Data Quality Control and Library Research on Political Parties

KENNETH JANDA

As introduced by Naroll in his book (1962) by that title, "data quality control" refers to the systematic evaluation of reliability among written reports of human behavior and social phenomena. In Naroll's words:

Data quality control deals not with individual reports but instead with groups of reports compiled by various authors under various conditions. It depends upon the assumption that some records are made under conditions of higher apparent trustworthiness than others. (p. 2)

For more than a century the only systematic method of evaluating the reliability of reports has been the method of internal and external criticism developed by classical historians. (pp. 1-2)

To evaluate report reliability, I propose instead to apply the general spirit and philosophy of statistical production quality control, as widely used in industry. The general spirit of such quality control is to test regularly, by sampling methods, the hypothesis that something is seriously wrong with production methods. (p. 10)

In data quality control the behavioral scientist working with written records tests Indications of unreliability to see if something is seriously wrong with the statements in the records. To be sure, there is an essential difference in the position of the comparativist and the industrial quality control engineer. (p. 11) . . . Therefore I propose to use the term "control" somewhat more broadly than it is used in industry. By "control of errors" I shall mean not only their detection but also the methods taken to counteract their effect on the results of the study. (p. 12)

Naroll proceeded to demonstrate the utility of data control in a cross-cultural survey of cultural stress in thirty-seven societies as reported in existing ethnographies. His study used six main control factors:

(1) collection of specific case reports by the ethnographer, (2) use by the ethnographer of (direct observation and personal participation in an ongoing culture as a major source of field data, (3) length of stay in the field among the people studied by the ethnographer, (4) familiarity of the ethnographer with the language of the people studied, (5) role of the ethnographer among the people studied (such as social scientist, missionary, government official), and (6) explicitness and generality of the ethnographer's report on the trait in question, with the concomitant presence or absence of a need for inference by the comparativist in order to classify the report. (pp. 14-13)

Analyzing ethnographic data on cultural stress according to these control factors, Naroll found that some factors were "provisionally validated" as sources of reporting bias while others showed no significant relationship to reports of cultural stress. For example, the length of stay (control factor 3) was positively related to witchcraft attribution, but there was no evidence that professional social scientists were more trustworthy ethnographic reporters than missionaries (control factor 5).

As Naroll himself pointed out, most of his control factors were "characteristically applicable to the ethnographic data collection process and may be of little use in controlling other kinds of data reports" (p. 26). Nevertheless he proposed that the methodology of data quality control had general applicability to social science research based on library materials, and he suggested a number of control factors which might be applicable to historiography. More important

than the specific control factors presented in the book was the attention called to the problem of data reliability and to methods of dealing with the problem.

The idea of data quality control has since been applied by Koh (1966) to social science bibliographic references on Korea and by Textor (1968) in his computerized *Cross-Cultural Survey*. This paper outlines another, and more extensive, application of data quality control methodology to literature on political parties under study in the International Comparative Political Parties Project. While the Project departs considerably from the methodology originally set forth by Naroll, it is inspired by the same philosophy in its treatment of data reliability problems.

THE INTERNATIONAL COMPARATIVE POLITICAL PARTIES PROJECT

The ICPP Project was established at Northwestern University in 1967 for the purpose of conducting the first comprehensive, empirically-based, comparative analysis of political parties throughout the world. Data for this analysis are being collected and managed through a variety of information retrieval techniques applied to published and unpublished writings on party politics in ninety foreign countries. The objectives and over-all design of the project are detailed in another place (Janda 1968). This paper will describe only those features of the project which are essential for understanding its data quality control methodology.

Contrary to most cross-national political research, our unit of analysis is the political party rather than the nation-state. Instead of drawing a sample of parties for analysis, we intend to gather data on the *universe* of political parties, defined as those whose candidates won at least 5 per cent of the seats of the lower house of a national legislature in two successive elections in the time period 1950-1962. (A complete list of the parties presently identified for inclusion in the study is given in Janda 1968.) We plan to code each of some 250 political parties on a series of variables (e.g., ideological orientation, centralization of power, method of leadership selection) for subsequent keypunching and statistical analysis. Our primary source of data for coding parties on

these variables consists of such library materials as books, articles, theses, government documents, party documents and newspapers.

