HOUSING ASSISTANCE SUPPLY EXPERIMENT

A WORKING NOTE

This Note was prepared for the DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT, under Contract No. H-1789. It is intended to facilitate communication of preliminary research results. Views or conclusions expressed herein may be tentative and do not represent the official opinion of the sponsoring agency.

25th Year
Rand
SANTA MONICA, CA. 90406

WN-7885-HUD

DATA MANAGEMENT SYSTEM: PART I, FIELDWORK DATA AND DATA TRANSFER SPECIFICATIONS

G. Levitt

July 1972

This Note was prepared for the DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT, under Contract No. H-1789. It is intended to facilitate communication of preliminary research results. Views or conclusions expressed herein may be tentative and do not represent the official opinion of the sponsoring agency.

Rand
SANTA MONICA, CA. 90406

PREFACE

This Working Note was prepared for the Office of Research and Technology (ORT) of the U.S. Department of Housing and Urban Development, to describe the procedures and processes by which data will be prepared for analysis in the Housing Assistance Supply Experiment.

This Note is the first in a series describing a data management system for the Supply Experiment which provides procedures and facilities for acquiring, organizing, storing, and retrieving data generated by field surveys, enrollment and disbursement operations, and other data sources.

This Note fulfills the requirements set forth in Sec. II.B, Phase I, Task 5, subparagraphs (1) and (3) of HUD Contract H-1789.

The author would like to acknowledge the contributions of Sandra Barry, Elaine Renner, and Zahava Blum. The present draft was edited by Janet DeLand.

CONTENTS

I. INTRODUCTION

This document describes the procedures and specifications that will be used to achieve and maintain high-quality data for analysis in the Housing Assistance Supply Experiment.

The major data source for the Supply Experiment will be formal surveys conducted at each of two sites.* These survey data will be generated continuously over the course of the experiment, beginning with extensive baseline collections and continuing with annual and semiannual follow-ups on the experiment participants. Additional data will be collected on the occurrence of certain other scheduled and unscheduled events, such as participant dropouts and reentries, participant movement between areas with different monitoring protocols, and certain by-products of enrollment and disbursement procedures. Other less structured data will be generated through informal observation by site monitors and others participating in fieldwork activities.**

All data collected will eventually be put into computer readable form. This process will comprise a set of integrated steps which include data validation, data coding, data digitizing, data cleaning, and data storage. This document describes each of these steps, their control and coordination, and their relationship to the task of providing high-quality data for analysis.

---

*For additional information, see WN-7833-HUD, *Site Selection for the Housing Assistance Supply Experiment: Stage I*, Housing Assistance Supply Experiment Staff, The Rand Corporation, May 1972.

**Exact sample sizes for the several field surveys are not yet fixed, and will not be until specific experimental sites are selected. Our monitoring plans imply roughly the following numbers for the baseline surveys:

Household interviews (renters and homeowners)	5,000
Building evaluations (one per single or multiple dwelling)	3,500
Landlord interviews (one per rental structure)	1,500

After baseline, these surveys will be repeated annually. In addition, a record will be created by an office interview on each household enrolled in the allowance program, updated semiannually. We anticipate possibly 10,000 enrollees. Enrollment will begin after baseline surveys are completed.

Current plans assume that only field interviewing and quality control will be subcontracted. The responsibility for the other field activities (coding, keypunching, and data cleaning) remains unassigned. Because some--and perhaps all--of the work described herein will be subcontracted rather than done at Rand, we shall not go into extensive detail on the format and content of training manuals, codebooks, and operational procedures. Such material will be developed and described either in a revision of this report or by the subcontractor (with Rand's assistance).

II. DATA CONTROL

To maintain a high-quality stream of data, strict control must be exercised over the data through each of the different processing steps. A high level of control is necessary to minimize and eliminate (where possible) potential sources of errors, to protect respondent confidentiality, and to increase data system operational efficiency in terms of providing timely as well as clean data.

Rand and its subcontractor(s) will be jointly responsible for creating and maintaining a data control system whose major function will be to track the movement of data accurately. Such a system will enable the rapid location of any data currently being processed or stored and will make evident any loss or misplacement of data.

A log will be maintained at each of the different processing sites, including the field survey sites, data preparation sites, and analysis sites. Information to be logged will include document identification, the names of the persons possessing each document, the dates on which possession was assumed and released, and the reason for having possession. In addition, each survey instrument will have a face sheet attached to it on which each person assuming and releasing control will sign on and sign off.

The data control system will also be used to *monitor* operational efficiency. Statistics will be generated from error frequency and timing data logged by individuals. These data will be used to tune data system operations by indicating bottlenecks, locating poorly performing individuals, and detecting poorly defined operational procedures.

