41)

From Table 5 it can be seen that a little over half of the Wholesale trade difference cases had adequate descriptions. In 32 percent of the cases the descriptions in the schedule were inadequate for coding; and in 14 percent of the cases the coding materials themselves were inadequate.

Table 5.--Relation of Industry Description on Census Schedule Coding Materials,* Wholesale Trade Difference Cases

	Percent of difference cases		
Adequate Description			54
No coding error made (combination of correct code and referral)		25	
Highly similar words in alphabetical Index	14		
Easy inference required to match to coding materials	11		
Could have been correctly coded		29	
Highly similar words in alphabetical Index	20		
Easy inference required to match to coding materials	9		
Inadequate Description			32
Problems of detailed classification**		19	
Description more general than alphabetical Index	11		
Description more detailed than alphabetical Index	6		
Other	2		
Problems of major classification only		13	
Coding Materials Inadequate			14
Total difference cases (100%)			(918)

*"Coding materials" refers to the alphabetical Index and/or the "Company Name List."
**In some of these cases there can also be a problem of major group classification.

In 25 percent of these Wholesale trade difference codes none of the three coders assigned an incorrect code -- one or two of them assigned the correct wholesale code and two or one of them referred the case.

In 29 percent of the difference cases, one of the three coders assigned an incorrect code when there was no excuse for it. If the coders had followed their instructions explicitly, they would have arrived at the correct code.

In 19 percent of the cases the industry description on the schedule could not be properly matched to the Index for detailed classification (other than major industrial classification). In most of these cases the description on the Census schedule was too general: for instance, entered on the schedule would be "shipping" whereas one needed to know the product shipped or the shipping product manufactured in order to match the description to the Index. Only in half as many cases was the description on the schedule more detailed than the Index; most frequently separate parts of the description could be matched to different lines in the Index leading to different codes.

In 13 percent of the difference cases there was no trouble in matching the description to the Index, but information as to the major industrial classification was absent or incorrect.

In the remaining 14 percent of the difference cases the coding materials themselves were inadequate. In a sizeable proportion of these cases the "Company Name Lists" were in error.

Some Suggested Actions for Improvement

While the above observations apply primarily to codes in wholesale trade, they indicate some actions which can be taken to improve the reliability of industry coding. The first of these has to do with improvement of the description of industry activity provided to the coder. The emphasis would be upon the respondents and would utilize Census public relations media in informing them how the question should be answered and the importance of answering it correctly. An accompanying action would be intensification of training and control of interviewers in requiring and obtaining correct answers to the industry activity question.

It was noted earlier that the major industrial classification was not explicitly provided for on the Census schedule. It was asked explicitly on the self-enumeration questionnaire. It is possible that the description provided to the coder will be improved if the question is included on the Census schedule. In considering a proposal such as this, however, it is necessary to consider whether dis-economies arising from such a change will be more than offset by improvement in coder reliability. This is a question which must be investigated.

Further Research

This has been a preliminary report presented for the purpose of indicating the procedure in a study of coding reliability. The analysis in depth of wholesale trade difference cases, above, is only a part of that analysis with future emphasis to be placed upon the effect of a lack of reliability upon detailed published Census figures.

In addition to the above, the study will continue largely as follows:

a. Complete the analysis outlined above for all industry codes having low Indices of Consistency.

b. Subject all occupation codes having low Indices of Consistency to similar analysis.

c. Analyze a sample of difference cases, both industry and occupation, for codes having high Indices of Consistency.

d. Analyze a sample of AAA cases to determine how frequently consistency of response led to an incorrect code. Research to date has indicated that at least for occupation codes the incidence of error in AAA cases is extremely small.

e. Investigate the correctness of codes applied in the Census for cases in which all three coders referred.

f. As a result of the above analyses, provide a comprehensive list of changes which can be made to improve the reliability of coding.

FOOTNOTES

1/ The industry and occupation coding specifications and training materials were developed by the Occupation and Industry Section of the Economic Statistics Branch of the Population Division of the Bureau of the Census. Members of this Branch, in particular William J. Mulligan, Stanley Greene and Mrs. Gladys M. Dodd, have cooperated in this analysis of coding consistency.

2/ U. S. Bureau of the Census, 1960 Census of Population, Alphabetical Index of Occupations and Industries, Washington, D. C., February 1960. (Revised Edition, October 1960.)