This material is stored in the information files of the ICPP Project as copies of the original pages tagged with three-digit code numbers in a fashion quite similar to the practice of the Human Relations Area Files (Murdock 1961). The methodology for handling our files, however, is quite different. We record the pages *and their corresponding code numbers* on 16-mm microfilm for instantaneous retrieval with Eastman Kodak's MIRACODE information retrieval system (Janda 1967a). The basic components of the MIRACODE system are a special microfilm camera and microfilm reader. A film magazine containing information on party politics in a given country can be searched by the MIRACODE reader for logical combinations of codes assigned to individual pages, which are then selectively displayed for the researcher. Approximately one thousand pages of material can be stored on one 100-foot magazine and searched for specified combinations of code numbers in ten seconds.

COVERAGE OF THE LITERATURE AND QUALITY OF THE DATA

Because our data are drawn from library materials, we are dependent upon both the *coverage* and *quality* of the literature on political parties. Coverage of the literature in our files can be assessed rather precisely by reference to the frequency distributions of coding categories across all the pages in a given country's information file. Table 1 shows that distribution for the file on party politics in Guinea, which contains 699 pages from thirty-four documents (Skogan 1967). The table shows, for example, that we have 134 pages in the Guinean files that deal with the party's "issue orientation" (code 530) but only 2 pages that discuss the way the party "raises funds" (code 260). A similar evaluation has been made for all nine countries processed for inclusion into the files to date (Janda 1967b). Our experience so far suggests that our missing data problems will be most severe for variables dealing with party organization and structure.

More relevant to the purpose of this paper is the separate question of the quality of the information that we do have in our files. We

Table 1: FREQUENCY DISTRIBUTIONS OF SUBSTANTIVE CODES USED IN INDEXING ALL 699 PAGES OF LITERATURE FOR GUINEA

FREQUENCY OF USAGE BY ASCENDING CODE NUMBERS

Code	Description	FREQ.	PCT.
000	Definition of a political party	1	0.1
010	Typology of political parties	3	0.2
020	Purpose of studying parties	2	0.1
100	When it was formed	11	0.8
110	Who formed it, base of support	7	0.5
120	Why was it formed	1	0.1
130	How was it formed	5	0.3
140	Political history of party	8	0.6
150	Organizational history of party	12	0.8
200	Selects candidates, party officials	5	0.3
210	Conducts election campaigns	7	0.5
220	Builds party policy and organi-zation	33	2.3
230	Influences government policy	17	1.2
240	Propagandizes its goals and activity	25	1.7
250	Discipline, maintains group unity	34	2.4
260	Raises funds	2.	0.1
270	Causes demonstrations, riots	2	0.1
280	Stands between gov't and citizen	14	1.0
290	Social activities of party	36	2.5
300	Party supporters	27	1.9
320	Party members	24	1.7
330	Party workers and activists	6	0.4
340	Party candidates	1	0.1
350	Party members in government posts	3	0.2
360	Party leaders and officials	120	8.3
370	Party factions	6	0.4
380	Organizational support	12	0.8
390	Group support	1	0.1
400	Local party organization	22	1.5
420	Regional party organization	17	1.2
430	National party convention, Congress	17	1.2
440	National party committee	34	2.4
460	Ancillary organizations	60	4.2
470	Party structure, functional or not	15	1.0
480	Articulation of party structure	22	1.5
490	Centralization of power	66	4.6
500	Gain control of government	6	0.4
520	Place members in government position	1	0.1
530	Issue orientation	134	9.3
540	Ideological orientation	37	2.6
550	Subvert the government	1	0.1
560	Efficiency and effectiveness	5	0.3
600	National crises	5	0.3
610	Issues of consensus or cleavage	14	1.0
620	Electoral system	6	0.4
630	Popular participation	7	0.5
640	Political norms and attitudes	24	1.7
650	Administrative bureaucracy	9	0.6
660	The executive	23	1.6
670	The legislature	11	0.8
680	Govt. structure, political history	88	6.1
690	Geographical allocation of authority	13	0.9
700	Economic environment	76	5.3
710	Geographical environment	10	0.7

