Benchmark Comparisons of Mainframe versus Microcomputer Analysis of Social Science Data Using *SPSS/PC*

Josefina J. Card and Sabrina Lin

Background: The Data Archive on Adolescent Pregnancy and Pregnancy Prevention

The rate of pregnancies and out-of-wedlock childbearing among U.S. teenagers is among the highest in the world. Every year the lives of several hundred thousand young people are changed by an unplanned event. Several federal agencies—the National Institute of Child Health and Human Development, the Office of Adolescent Pregnancy Programs, the Bureau of Health Care Delivery and Assistance, and the Office of Child Development—and private foundations—Ford, Rockefeller, Hewlett, Mott—have poured millions of dollars into research and demonstration programs aimed at understanding and combatting this problem. The rate of unwanted pregnancies is also high among socioeconomically disadvantaged individuals, and government-sponsored family planning programs have been developed to help serve these individuals' needs.

Josefina J. Card is president of Sociometrics Corporation in Palo Alto, CA. She received her Ph.D. degree in social psychology from Carnegie-Mellon University in 1971 and from 1973–1983 served on the research staff of the American Institutes for Research in the Behavioral Sciences. Sabrina Lin is a research associate at the Sociometrics Corporation. Holding a Master's degree in statistics from Stanford, she is now a graduate student in psychology at that university.

The authors wish to acknowledge the many contributions of the Mainframe-DAAPPP archivists, Drs. Phil Ritter and Starr Silver. Dr. Silver put us in touch with this valuable new journal. In addition, she and Dr. Stuart K. Card of the Xerox Palo Alto Research Center reviewed an earlier draft of this paper and made helpful suggestions for improvement. We thank them for their input.

This paper was produced with funds provided by Small Business Innovation Research Grant No. APR 000923-01-0 from the Office of Population Affairs. For additional information on the procedures and products of the Data Archive on Adolescent Pregnancy and Pregnancy Prevention, contact the authors at Sociometrics Corporation, 3191 Cowper Street, Palo Alto, CA 94306 (415-321-7846).

Social Science Microcomputer Review 3:4, Winter 1985. CCC 0885-0011/85/$1.50.

The Data Archive on Adolescent Pregnancy and Pregnancy Prevention (DAAPPP) has been established by the Office of Population Affairs, the federal agency responsible for the coordination of population-related activities, to encourage research on these timely topics of adolescent pregnancy and family planning service delivery. DAAPPP's goal is to make available to the community of scholars those extant data most capable of helping these fields advance. Such an archive should have many benefits. For little additional cost (relative to the data collection costs already incurred by the exemplary studies): (1) It should accelerate the growth of scientific knowledge about adolescent pregnancy, pregnancy prevention, and family planning; (2) It should provide practitioners, service providers, administrators, and policymakers with a larger scientific base on which to build their work; (3) It should facilitate the dissemination of model approaches to combatting the problems of adolescent pregnancy, pregnancy prevention, and family planning; (4) It should encourage investigations by new investigators with different perspectives and professional backgrounds: for example, individuals interested in the problems of adolescent pregnancy and family planning, who have first-hand knowledge about these problems, but who find themselves in circumstances where they are not capable of doing independent research (because they do not have access to research funds or available data). In short, the archive should accelerate the amount of research being done on the problems of adolescent pregnancy, pregnancy prevention, and family planning; accentuate common findings relating to ways of combatting these problems; enlarge the body of researchers working on the problems; and in these ways help advance the field. These advances should be accomplished in an exemplary, cost-effective manner, because data collection costs will not have to be duplicated, and because comparison of results across multiple investigations will be greatly facilitated.