3/ For example: There are codes for about 90 differently described automobile company entries; there are codes for over 300 different types of occupation descriptions containing the word "engineer" although there were only ten different codes into which engineers can be classified. There are listed about 280 different types of college teachers, 280 different types of "inspectors," 220 "repairmen," 140 "mechanics." There are also such single-entry items as "Krippendorfer" (coded as an "operative and kindred worker, not elsewhere classified" in the "Leather and leather products: footwear, except rubber" industry); "Osmosis man" (coded also as an "operative and kindred worker, n.e.c." but in the miscellaneous food preparation and kindred products industry); and "Shill" (coded as an "attendant, recreation and amusement" in the miscellaneous entertainment and recreation services).

4/ Cf. M. H. Hansen, H. H. Fasteau, J. J. Ingram and G. Minton, "Quality Control in the 1960 Censuses," New Frontiers in Administrative and Engineering Quality Control, ASQC, Milwaukee, 1962, pp. 323-339.

5/ Max Bershad of the Statistical Research Division of the Bureau of the Census developed the model for this Index. See Appendix A for a fuller presentation. This is similar to the Index presented at this meeting by L. Pritzker, R. Hanson, "Measurement of Errors in the Censuses of Population and Housing."

6/ If one coder assigned the code under consideration which turned out to be correct, and the other two coders referred the case because the description was ambiguous, the case was treated as though there were no incorrect code assigned.

APPENDIX A: Derivation of the Index of Consistency

Let: j = the document

$j = 1, 2, 3 \ldots\ldots N$

i = the coder

$i = 1, 2, 3 \ldots\ldots K$

i' = a coder other than the one selected as i

Let: X_{ij} = 1 when coder i classifies the j^{th} document as code h.

= 0 when coder i classifies the j^{th} document as other than code h.

$$X.. = \frac{\sum_{i=1}^{K} \sum_{j=1}^{N} X_{ij}}{KN}$$

For perfect consistency in the use of code h,

$$E(X_{ij} - X..)^2 = E(X_{ij} - X..)(X_{i'j} - X..)$$

$$\text{or } 1 = \frac{E(X_{ij}-X..)\,(X_{i'j}-X..)}{E(X_{ij}-X..)^2}$$

and a measure of consistency for code h would be:

$$C = \frac{E(X_{ij}-X..)\,(X_{i'j}-X..)}{E(X_{ij}-X..)^2}$$

$$= \frac{E(X_{ij}\,X_{i'j} - X..^2}{E(X_{ij}^2) - X..^2}$$

Since X.. in this study will be very small, the measure of consistency becomes

$$C = \frac{E(X_{ij}\,X_{i'j})}{E(X_{ij})^2} = \frac{E(X_{ij}\,X_{i'j})}{E(X_{ij})} = \frac{E(X_{ij}\,X_{i'j})}{X..}$$

The numerator and denominator of C will be estimated from a sample of n documents and of coders, each sample document having been coded independently by three different people.

For three coders, an estimate of $E(X_{ij}\,X_{i'j})$ for code h is

$$\frac{\sum_{j=1}^{n} (x_{1j}x_{2j} + x_{1j}x_{3j} + x_{2j}x_{3j})}{3n}$$

For code h = A, and other codes designated by B,

$$\frac{\sum_{j=1}^{n}(x_{1,j}x_{2,j} + x_{1,j}x_{3,j} + x_{2,j}x_{3,j})}{3n} = \frac{(n_{AAA} + n_{AAB}) + (n_{AAA} + n_{ABA}) + (n_{AAA} + n_{BAA})}{3n}$$

$$= \frac{3n_{AAA} + n_{AAB} + n_{ABA} + n_{BAA}}{3n}$$

$$= \frac{3n_{AAA} + n_{\overline{AAB}}}{3n}$$

An estimate of X.. is

$$\bar{x} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{3} x_{ij}}{3n} = \frac{3n_{AAA} + 2n_{\overline{AAB}} + n_{\overline{ABB}} + n_{\overline{ABC}}}{3n}$$

where C is any code other than A or B.