FREQUENCY BY RANK-ORDER OF USAGE

RANK	CODE	FREQ.	PCT.
1	530	134	9.3
2	360	120	8.3
3	680	88	6.1
4	700	76	5.3
5	720	68	4.7
6	490	66	4.6
7	460	60	4.2
8	890	41	2.8
9	880	38	2.6
10	540	37	2.6
11	290	36	2.5
12	250	34	2.4
13	440	34	2.4
14	220	33	2.3
15	300	27	1.9
16	240	25	1.7
17	320	24	1.7
18	640	24	1.7
19	660	23	1.6
20	400	22	1.5
21	480	22	1.5
22	770	21	1.5
23	760	20	1.4
24	230	17	1.2
25	420	17	1.2
26	430	17	1.2
27	850	17	1.2
28	750	16	1.1
29	470	15	1.0
30	280	14	1.0
31	610	14	1.0
32	690	13	0.9
33	150	12	0.8
34	380	12	0.8
35	740	12	0.8
36	100	11	0.8
37	670	11	0.8
38	820	11	0.8
39	860	11	0.8
40	710	10	0.7
41	730	10	0.7
42	840	10	0.7
43	650	9	0.6
44	140	8	0.6
45	110	7	0.5
46	210	7	0.5
47	630	7	0.5
48	810	7	0.5
49	330	6	0.4
50	370	6	0.4
51	500	6	0.4
52	620	6	0.4
53	130	5	0.3
54	200	5	0.3

Table I (continued)

CODE		FREQ.	FCT.	RANK	CODE	FREQ.	PCT.
720	Social environment	68	4.7	55	560	5	0.3
730	Religious conditions	10	0.7	56	600	5	0.3
740	Social norms and attitudes	12	0.8	57	870	5	0.3
750	Activities of the military	16	1.1	58	10	3	0.2
760	Activities of the students	20	1.4	59	350	3	0.2
770	Activities of the trade unions	21	1.5	60	780	3	0.2
780	Volunteer activity, interest groups	3	0.2	61	830	3	0.2
800	Number of parties	1	0.1	62	20	2	0.1
810	Election results	7	0.5	63	260	2	0.1
820	Stability of parties in system	11	0.8	64	270	2	0.1
830	Interparty competition	3	0.2	65	0	1	0.1
840	Interparty cooperation	10	0.7	66	120	1	0.1
850	Origin, support, history of system	17	1.2	67	340	1	0.1
860	Status of the party in party system	11	0.8	68	390	1	0.1
870	Typology of party systems	5	0.3	69	520	1	0.1
880	International party system	38	2.6	70	550	1	0.1
890	Other parties	41	2.8	71	800	1	0.1

FREQUENCY DISTRIBUTIONS COMBINED INTO MAJOR CODING CATEGORIES

CODE		FREQ.	FCT.	RANK	CODE	FREQ.	PCT.
000	Definitions, functions, theory	6	0.4	1	400	233	17.5
100	How does a political party begin	44	3.1	2	700	236	16.4
200	What does a political party do	175	12.1	3	300	200	13.9
300	Who belongs to the party	200	13.9	4	600	200	13.9
400	How is the party organized	253	17.5	5	500	184	12.8
500	What does party seek to accomplish	184	12.8	6	200	175	12.1
600	Conditions-political environment	200	13.9	7	800	144	10.0
700	Conditions-social, econ, geographic	236	16.4	8	100	44	3.1
800	Conditions-party system	144	10.0	.9	0	6	0.4

approach the problem of data reliability (and validity) first by scoring each source document on a series of twenty-two "data quality" variables. Sample variables, which are analogous to Naroll's "control factors," are "place of publication, original language of source," "position of author," "source of data," "scope of study," "field research," and subjective judgments of the document's over-all "quality," "ideological orientation," and "objectivity." A complete listing of the data quality codes is contained in the Appendix. For the most part, these coding categories constitute no more than nominal scales, although some of the scaling is ordinal, with higher numbers implying higher ratings on those variables.

Each document is scored on the data quality variables by the same analyst who indexes it for our **MIRACODE** system. As a matter of routine, approximately 10 per cent of the pages in our files are reprocessed by a second analyst who reindexes the pages done by the first analyst and recodes his data quality variables for the same document.

Two special computer programs are then used to calculate both *interindexer* and *intercoder* reliabilities for the reprocessed material. In contrast to the rather low inter-indexer reliabilities around *.50* that have characterized our indexing of test for **MIRACODE** retrieval, we have encountered more generally acceptable intercoder reliabilities above .80 for data quality variables. (The distinction between "interindexer" and "intercoder" reliabilities is discussed in depth in Janda 1967b. There a lower level of reliability is defended as being "acceptable" for indexing than for coding.)

