

The Data Management System

*Norman A. Constantine, Wendy L. Constantine,
Michael R. Wrona*

Research data management systems are employed to collect, control, manage, analyze, and archive research data. They consist of personnel, procedures, hardware, and software which together work as an integrated whole. To the extent that systems have been carefully planned and pilot tested and are supported with adequate resources, these systems will be capable of producing reliable, valid, and useful information.

In this chapter we describe the research data management systems developed and operated by the IHDP. This presentation is made in the context of general principles of design, implementation, and operation which are applicable to research data management systems for many types of controlled clinical trials and other longitudinal studies.

We begin with a brief history and overview of the IHDP data systems division, followed by a summary of data collection instruments and procedures. (A complete description of data collection procedures appears in Chapter 28.) We then delineate procedures to maintain the quality of the data. Next, we discuss data management, database organization, and data analysis. Finally, we describe the database documentation and data archives.

from: R. T. Gross, D. Spiker, & C. Haynes (Eds.), *Helping Low Birth Weight, Premature Babies: The Infant Health and Development Program*. Stanford, CA: Stanford University Press. (1997).

Overview

The systems described in this chapter were developed and operated at the IHDP NSO at Stanford University to process all of the IHDP research and evaluation data. These data were collected with 61 data collection forms across nine assessment time points, from birth through age three. The fundamental objective of these systems was to manage and analyze the study data in an accurate, efficient, and timely manner. (Process data regarding implementation of the intervention for the treatment group were collected and processed separately by the PDO at the Frank Porter Graham Child Development Center, as described in Chapter 27.)

System History and Personnel

The design and development of the IHDP data systems began in December 1984 with the recruitment of the field operations and data systems directors. This coincided with the beginning of randomization and the arrival of the first completed baseline data collection forms from the sites (computerized systems for randomization and electronic mail were already in place). Temporary data systems were quickly set up early in 1985 to process the incoming data needed to monitor recruitment, randomization results, and data collection quality, as well as to perform an initial validity study on a measure used to determine eligibility for randomization (see Constantine, Kraemer et al. 1987). While it is prudent to allocate one or two years for adequate planning and pilot testing of a new large-scale system (see, e.g., Helms 1978; Meinert 1986; Metzger 1973), in our situation, as is often the case, time was simply not available.

During 1985, recruitment of data control and database management personnel began, and the temporary data systems were "pilot tested" on the early data. In 1986, staff recruitment continued, and the design and implementation of the formal data systems began. By mid-1986 most of the data control and database management staff had been recruited, and most of the formal data systems were in place. The systems operated continuously through early 1990, after which final data archiving and documentation were completed.

In addition to the directors of field operations and data systems, the core data systems personnel over most of the life of the project included two database analysts, a statistical analyst, a systems operations manager and assistant, a field operations coordinator, a data control manager, and four part-time data control technicians. Consultants, contract programmers, and contract analysts also were employed as needed for specialized coding, systems design assistance, and statistical analysis.

Data Flow

The remainder of this chapter is organized around the natural flow of the data through the five major steps of: (1) data collection, (2) data quality control, (3) database management and organization, (4) data analysis, and (5) data archiving. This data flow is represented in Figure 30.1.

Data Collection Instruments and Forms

Detailed information about the instruments and data forms is contained in Chapter 28. For purposes of understanding the nature of the data to be managed and analyzed, a few key points are noted here.

First, at each assessment point data were collected using specially designed forms which included standardized, validated instruments together with additional items that were developed or modified for this study (e.g., Family Interview form; Interval Health form). Second, instruments and data forms were extensively pretested, pilot tested, and revised as necessary. Finally, to ensure consistent data collection procedures across sites, each data form was accompanied by a detailed written in-