Therefore, an estimate of

C is $$\frac{3n_{AAA} + n_{\overline{AAB}}}{3n_{AAA} + 2n_{\overline{AAB}} + n_{\overline{ABB}} + n_{\overline{ABC}}}$$

APPENDIX B: Wholesale Trade Codes and Associated Same Product Manufacturing and Retail Trade Codes

Wholesale Code		Associated Codes of Same Type Product		
Code	Description	Code	Description	Percent
606	Motor vehicles and equipments	267	Manufacturing, durable goods, transportation equipment, motor vehicles and motor vehicle equipment	15
		656	Retail trade - motor vehicles and accessories retailing	34
607	Drugs, chemicals and allied products		Manufacturing, non-durable goods, chemicals and allied products	
		406	- Synthetic fibers	-
		407	- Drugs and medicine	10
		408	- Paints, varnishes and related products	4
		409	- Miscellaneous chemicals and allied products	19
		658	Retail trade, drug stores	9
608	Dry goods and apparel		Manufacturing, non-durable goods,	
		349	- Yarn, thread and fabric mills	17
		B	- Apparel and other accessories	21
		367	- Miscellaneous fabricated textile products	*
		646	Retail trade, apparel and accessories stores except shoe stores	7
609	Food and related products		Manufacturing, non-durable goods, Food and kindred products	
		306	- Meat products	14
		307	- Dairy products	10
		308	- Canning and preserving fruits, vegetables and seafoods	10
		309	- Grainmill products	1
		316	- Bakery products	2
		317	- Confectionery and related products	1
		318	- Beverage industries	3
		319	- Miscellaneous food preparation and kindred	2
		326	- Not specified food industries	2
		F	Retail trade - Food stores, except dairy products	12
616	Farm Products - Raw Materials	A	Agriculture	7
617	Electrical goods, hardware and plumbing		Manufacturing, durable goods	
		259	- Electrical machinery, equipment and supplies	16
		649	Retail trade - Household appliances TV, and radio stores	8
618	Machinery, equipment and supplies		Manufacturing, durable goods, Machinery, except electrical	
		256	- Farm machinery and equipment	2
		257	- Office, computing and accounting machines	5
		M	- Miscellaneous machinery	10
		666	Retail trade - Hardware and farm equipment stores	4
619	Petroleum products		Manufacturing, non-durable goods	
		416	- Petroleum refining	24
		419	- Miscellaneous petroleum and coal products	*
		657	Retail trade - gasoline service stations	9
		687	- Fuel and ice dealers	11

*Less than 0.5 percent.

APPENDIX C: Distribution of Codes Associated With Each Wholesale Code by Major Industrial Classification

	606	607	608	609	616	617	618	619	626	629
Referrals	.22	.18	.14	.19	.21	.27	.34	.30	.18	.45
Agriculture	*	-	-	.05	.07	-	*	-	.04	-
Forest and Fisheries	-	-	-	*	-	-	-	-	-	-
Mining	-	*	-	*	*	-	.01	.04	.01	-
Construction	-	*	-	*	-	.03	.01	-	.02	*
Manufacturing	.24	.39	.45	.45	.13	.25	.26	.26	.28	.06
Durable goods	.20	.04	.01	*	*	.24	.25	.01	.14	.02
Nondurable goods	.04	.35	.44	.45	.13	.01	.01	.25	.14	.04
Transp. Commun. and other Public Utilities	.04	.02	.03	.06	.16	.03	.01	.05	.03	.05
Wholesale and Retail Trade	.46	.37	.34	.23	.34	.37	.28	.34	.40	.37
Wholesale	.10	.20	.19	.08	.20	.14	.14	.13	.18	.30
Retail	.36	.17	.15	.15	.14	.23	.14	.21	.22	.07
Finance, Ins. and Real Estate	.01	*	.01	.01	.04	*	*	*	*	.01
Business and Repair Services	.03	.01	.01	.01	.02	.04	.07	.01	.02	.03
Personal Services	-	*	.01	-	*	.01	*	*	.01	*
Entertainment and Recreation Services	-	-	-	*	-	-	*	-	*	*
Professional and Related Services	-	*	-	*	.02	*	.01	*	*	.01
Public Administration	-	*	-	*	*	*	-	-	-	.01
Industry not reported	-	-	*	-	*	-	*	*	*	*
Total **(1.00)	(415)	(328)	(319)	(1,587)	(462)	(697)	(1,018)	(634)	(1,729)	(566)

*Less than .005.

**These totals are slightly less than the totals of the difference cases presented in Table 2 because excluded here are cases where the "associated code" was a blank or an impossible code.