The DAAPPP collection, which will eventually consist of data and documentation from 80 to 100 of the most outstanding relevant studies, is being chosen with the help of a National Advisory Panel of experts in the fields of adolescent pregnancy, family planning, and data collection and storage. Each database in the collection is being processed and documented in a standard way prior to inclusion to facilitate use. Machine-readable program statements are provided that transform the raw data into a system file capable of being analyzed with *SPSS-X* on mainframe computers or *SPSS/PC* on microcomputers. A user's guide that describes the data set, its origins, strengths, and weaknesses is also provided. The entire collection will eventually be accompanied by an educational package that will allow the archive to be used not only for research purposes but for instructional purposes as well. At the moment, machine-readable data, documentation, and *SPSS/SPSS-X* program statements (for mainframe use) are available for the 52 data sets listed in Table 1.

Table 1 List of data sets currently available from DAAPPP

Data Sets Processed and Documented by DAAPPP Staff

1 1971 U.S. National Survey of Young Women: Selected Variables
2 1976 U.S. National Survey of Young Women
3 Project TALENT: Consequences of Adolescent Childbearing for the Young Parents' Future Life, 1960–1974
4 Detroit Mother-Daughter Communication Patterns: Mother File, 1978
5 Detroit Mother-Daughter Communication Patterns: Daughter File, 1978
6 Philadelphia Collaborative Perinatal Project: Economic, Social, and Psychological Consequences of Adolescent Childbearing, 1959–1965
7 Nashville General Hospital Comprehensive Child Care Project, 1974–1976: Selected Variables
8 State Policy Determinants of Teenage Childbearing, 1979
9 1980 U.S. Survey of Services Provided by Adolescent Pregnancy Programs
10 1982 Evaluation of OAPP Adolescent Pregnancy Programs
11 1980 U.S. Current Population Survey: Selected Variables—Women
12 1980 U.S. Current Population Survey: Selected Variables—Men
13 1980 U.S. Current Population Survey: Selected Variables—Children
14 1982 U.S. Current Population Survey: Selected Variables—Women
15 1982 U.S. Current Population Survey: Selected Variables—Men
16 1982 U.S. Current Population Survey: Selected Variables—Children
17 1977 U.S. Current Population Survey: Selected Variables—Women
18 1977 U.S. Current Population Survey: Selected Variables—Men
19 First U.S. Health and Nutrition Examination Survey, (HANES), 1971–1975
20–24 National Longitudinal Study of Youth (NLSY), 1979–1982: Selected Variables (Waves 1–4), and Supplementary Variables
25 1981 U.S. Survey of Title X-Funded Family Planning Clinics
26 1982 National Survey of Family Growth (NSFG), Cycle III (Women Aged 15–44)
27 1982 National Survey of Family Growth (NSFG), Cycle III (Women Aged 15–19)
28 1979–1980 U.S. Survey of Unmarried Women Under 18 in Family Planning Clinics
29 Effects of Organized Family Planning Programs on U.S. Adolescent Fertility
30 Johns Hopkins Study of Repeat Adolescent Pregnancy, 1976–1982
31 1972–74 Ventura County Survey of Unmarried Pregnant Women Aged 13–20
32 1982 San Jose, Calif. Study of Adolescent Perinatal Risk Behavior
33 1981–1982 Evaluation of OAPP Adolescent Pregnancy Programs: Individual Level Data I
34 1981–1982 Evaluation of OAPP Adolescent Pregnancy Programs: Individual Level Data II
35 1979–1981 Philadelphia Study of Psychological Factors Associated With Adolescent Fertility Regulation—Females
36 1979–1981 Philadelphia Study of Psychological Factors Associated With Adolescent Fertility Regulation—Males
37–38 The National Survey of Children, 1976
39 Florida–Puerto Rico Study of Adolescent Pregnancy and Neonatal Behavior, 1978
40 Maricopa County, Arizona Study of Child Maltreatment Risk Among Adolescent Mothers, 1976–1978
41 1955 Growth of American Families: Married Women
42 1955 Growth of American Families: Single Women
43 1960 Growth of American Families
44 1979 U.S. National Survey of Young Women
45 1979 U.S. National Survey of Young Men

Data Sets Acquired from the Western Washington Archive on the Antecedents of Teenage Pregnancy