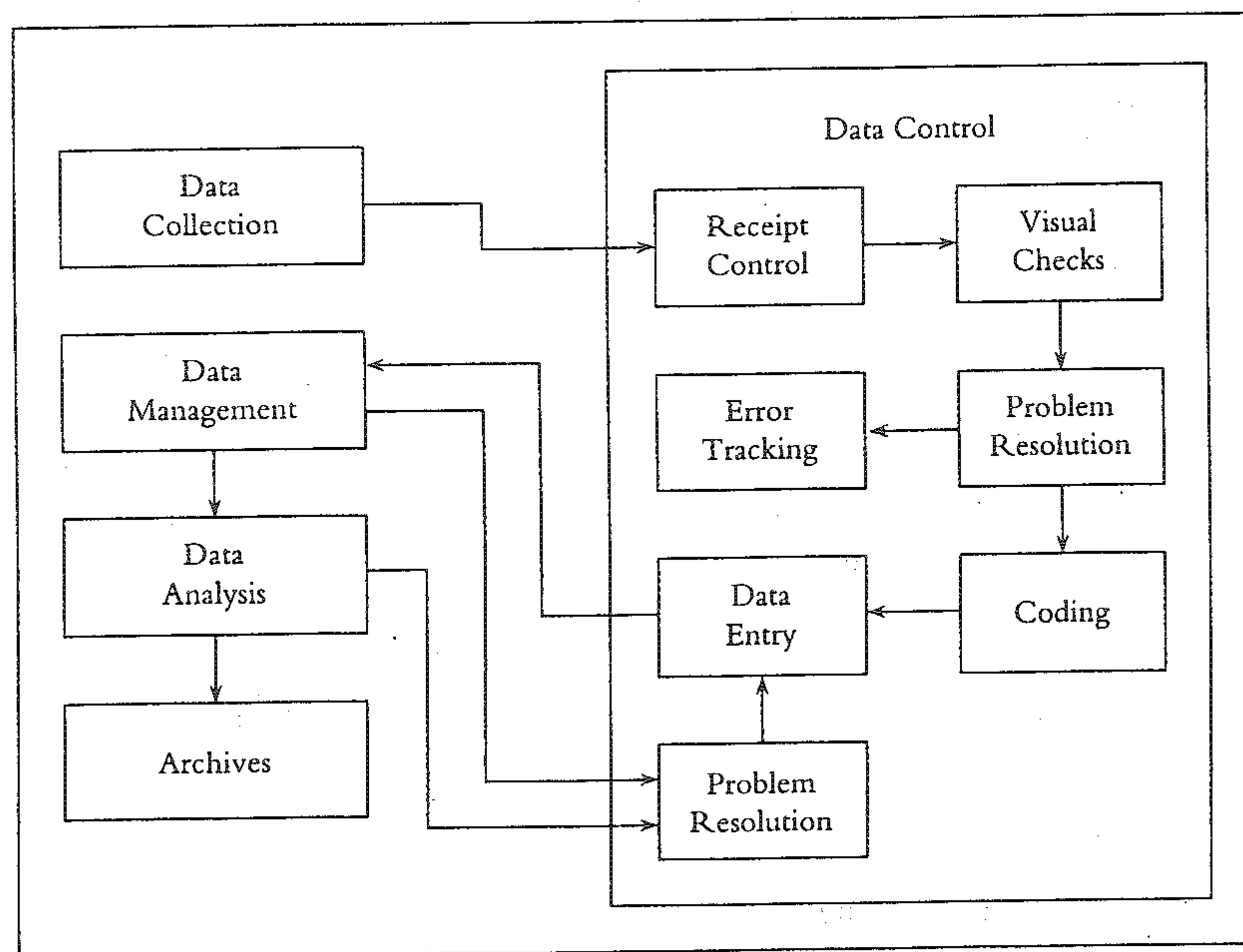


Figure 30.1. IHDP data flow

struction manual (explaining how items were to be administered and scored), and extensive training programs were implemented. These are described fully in Chapters 28 and 29. By the end of the IHDP, more than 60,000 data forms had been processed at the NSO.