To insure cooperation, supervisors at each of the processing sites will be responsible for validating individual compliance with control procedures. Equally important, operational personnel will have to be convinced that the control system exists as a support aide, rather than as a "big brother."

Because the control and monitoring system itself will affect overall operational efficiency and costs, the amount of data generated by it at any one time will be kept to a minimum. The data to be generated

will be determined by system operating efficiency; more monitoring data will be collected when a degradation in either data quality or timeliness occurs, and less data will be generated as the system functions more smoothly. Table 1 summarizes the data that will be collected under the different circumstances described.

Table 1

DATA TO BE COLLECTED FOR EACH FIELD DOCUMENT AT EACH PROCESSING STAGE

	Monitoring Data	
Control Data	Continuous	When Needed
Sign-ons (date, time, name) Sign-offs (date, time, name) Document identification Processing stage Reason for possession Supervisor validation	Actual time spent Total error count (if appropriate)	Time spent on each step Time spent on individual questions Nature of errors Number of errors detected at each step General comments

III. DATA PROCESSING STEPS

INITIAL FORM OF THE DATA

All data will be initially delivered in a form preorganized for computerization. Field survey data will be documented on questionnaires that have been constructed partly for this purpose. These survey data will also have undergone several field quality checks before reaching the coding stage of computerization. The quality checks will include interviewer review, interview supervisor review, and reinterviewing validation of an unspecified percentage of the surveys conducted.

Data gathered through less formal mechanisms will be documented on separate forms. The recording of informal observations, for example, will include the data observed, their sources (names, places, etc.), times, dates, subject indexes, and security and confidential information.

DATA CODING

Field data will first be organized into a concise and standard form that can be readily comprehended and handled statistically. The full process of coding (i.e., the classification and standardization of data for analysis) includes the following activities:

1. Semantic analysis of responses to open-ended questions, i.e., creation and assignment of code representations to verbal answers.
2. Conversion to standard forms for "closed-ended" responses, e.g., rounding numbers and converting to standard units of measurement.
3. Transfer of responses to coding forms if this should prove necessary for efficient digitization.
4. Encoding of names and places in order to secure respondent confidentiality.

Coding will be performed by personnel having special training and extensive preparation concerning each of the survey questionnaires they

will be handling. Training manuals for the questionnaires will be created that will include an introduction to the general area of survey research, the relevant techniques of content analysis, the kinds of codes that will be used and their uses, typical coding problems, and standard coding operational procedures.

Codebooks describing the codes required for the analysis of each questionnaire will also be created. A codebook will contain the coding schemes for each open-ended question of a questionnaire; rules for rounding, normalizing, and converting data; and rules for positioning data in the punched-card fields to which they will be transcribed (e.g., right-justifying numerals and inserting trailing or leading zeros). If, in addition, a coding transfer form is required, the codebook will specify the card and card columns onto which the data are to be transcribed.

While Rand and/or its subcontractor(s) will be primarily responsible for the development of coding schemes, the coders will be responsible for identifying situations that do not fall within the bounds of the code being used. Under such circumstances, coders will report to their supervisors, who will coordinate efforts to resolve conflicts and ambiguities. All coding decisions will be documented and incorporated in the questionnaire codebook.

For questions whose answers require complicated codes (e.g., multipoint scaling), a coder reliability check on a subsample will be made by either the coding supervisor or another coder. Again, the resolution of any ambiguity resulting from these code checks will be coordinated by the coding supervisor.

Coders will also be responsible for the encoding of any questionnaire data that might compromise the confidentiality of a respondent (specifically, respondent name and address). Codes for these encodings will be maintained and supplied from a central list.* Assignments made by the coders will in turn be documented and returned promptly for safekeeping.

Coded documents will be passed on to keypunching. Documents that fall below an acceptable standard (to be established by Rand) will be returned to the field for reinterviewing if possible.

---

*The location of and responsibility for this information is under current study.

DIGITIZING (KEYPUNCHING)

The coded data will be keypunched to produce either a deck of punched cards or a tape reel. The coding forms will be designed to organize the data in a form that will reduce keypunching time and errors. The organization will include the following:

1. Each page of each questionnaire will be identified by source and page number so that questionnaires that have missing or extra pages can be detected.
2. The answers to every n^{th} question will be labeled to indicate the card number and card columns in which those answers are to be punched.
3. All possible responses to closed-ended questions will be listed horizontally at the far right of the page, below the respective questions.
4. Without reducing clarity, as many questions as possible will be listed on each questionnaire page, and only one side of each page will be used.
5. Questions requiring multipart responses (e.g., name, age, and sex of each household member) will be organized so that the parts and their order are easily recognized as being related to each other and belonging to the same question.