CONTROLLING DATA QUALITY: THE PROBLEM OF BIAS

Once the data quality variables are assigned to the documents and keypunched for computer processing, they are used in two ways to "control" or monitor the quality of information in the files. First, usage of the codes is tabulated for all the documents on party politics in a given country. This pro-

Table 2
DATA QUALITY CODES APPLIED TO 34 DOCUMENTS INDEXED FOR GUINEA

DATA QUALITY CODES	MOST FREQUENT:	(N)	2nd MOST FREQUENT:	(N)	3rd MOST FREQUENT:	(N)	4th MOST FREQUENT:	(N)
Document Type	Section in book	12	Journal article	11	News feature	5	News item	2
Place of Publication	United States	25	Guinea	3	Other	3	France	2
Original Source Language	English	29	French	S				
Position of Author	Academic	23	Party official	3	Journalist	3	Not applicable/no information	5
National Background	United States	24	Guinea	5	France	1	Geographical area	
Language Resources	Cites native sources	17	Native sources	5	Not applicable/no information	12		
Date of Data	1960-1964	18	1955-1959	9	Post-World War II	6	Not applicable/no information	1
Data Source Types	Government or party document	19	Secondary sources	19	Electoral or ecological data	4	Personal experience	4
Quantitative Analysis	No quantification	23	Raw data in context	7	One table	3	Two or more tables	1
Theoretical Treatment	No propositions	33	1 + propositions					
Traditional Scholarship	No footnotes	17	Between one and two per page	6	Less than one per page	6	Between two and three per page	4
Nature of Sources Cited	No footnotes	17	Primary source—party document	15	Primary source—private record	1	Secondary source—news	1
Scope of Study	Single country	19	Single party	5	Area survey	1	Comparison of parties	
Field Research	More than one year in country	10	None	8	Geographical area	5	Less than one year in country	4
Overall Judgment of Quality	Medium	16	High	11	Low	7		
Author's Ideology	Centrist	19	Leftist	5	Rightist	1	Not applicable/no information	9
Author's Objectivity	Objectivity not questioned	26	Value-oriented	6	Not applicable/no information	2		

vides a statistical profile of the file as given in Table 2, which shows the data quality summary of thirty-four documents indexed for Guinea. While this overview is useful for assessing the state of the literature on party politics in a given country, it can be misinterpreted because of its gross nature. Consider Skogan's comments on Table 2:

There appears to be a lack of extensive field research within Guinea itself. Although Table 2 lists "more than one year in country" as its most frequent document code, this is a reflection of work of one man, Victor DuBois, and the extensive contribution that his works have made to our files. Using authors rather than documents as our unit of analysis, we find that most do not appear to have spent much time in Guinea. (Skogan 1967)

Despite such possibilities for misinterpretation, this data quality summary has value for macro-level evaluations of the data in the files, although it does not readily allow for "control of errors" in Naroll's sense of "counteracting their effect on the results of the study."

The second usage of the data quality codes in controlling the quality of information obtained from the files—and certainly in keeping with Naroll's meaning of "control"—lies in the coding of individual parties on variables for comparative analysis. Here the data quality variables will facilitate decisions about the proper way to code or score a party on a given variable in the face of conflicting statements retrieved from our files by the MIRACODE system.

The MIRACODE equipment will retrieve varying numbers of pages containing information relevant to the coding of a given party on a given variable—e.g., "centralization of power" as measured on a three-point scale: "low," "medium," or "high." Insofar as the retrieved information is relevant to the decision, the researcher is expected, at the preliminary stage of coding, to score the party for each "hit" on the MIRACODE reader. As a result, he may record different scores on the party's coding sheet for "centralization of power" after searching all the material in the files.

When diverse sources disagree in statements about a party, we will seek to determine the basis of disagreement through a special analysis of variance computer program, which treats the data quality codes as independent variables predicting to varying party codes as the dependent variable. The program will try to identify the existence of systematic differences among data quality variables which account for variance in the dependent variable as coded from information in the files.

An example may clarify the procedure: The MIRACODE system may retrieve a total of thirty pages indexed for "party members" (code 320) pertaining to party X. Perhaps twenty-five of these pages would be relevant to coding the party on "severity of membership requirements." Assume that ten of these twenty-five report that membership in the party does *not* require the payment of dues, while fifteen other documents state that dues *is* a membership requirement. By analyzing the source of variance in our coding of this variable, we may discover that the discrepancy is explained by a data quality variable, e.g., "position of author"—with academics reporting no dues requirement and former party officials revealing that members are indeed required to pay dues to stay in good standing.

This example is offered only to illustrate the general procedure for using our data quality variables in "quality control" of the data we generate through library research. Even if we identify systematic sources of variance underlying disparate coding decisions, we have no method for "automatically" determining which coding decision is the *valid* one. With respect to judgments of validity, we are left in the age-old position of library researchers confronted with disagreeing sources; we use a variety of criteria, usually depending on the variable in question, to assess source validity. In the above example, we would probably decide to code "severity of membership requirements" according to the statements furnished by former party officials, who presumably constitute a "better" source for this variable than academics. Often, the disclosure of systematic differences among sources initiates focused analysis to resolve the discrepancy.