Data Quality Control

Data quality control is a critical, but often insufficiently developed, part of research data systems. In a longitudinal study such as the IHDP, data quality control has two primary objectives. The first is to identify and correct data errors. This is facilitated by structured data inspection and remediation procedures. The second objective is to reduce the incidence of errors made in the data collection process, thereby increasing data systems productivity. Achieving this objective requires providing regular data quality feedback to the data collectors.

Every completed IHDP data collection form was reviewed for completeness and accuracy by the site's study coordinator (for unmasked data) or by the site's evaluation coordinator (for masked data), and any problems detected on the form were resolved. The site's study or evaluation coordinator then sorted the forms by type, completed transmittal logs for each form type, and shipped the forms to the NSO by air express.

The data control group at the NSO preprocessed the data collection forms before data entry, and maintained hard copy files of these forms. Preprocessing activities included logging the data collection forms into the computerized receipt control system, visually checking the forms for errors, coding open-ended responses as necessary, and creating and logging data-entry batches. In addition, the data control staff checked and resolved computerized error and missing value reports. (These activities are described in the Database Management section below.) The remainder of this section describes the preprocessing activities performed by the data control staff.

Receipt Control

Receipt control systems document the receipt of data collection forms. These systems are used to provide timely feedback to the sites about specific missing forms and also to monitor compliance with research protocols, e.g., to determine the proportion of forms collected within their targeted assessment date windows. The NSO's receipt of each data collection form was documented through a separate computerized receipt control and assessment monitoring system (see Constantine, Shing et al. 1987), which was essential, because most forms were not keypunched and entered into the main database system immediately upon receipt.

Upon receipt of each shipment of data collection forms, the content of the shipment was checked against the site's shipping log, and any discrepancies were noted and resolved. One copy of the shipping log was returned to the site and the other filed at the NSO. For each form received, the following information was recorded on a data entry coding sheet: ID number, form number, date form completed, and date form received. These data entry coding sheets were keypunched weekly.

Three reports provided sites with monthly feedback on receipt control of data collection forms: (1) the window monitoring list reports provided case-by-case documentation of whether a form was received and whether it was completed within its targeted assessment date window, (2) the window monitoring summary reports showed the percentage of forms received within the window for each assessment and provided a basis for comparing the performance of one site with another, and (3) the assessment monitoring system graphs presented a visual representation of site performance over time in conducting assessments and completing them within the window. All three reports were based on the date each study infant reached the appropriate CA for that assessment together with the allowable window of time around that date. Monthly mailings to each site included the window monitoring list reports for the site, the window monitoring summary reports for that site and for the seven other sites, and the assessment monitoring system graphs for that site and for the combined sample. These reports provided site data collection staff with timely feedback about their performance. Usually they were a source of positive news and were gratifying. Sometimes they illustrated problems, and were used to determine which sites needed further training and assistance.

Visually Checking for Errors

As the lag time between data collection and error checking increases, the likelihood of retrieving missing data or correcting errors decreases. For this reason, prompt error checking is essential. At the NSO, the data collection forms received each week were checked visually for errors using item-by-item visual check specifications.

First, the four-digit infant ID number and birth date recorded on the form were visually checked for agreement against a master inventory. This birth date check was used to verify the identity of each subject without revealing the subject's name to NSO staff. An electronic mail query was sent to the site immediately if this check revealed that the data were not consistent. This type of initial ID check is essential, because such an error is one of the most serious that can occur.

Next, three additional types of errors were checked for: missing values, inappropriately followed instructions, and incomplete or illegible data. For each potential error detected, a Data Quality Control form (see Appendix 30A) was completed. One copy of this four-copy, carbonless paper form was retained at the NSO, and

the three remaining copies were sent to the site to obtain clarification or correction. After retrieving the required information, the site noted the clarification or correction on the Data Quality Control form, retained two copies, and sent back the third to the NSO. When this copy was received, the resolution was noted on the original data collection form and the Data Quality Control form was filed.

Most problems discovered during visual checking were resolved before data entry; however, occasionally problems were resolved afterward. In the latter case, of course, corrections were applied to the computer record as well as to the paper forms.