All keypunched data must be verified before it can be considered ready for computer processing. Keypunchers must, in addition, be of senior level, preferably with a minimum of two years of general keypunching experience plus special training in the area of survey data.

DATA CLEANING

After digitization, all data will be computer validated and edited for wild codes, interquestion contingencies, and longitudinal consistency (if necessary). By *wild codes*, we mean any response to a closed-ended question that does not fall in the range or set of categories allowed for that question; by *interquestion contingencies*, we mean the

logical consistency of responses to questions that are related (e.g., catching reference to a pregnant male); and by *longitudinal consistency*, we mean the comparison of longitudinal data for reasonable consistency over time (e.g., length of residence).

Data cleaning, the process of computer validating and editing as described above, includes a sequence of related operational steps:

1. Preparation of the validation specification. All allowable codes and interquestion relationships must be specified in a computer readable form and input to the validation segment of the cleaning program for processing and storage. As part of the cleaning specification, it will be possible to specify the following for each response or set of responses:

- o Allowable code ranges and category representations,
- o Allowable syntax for dates, names, etc., and data representations (numerics and alphabetics),
- o Allowable relationships between data using logical connectives (i.e., negation, conjunction, and disjunction).

The processing program will output the validation specifications in a form suitable for review and correction. This output will be kept as a permanent part of the questionnaire documentation.

2. Testing validation specifications. The validation specifications will be tested on a data sample constructed to generate a variety of different errors. If the validation specifications catch all errors as designed, the specifications will be considered ready for operational use.

3. Validation of field data. Keypunched field data will be processed by the cleaning program using the specifications tested in Step 2. Output from the program will indicate the number and nature of the errors detected for each case processed. Data containing errors that cannot be resolved will be returned to the field for reinterview

where possible. Keypunching and other transcription errors will be corrected from source-document information. The output from this step will be annotated and kept as a permanent part of the control system operation.

4. Editing field data. Once data corrections have been specified, they must be transcribed into a computer readable format and used by the editing part of the cleaning program to update the source data for which they were created. Output from the cleaning program will indicate the nature of the correction made and whether or not it was accepted. If the correction is not accepted, the specification will be appropriately modified and the edit rerun. If all corrections are accepted, the edited data will be revalidated as an additional check. All output generated in Step 4 will also be recorded as part of the data control operation.

5. Creating marginals. Tabulations and marginals (e.g., sums, averages, etc.) will be created for use as an additional validity check and as a means of providing some immediate feedback on questionnaire performance. These marginals will be saved for future analysis.

These operational steps will be recorded in a manual for data cleaning. This manual will describe all phases of the use of the cleaning program. In addition, special coding forms will be created for documenting coding instructions, specifying validation specifications, and specifying editing corrections.

Cleaning software employed will either be provided by the fieldwork subcontractor and/or purchased and modified for Rand's use by Rand staff. The actual course taken will depend on whether data cleaning is subcontracted or done at Rand.

DATA STORAGE AND TRANSFER

Rand will store all data it processes (including backup copies) on computer tape--punched cards will not be used. Thus, if any part of the initial data cleaning is subcontracted, that subcontractor will be required to transfer data to Rand on magnetic tape. Tape formats,

density, blocking sizes, and recording mode will be established by Rand. Rand will also provide the tape reels which will be circulated between it and a subcontractor at a rate yet to be determined. Some of these considerations will, of course, be irrelevant if Rand assumes responsibility for data cleaning.

Rand will establish an archive in which all project data and literature will be housed. Included in the archive repository will be

1. All completed and validated questionnaires.
2. Computer readable copies of all completed and validated questionnaires (stored on magnetic tape).
3. All manuals, training guides, codebooks, and documents describing operational procedures.
4. Data control information, including logs and computer listings.
5. Project reports and papers, and computer analysis printouts.

The archive will also contain all computer readable data files (and their documentation) constructed specifically for analysis. It should be noted that these files will differ substantially from the source data generated by the cleaning process. Data for analysis will have been restructured and formatted to reflect the course of analysis as it proceeds. The data management techniques for processing data in this manner will be described in a subsequent version of this report.

IV. SUMMARY

FIELDWORK

The major source of data for the Supply Experiment will be formal surveys. These data will arrive from the field to be prepared for computer analysis through an operational process that includes data coding, keypunching, and computer data cleaning.