CONTROLLING DATA QUALITY: THE PROBLEM OF MEASUREMENT ERROR

In addition to guarding against the intrusion of bias or systematic error in the data generated through library research, there is

the additional problem of guarding against more or less random measurement error, which is especially vexing in the age of team research, computer analysis, and data banks. Holes in punch cards and magnetic spots on computer tape convey an awful definitiveness and finality—regardless of the uncertainty that may have attended the actual coding process. Once a coding decision is reached, the variable score is enshrined in paper or plastic for subsequent analysis and, usually, unquestioned acceptance. In the ICPP Project, however, we recognize and allow for the fallibility of our data by means of an "adequacy-confidence" scale, which expresses our evaluation of the quality of the data in our files that underlie each variable code. Our primary information resource for coding any party on any given variable will be the hundreds of pages we have indexed and microfilmed on party politics in the country under concern. Obviously, the literature will vary in its adequacy for providing information with which to make coding judgments, and our analysts will have more confidence in coding some variables than in coding others. We intend to reflect the adequacy of the documentation underlying any given variable and party and our analysts' confidence in their coding judgments by accompanying each variable with an "adequacy-confidence" rating, as scored by those who coded the variable.

Every variable for every party will be coded independently by each member of the two-man research team that has read and indexed the literature in the file for that country. The *variable* code that is eventually keypunched for statistical analysis derives from their independent coding operations. When the coders agree on a variable code, that code will obviously be entered for the variable. When they disagree over coding the variable, an attempt will be made to resolve their disagreement through discussion, involving outside coders if necessary. The *adequacy-confidence* code that is assigned to the final variable code also derives from both analysts' adequacy-confidence codes, which are independently assigned when the variable is coded. Intercoder conferences and involvement of outside coders are also used to resolve disagreements in rating variables on the adequacy-confidence scale.

The adequacy-confidence scale was designed to reflect four factors that seem especially important in determining the researcher's belief in the accuracy or truth value of the coded variable-as well as can be determined through library research. These factors are (1) the number of sources that provide relevant information for the coding decision, (2) the proportion of agreement to disagreement in the information reported by different sources, (3) the degree of discrepancy among the sources when disagreement exists, and (4) the credibility attached to the various sources of information.

The first three factors deal more with the "adequacy" of the literature that can be cited to document the variable code, and the fourth deals more with the analyst's confidence in coding the variable. In an effort to "objectify" our measure of the researcher's belief in the accuracy or truth value of the coded variable, we have operationalized the adequacy-confidence scale primarily in terms of the first three factors: (1) number of sources, (2) proportion of agreement, and (3) degree of discrepancy. However, this operationalization is intended only to guide the researcher in arriving at his adequacy-confidence rating when the fourth factor (source credibility) is held constant across documents. If the credibility factor, ignored in our operationalization, interacts sufficiently with the information sources to cause the researcher to be more or less confident in his coding than the operationalization formula would suggest, then he is free to revise the adequacy-confidence rating accordingly.

The credibility factor is kept out of the operationalization due to the great difficulty in fashioning an acceptable scale for a position in n-dimensional attribute space, created from the several subfactors contributing to source credibility, of which three seem especially important: (1) amount of attention given to the variable in the source, (2) adequacy of the research underlying the author's observation, and (3) the integrity and objectivity attributed to the author. These three factors, and certainly others, can interact in a variety of ways to affect the researcher's evaluation of source credibility, and we have not attempted to spell out rules for handling the combinations and subtleties involved in any such evaluation. Instead, we are constrained to leave source credibility operate as a subjective variable in tempering the re-

searcher's belief in the truth value or accuracy of the variable code after reference to the more objective operationalization.

In general, if the "credibility gap" between sources is not great, it is expected that the researcher will score the coding judgment according to the objective operationalization of the adequacy-confidence scale. But when he feels that the credibility of the sources is such that straightforward application of the operational definition results in a confidence code value that does not reflect his own belief in the truth value or accuracy of the variable code, then he should revise his adequacy-confidence code accordingly.

To guide the researcher in interpreting the graduations in the adequacy-confidence scale, a conceptual definition of each scale category is presented in Table 3 with the operationalizations of the coding categories.