1 Adolescent Sexual Behavior: Context and Change
2 Contraceptive Use Effectiveness: The Fit between Method and User Characteristics, Study 1
3 Contraceptive Use Effectiveness: The Fit between Method and User Characteristics, Study 2
4 Comparison of Participants and Dropouts from a Teen Contraceptive Program
5 Psychosocial Factors in Adolescent Pregnancy
6 Adolescent Development and Contraceptive Use
7 Adolescent Socialization and Heterosexual Behavior

Benchmark comparisons of Data Analysis using Mainframe-DAAPPP vs. Micro-DAAPPP

To investigate the feasibility and utility of creating the microcomputer version of DAAPPP and to assess the relative (time, cost, and convenience) benefits and costs of using Micro-DAAPPP versus Mainframe-DAAPPP, we conducted a series of benchmark analyses on three prototypical data sets of varying size (in terms of number of variables and number of cases):

—DAAPPP Data Set No. 8: State Policy Determinants of Teenage Childbearing (171 variables; 52 cases);
—DAAPPP Data Set No. 25: Survey of Title X Family Planning Agencies (102 variables; 351 cases);
—DAAPPP Data Set No. 27: National Survey of Family Growth, Women Aged 14–19 (246 variables; 1,888 cases).

We conducted a range of parallel analyses on these data bases, including descriptive and categorical analyses (FREQUENCIES and CROSSTABS), comparison of groups analyses (ONEWAY and ANOVA), multivariate analyses (CORRELATION and REGRESSION), non-parametric analyses (NPAR TESTS), and utility procedures (SORT CASES). To study the effect of sample size on time required to do microcomputer statistical analysis, we created two analysis files from Data Set 27: a file with data from a subset of 1,000 cases (henceforth 27a), and a simulation file consisting of file 27a copied to yield 2,000 cases (henceforth 27b).

The time required to run the various benchmark analyses using *SPSS/PC* with an 8087 math coprocessor is given in Table 2. For both Data Set Nos. 8 and 25, the majority of analyses took less than a minute to run. For Data Set No. 27a, analyses took, on the average, between 2 and 3 minutes. Analyses with Data Set No. 27b took about twice as long as identical analyses run on 27a, showing that time to perform comparable analyses with *SPSS/PC* goes up proportionately with sample size. Overall, only an impressively small amount of time (well under six minutes) was required to receive results for the full range of "typical" analyses.

The only exceptions to this rapid turnaround were unusual runs that would not be run by a typical user with any regularity. For example, it took between 5 and 46 minutes to get FREQUENCIES for all variables in the files studied. A user is bound to do such a comprehensive run only once, prior to conducting the specific analyses of interest. Similarly, the SORT procedure, one not generally used by social scientists, took under a minute for the smallest data set but 37 minutes for a 5-way sort on the largest data set.

The significance of the cost savings involved by doing analyses on a microcomputer versus a mainframe can be gleaned from Table 3. Table 3 gives actual costs incurred when we tried to do illustrative "simple" and "exhaustive" analyses on the Stanford University mainframe computer. Before analyses could even be undertaken, we

Table 2 Time required for *SPSS/PC* analyses

Analyses conducted	Data set number 8 (171 vars, 52 cases)	25 (102 vars, 351 cases)	27a (199 vars, 1000 case subset)	27b (199 vars, 2000 case simulation file)
Descriptive and categorical procedures				
1. FREQUENCIES				
All variables	22:52	4:48	23:55	46:36
6 variables	0:10	0:38	2:24	5:22
2. CROSSTABS				
4 or 5-way	0:16	0:38	2:11	4:18
3-way	0:11	0:33	2:27	4:50
Comparison of groups procedures				
1. ONEWAY				
Many groups	0:11	0:25	2.10	4:15
Few groups	0:11	0:25	2:10	4:15
2. ANOVA				
Multi-way	0:49	1:07	3:04	5:11
2-way	0:11	0:27	2:10	4:11
Multivariate procedures				
1. CORRELATION				
100 variables	0:56	4:09	2:54	5:45
6 variables	0:14	0:33	2:12	4:21
2. REGRESSION				
10 predicators	0:14	0:37	2:12	4:21
1 predictor	0:14	0:33	2:12	4:21
Non-parametric procedures				
1. NPAR TESTS				
Median Test	0:10	0:29	2:18	4:35
Mann-Whitney Test	0:10	0:29	2:12	4:20
Utility procedures				
1. SORT CASES				
5 categories	0:39	2:55	16:44	37:12
1 category	0:39	2:55	13:08	28:00