Figure 1 illustrates the overall operational data flow. Each processing stage or section will have a supervisor whose main function will be to insure the quality and timeliness of data as the data pass through his station. The means by which each of these sections will communicate will depend largely on the extent of the work subcontracted.

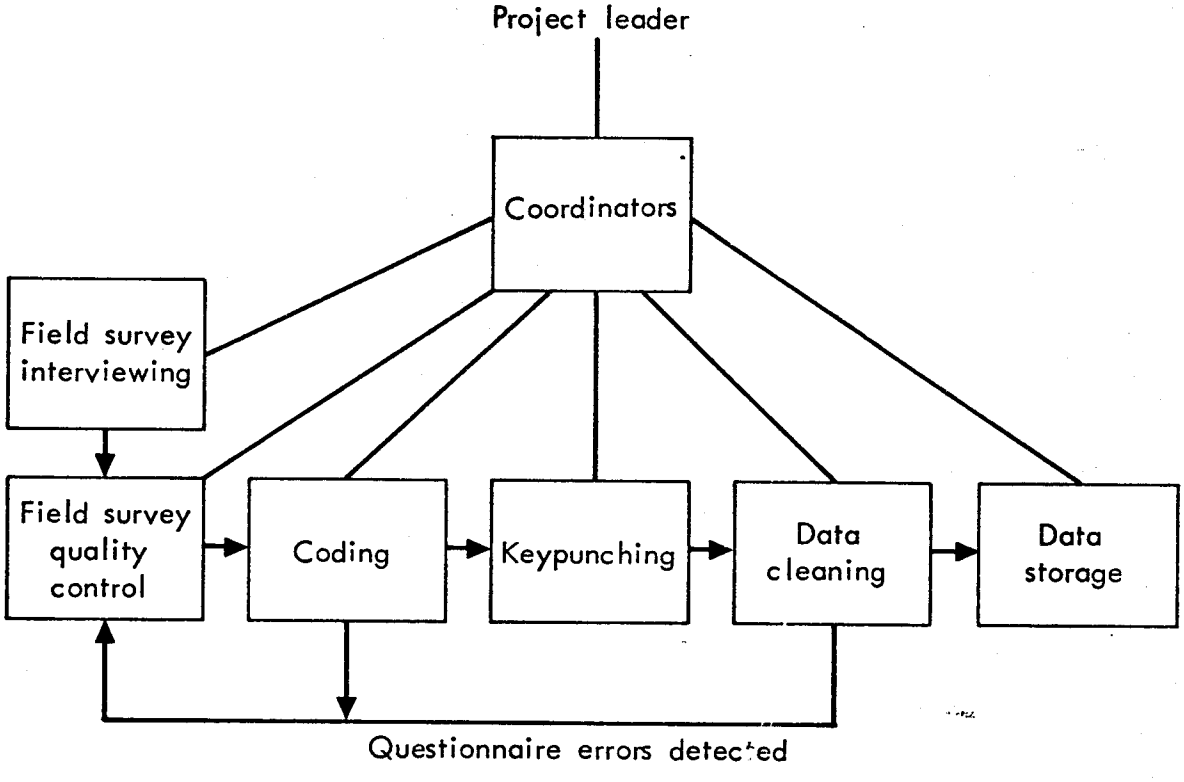

Fig. 1 — Fieldwork overview

In any case, we expect to rely heavily on the telephone and the mails whenever possible. For data that must be transported over distances, we believe that air transportation will be adequate.

Coordination of the fieldwork effort will be the combined responsibility of the Experiment field operations manager, the Experiment data systems manager, and a member of the fieldwork subcontractor's staff.

CONFIDENTIALITY

The maintenance of confidentiality will be crucial to having an uninterrupted flow of high-quality data from the field. We believe that confidentiality will be maintained through the combined efforts of all individuals involved in fieldwork operations and through simple but effective technological devices. Proposed measures for ensuring confidentiality are summarized below:

1. Heavy emphasis on staff indoctrination on the importance of maintaining confidentiality.
2. Data control--the close tracking of all respondent data in order to minimize or eliminate data loss and mishandling.
3. Encoding of any respondent data that could conceivably be linked to the respondent source.
4. Storage of data away from the collection sites.
5. Speedy movement of data through processing to storage.
6. Allowing no more than one copy of a field document to be circulated at any time.
7. Limiting the information supplied in cases where a questionnaire must be identified by true respondent name and returned to the subcontractor for additional fieldwork (e.g., reinterviewing) to only what is required to complete the fieldwork accurately.
8. Ultimate storage of all field data at Rand. Any copies requested by other organizations or researchers will require appropriate approval before release. Such copies, of course, will contain no data that can be traced back to a respondent source.

pt.1