The analysis of variance approach discussed earlier is in order only for variables that rated from "2" to "5" (and possibly "7") on the adequacy-confidence scale. Other scale values suggest a lack of disagreement within the literature, leaving no "variance" to be explained by the data quality codes through the analysis of variance model. Even for relevant adequacy-confidence codes, the analysis of variance approach is useless in detecting error if the observations in the literature are too few to support a statistical analysis. In these cases—which may turn out to be most cases—the adequacy-confidence scale is used in two less elegant methods for "control of errors" in Naroll's sense of "counteracting their effect on the results of the study."

The first and simpler method is to study scatter diagrams or contingency tables for the presence of deviant cases as disclosed by distance from the regression line or entries in cells off the diagonal. Assuming that the diagrams or tables are constructed to show

Table 3
ADEQUACY-CONFIDENCE SCALE

Code	Category label	Conceptual definition	Operational definition
BLANK	**Inapplicable**	Variable does not apply to the party coded	
1	**Inadequate**: no data	No information is contained in the file on the variable being considered	
2	**Inadequate**: disagreement	Disagreements ace found in the file which are not resolvable by reference to source credibility. The disagreement might be resolved by more data, but the information in the file in inadequate	
3	**Barely adequate**: lowest confidence	It is possible to cite this code as the most probable among alternatives, but further research could easily produce a finding at great variance from this one	Two situations ran produce this code: (I) There is a 1:1 division between sources with a "great" discrepancy in the suggested codes, but one code can be favored on the basis of source credibility. (2) Data are incomplete in some way, but a code can be inferred from available information
4	**Adequate**: low confidence	There is a disagreement in the literature which suggests that the code might not be supported by further research, although the alternative is rot greatly discrepant	There is a 1:1 division between sources with a 'medium' discrepancy* in suggested codes, but one code can be favored on the basis of source credibility
5	**Adequate**: low to medium	There is no strong agreement in the literature on this particular code, but further research is likely to support the code or one close to it	Three situations can produce this code: (I) No source has complete information, but a summary code can be made from data from two or more incomplete sources. (2) There is a 2:1 division between sources without regard to degree of discrepancy. (3) There is a 1:1 division between sources with a 'small' discrepancy' in suggested codes, but one code can be favored on the basis of source credibility
6	**Adequate**: medium confidence	The rode Is not extensively documented is the literature, but there is no disagreement in evidence. Further research would likely support the code, but there are no strong grounds to rule out possible disagreement	One source cites the summary code with no disagreement in evidence
7	**Adequate**: medium to high	Although the code is quite well documented, the judgment is placed in some doubt because it is nut unanimous. Disagreements might occur in further research, bat the code would likely be supported	There is at least a 3:1 division between sources, without regard to the degree of discrepancy, and the overwhelming evidence favors the code
8	**Adequate**: high confidence	Since documentation of the code is good and no disagree. ments are apparent, it is probably accurate, although additional documentation is desirable	Two sources agree on the code and no source disagrees
9	**Adequate**: highest confidence	The variable code is extremely well documented and no disagreements are apparent; belief in the accuracy of this code is about as high sauna could expect in the absence of original field research	Three or more sources agree on the code and no source disagrees

*The degree of discrepancy is applicable only to variables of an ordinal or interval nature. Whether a discrepancy is classified as "small," "medium" or "great" depends on the particular variable and is established in the operational definitions for each variable, which must be referred to in order to determine or interpret the degree of discrepancy.

the relationship between two variables linked by theory, the presence of deviant cases suggest either measurement error or exceptions to the theory. By examining the cases for their adequacy-confidence scale scores, which can be displayed by appropriate computer routines, the researcher might be able to determine if apparent exceptions to his theory rate low on the scale and represent probable measurement error or if the data seem solid and the theory dubious.

The second method of controlling error involves a "stepwise" approach to the calculation of correlation coefficients. By means of flexible computer programs for including and excluding cases from analysis on the basis of their adequacy-confidence scores, correlations can be calculated first for the "best" data, then again for progressively larger sets of data as the quality restriction

is relaxed. Assuming that measurement error (as expressed by the adequacy-confidence scale) is random and the hypothesized relationship is true, smaller correlation coefficients should be generated from each progressive relaxation of data quality. If the correlations should happen to increase, serious attention should be given to bias among data at the lower end of the adequacy-confidence scale.

Although problems inherent to library research are not unique to the ICPP Project, the scope of our activities is such that we must develop systematic procedures for evaluating the information that resides in and emerges from our files. We have adopted the methodology of data quality control to help us cope with the problem of data reliability.