had to pay for costs to store the data on disk (costs which come to $1.24 per month for the smallest data set, $17.49 for the largest). We could have chosen to keep the data on a mainframe tape to avoid such storage charges. We would then have had to pay over $5.00 to download the data from tape to disk every time we accessed the data! Costs to perform one illustrative "simple" analysis (a two-factor, one-way analysis of variance) ranged from $3.45 to $8.48. Costs to perform one illustrative "exhaustive" analysis ranged from $10.71 to $36.71.

In addition to these cost savings, other benefits ensued from microcomputer use. We performed our microcomputer analyses from the convenience of our own office, and results were printed on our printer a foot away from where we were sitting. In contrast, we had

Table 3 Illustrative costs saved by performing data analyses on a microcomputer

	Data set number 8[a]	25[b]	27[c]
Data storage and access costs			
Monthly disk storage	1.24	$ 2.31	$17.49
or tape download (per request to download data from tape to disk to avoid disk storage charges)	$ 5.71	$ 5.71	$ 6.09
*Data analyses costs**			
Illustrative simple analysis: One-way analysis of variance with 2 variables	$ 3.45	$ 3.98	$ 8.48
Illustrative exhaustive analysis: Frequencies for all variables in file	$36.71	$10.71	$24.11

[a]171 variables, 52 cases
[b]102 variables, 351 cases
[c]199 variables, 1,888 cases
*These cost estimates do not include costs incurred while setting up the analyses at a terminal, nor do they include costs incurred by each failed execution attempt.

to access Stanford's mainframe via a terminal; results were printed on Stanford's printer two miles away. A staff member then had to be dispatched to Stanford to get the printouts. Thus courier (or data-retrieval) costs should be added to the data storage, access, and analysis costs given in Table 3.

SPSS/PC is an interactive system, we were free to correct mistakes or fine-tune desired runs immediately. In contrast, we had to wait much longer (the exact length of the wait varying as a function of how much money we were willing to pay to raise the "priority status" of our job) for mainframe results. Table 4 lists mainframe cost rates used in this study for comparison purposes.

Table 4 Mainframe service rates in benchmark study

	Day (M–F 6 AM–6PM)	Other (M–F 6 PM–6 AM Weekends–Holidays)
Processing*	$ .62/second	$ .27/second
Tape mount	$ 2.00/mount	same
Tape storage	$ 3.00/month	same
Cards read	$ 1.00/thousand	same
Cards punched	$10.00/thousand	same
Card images submitted	$ .10/thousand	same
Telenet/Tymnet connect	$ 8.50/hour	same

*Interactive and batch are expressed in 3081-D equivalent seconds
Disk Storage: $0.0085/kilobyte per month (1 track = 47 kilobytes and 1 block = 2 kilobytes)

Summary and Conclusions

Our conclusion is that Micro-DAAPPP is not only feasible, but convenient, useful, and cost-effective as well. Micro-DAAPPP will be the first comprehensive data archive ever to be constructed for personal computers. Micro-DAAPPP will also be the first data archive ever to be sold as a self-contained package of data, documentation, instructions for use, and educational exercises. These innovations should further research on the problems of family planning and adolescent pregnancy, provide an invaluable resource for the teaching of statistical data analysis, enlarge the market for the Data Archive on Adolescent Pregnancy and Pregnancy Prevention, decrease costs associated with acquiring and using DAAPPP data, facilitate the ease and convenience with which even nonseasoned data analysts and researchers can access and use DAAPPP data and ensure DAAPPP's useful life for many years to come.