APPENDIX

Data Quality Control Codes

Columns	Variable

1-18   SENIOR AUTHOR'S LAST NAME AND IN!-TIALS

19-20   YEAR OF ORIGINAL PUBLICATION

21-23   COUNTRY CODE

24-26   DOCUMENT CODE

27-29   INDEXER CODE

30 TYPE OF DOCUMENT

    0   not otherwise classified
    1   reference *source-Facts On File,* Keesings Archives, etc.
    2   newspaper or magazine item-popular periodical
    3   newspaper or magazine feature story-popular periodical
    4   party document-constitution, platform
    5   government documents-reports, statistical abstracts
    6   journal article
    7   article or chapter in book (used for reprints of journal article)
    8   thesis or monograph
    9   book

31-32   PERIODICAL CODE-specific for each country

33 PLACE OF PUBLICATION

    blank don't know (missing data)
    0   not otherwise classified (use also when not applicable)
    1   United States (except if 2 is applicable)
    2   in colonizing country (U.S., Britain, France, Germany, Spain, Portugal, Netherlands)
    3   in area of world where country exists-i.e., Latin America, Africa, Europe, Asia
    4   in country studied

34 ORIGINAL LANGUAGE OF SOURCE

    0 not otherwise classified
    1   English
    2 French
    3   Spanish
    4 German
    5   language of country studied (if two apply, favor using this code)

35 AUTHORSHIP

    0 no author named
    1   one author
    2   two authors
    3   three or more authors
    4   corporate author (e.g., Bulgarian National Committee)

36 POSITION OF FIRST-NAMED AUTHOR

(favor higher code if two apply)

blank no information (missing data)

0 not otherwise classified

1 journalist

2 government official in country studied

3 ex-government official

4 party official in country studied

5 ex-party official

6 academic

37 **PRESUMED NATIONAL BACKGROUND**- judged from last name and source of publication

blank not applicable-no author given

0 no judgment made/not otherwise classified

1 United States (except if 2 is applicable)

2 from colonizing country-U.S., Britain, France, Germany, Spain, Portugal, Netherlands

3 from area of world where country exists—e.g., Latin America, Africa, Europe (use if in doubt of 4)

4 from country studied

38 **EVIDENCE** OF USE OF LANGUAGE RESOURCES

blank not applicable (use for general theory, not Country studied)

0 no information

1 coder infers author has no ability in native language

2 cites translated materials, worked with interpreter

3 cites native language sources, uses native language phrases in text (excluding the native names of political parties)

4 uses native interviewers to collect survey information

5 document itself translated from native language or written by native in English

39 **DATE OF MAJOR PORTION OF DATA**

(code later period if other choice cannot be made)

blank not applicable (use for general theory)

0 not otherwise classified

1 prior to World War II (1939 or earlier)

2 1940-1944

3 1945-1949

4 1950-1954

5 1955-1959

6 1960-1964

7 1965-present

8 post-World War II (give preference to above categories)

40 **NOT USED**

41-49 **CODE FOR DATA SOURCES (entered in** columns 41-49, ranked by importance)

blank not applicable (use for speeches, election reports, etc.)

0 no data sources revealed

1 not otherwise classified

2 secondary sources-newspapers, books, journals, broadcasts

3 government publications or party documents

4 election returns or ecological data

5 roll call votes

6 sample survey of individuals

7 interviews with party officials or leaders

8 personal experience as participant observer

50 **NUMBER OF DATA SOURCES USED**

51 **NOT USED**

52 **QUANTITATIVE ANALYSIS SCORE**

0 no quantification involved

1 raw data or per cents reported in text but not in tables

2 one raw data or percentage table reported

3 two or more raw data or percentage tables reported

4 bivariate measures of association reported

5 multivariate statistics reported

53 **THEORETICAL TREATMENT SCORE**

0 no explicit propositions advanced or tested

1 general theory that discusses "relevant" variables, but does not state relationships among them

2 one or more explicit propositions advanced but not statistically tested

3 one or more explicit propositions advanced and statistically tested

4 enumeration of three or more propositions with common concepts into a body of theory

5 incorporation of three or more propositions with common concepts into a body of theory

54 **TRADITIONAL SCHOLARSHIP SCORE**

blank not applicable (speeches, election returns)

0   no footnotes cited or attribution of sources
1   less than 1 footnote per page
2   between 1 and 2 footnotes per page
3   between 2 and 3 footnotes per page
4   more than 3 footnotes per page

55   **NATURE OF SOURCES CITED IN FOOT-NOTES** (enter the highest when appropriate)

blank not applicable-no footnotes
0 not classified
1 tertiary sources-encyclopedias, references only
2 secondary sources-newspapers and magazines
3 secondary sources-books, journal articles
4 primary sources-party and government documents
5 primary sources-personal records, memoirs, interviews, data from unpublished sources

56   **CITATION OF DUVERGER (enter** highest applicable)

blank not applicable-no footnotes in text
0 footnotes, but none to Duverger
1 one footnote to Duverger
2 two or more footnotes to Duverger
3 mentions Duverger in the text
4 tests out Duverger's propositions or theory, modeled after Duverger's analysis, uses Duverger's concepts or "branch" and "caucus" parties, "majority bent" parties, etc.

57   **SCOPE OF STUDY** (use for whole document whether all is coded or not)

0 not otherwise classified
1 conceptual or theoretical, without emphasis on data and evidence
2 survey of parties or politics in given area, e.g., Latin America
3 comparative analysis of governmental systems
4 comparative analysis of political parties
5 study of a single country

6   study of a single party
7   news event

58-60   **FOCUS OF STUDY-MOST FREQUENT SUBSTANTIVE CODING CATEGORY USED**

61-63   **NUMBER OF TIMES MOST FREQUENT SUBSTANTIVE CODING CATEGORY USED**

64-66   **FOCUS OF STUDY-SECOND MOST FREQUENT SUBSTANTIVE CODING CATEGORY USED**

67-69   **NUMBER OF TIMES SECOND MOST FRE-QUENT SUBSTANTIVE CODING CATEGORY USED**

70   **FIELD RESEARCH**

blank not applicable or no information
0   evidence of no work in country studied
1   evidence of work in geographical area
2   spent less than one year in country
3   spent more than one year in country, or two trips of any length, or author writing in country
4   author a nonresident native of a country
5   author a resident of country

71   **CODER'S SUBJECTIVE JUDGMENT OF QUALITY OF SOURCE**

1   low
2   medium-code unless evidence points to high or low
3   high

72   **CODER'S SUBJECTIVE JUDGMENT OF IDEOLOGICAL ORIENTATION OF AUTHOR**

0   not classified on left-right dimension
1   leftist
2   centrist-code unless evidence points to low or high
3   rightist

73   **CODER'S SUBJECTIVE JUDGMENT OF AUTHORS OBJECTIVITY**

1   antiseptically objective—e.g., "scientific" analysis, mainly tabular presentation of data
2   no reason to doubt objectivity
3   values detectable
4   emotional language

74-76   **NUMBER OF PAGES CODED**

78-80   **CODING TIME IN MINUTES**

BIBLIOGRAPHY

JANDA, KENNETH

1967a Political research with **MIRACODE** a 16-mm. microfilm information retrieval system. *Social Science information* 6: 169-181.

1967b *A microfilm and computer system for analyzing comparative politics literature.* Paper delivered at the National Conference on Content Analysis, The Annenberg School of Communications, University of Pennsylvania, Philadelphia, November 16-18.

1968 Retrieving information for a comparative study of political parties. In William J. Crotty, ed., *Approaches to the study of party organization.* Boston, Allyn and Bacon. Reprinted in *information retrieval: applications to political science.* Indianapolis, Bobbs.Merrill.

KOH, HESUNG CHUN

1966 A social science bibliographic system: orientation and framework. *Behavior Science Notes 1:145-163.*

MURDOCK, GEORGE P., *et al.*

1961 *Outline of cultural materials.* New Haven, Human Relations Area Files.

NAROLL, RAOUL

1962 *Data quality control: a new research technique.* Glencoe, Ill., Free Press.

SKOGAN, WESLEY

1967 *Bibliography on party politics in Guinea, 19S0-1952.* In Kenneth Janda, ed., *ICPP Bibliography Series.* Evanston, International Comparative Political Parties Project, Northwestern University.

TEXTOR, ROBERT B.

1968 *A cross-cultural summary.* New Haven, Human Relations Area Files.

AUTHOR'S NOTE

Data on how 158 political parties were scored on the adequacy-confidence codes for 100 variables are reported in Kenneth Janda, *Political Parties: A Cross-National Survey* (New York: The free Press, 1980), Chapter 2, pp. 12-18. This chapter is available online at

www.janda.org/ICPP/ICPP1980/Book/PART1/Ch.02_VariableCoding/Ch.02p12.htm