The procedures described above were used to check all 40-week, 4-, 8-, 12-, and 36-month forms. For the 18-, 24-, and 30-month forms a sampling procedure was used. With these three assessments, the ID number and birth date were checked against the master inventory for all forms. Then, for each form type, every fifth form completed by each data collector was selected for the complete visual check. Checking of additional forms was performed as indicated by the results of these sample checks.

Problem Tracking

As Data Quality Control forms were completed, they were entered into a forms management system to facilitate tracking and to produce error summary reports (see LeTendre and Shing 1989). Four types of error summary reports were generated by this system:

1. Quarterly reports listing each Data Quality Control form by data collection form type and by the item containing the error. These enabled NSO field operations staff to identify items for which the site staff needed further training.
2. Quarterly reports for each site showing the average number of days that elapsed between transmittal of the Data Quality Control form to the site and receipt of a reply by the NSO. These reports were used by the sites and the NSO staff to monitor whether responses were sent within the two-week guideline.
3. Monthly reports listing by staff ID number each Data Quality Control form sent to each site during the previous month. These enabled site personnel to identify staff making consistent errors so that supplemental training could be provided.
4. Monthly reports summarizing the error rate for the last ten weeks and for the current calendar year at each site. These enabled the NSO to identify those sites needing to increase effort in reducing errors. These reports also enabled each site to compare the quality of their data collection with that of the other sites, serving as an incentive for each site to achieve maximum data quality.

Our experience indicates that this kind of monitoring does serve as incentive to improve the quality and timeliness of data collection. When this feedback was less

than optimal, however, it often created anxiety and had negative effects on staff morale. To counteract such negative side effects, written feedback should be accompanied by personal contact and support (e.g., phone calls, site visits).

Coding

In any data collection effort, efficient and user-friendly formatting of the data collection forms will help ensure data quality. In most instances, properly field tested precoded items are superior to open-ended items which require recording free-form responses verbatim for later coding. Precoded items are more efficient for the data collectors to complete, and can save significant time that would otherwise be expended on coding. At the same time, properly precoded responses increase reliability by standardizing response options for the respondent.

Most variables on the IHDP data collection forms were precoded. Of the small number of open-ended variables, approximately one-third were coded at the NSO, while the remaining responses were never coded.

Open-ended variables coded at the NSO were coded using either standardized systems or systems developed by the NSO. For example, with the Interval Health form, open-ended questions were included to ascertain the health problems of the infant during the interval since the last assessment. For these questions, the International Classification of Diseases, Revision 9 (ICD-9) coding system was used. This is a well-documented coding system used by hospitals and other health organizations to document illnesses, conditions requiring surgery, and other health conditions. An expert in this coding system was hired as a consultant to code the open-ended health questions collected with this form across all assessment points.

Coding categories for some open-ended items were developed by data control staff, sometimes with the assistance of the investigator who requested that the item appear on the form. To develop the codes, we examined verbatim responses and created codes for the responses that were given frequently. Responses occurring rarely were coded as "other" (approximately 10% of the responses to open-ended questions were coded in this category). To check coding reliability, a second person conducted independent coding of the same items, and a supervisor compared the coded items for consistency.

Database Management

Database management involves entering the data into the computer, constructing and cleaning the computerized data files, and performing data backups. Each of these three processes is discussed below.

Data Entry

To ensure accuracy, an off-site data entry service keypunched all forms twice. Most data collection forms were submitted for keypunching in two or more batches after data collection for that form type was completed for the entire sample, and after problems discovered during visual checks were resolved. The first batch consisted of a sample of approximately 50 completed forms which were used during the final development and testing of the data-cleaning computer programs for that form. Soon after, the remaining forms received were submitted in one or more additional batches.

Data Cleaning

Standard cleaning procedures for all IHDP data collection forms involved four sequential steps: visual checks, cumulative edit reports review, initial descriptive statistics review, and final descriptive statistics review. Visual check procedures were discussed above; the remaining three data cleaning procedures are described briefly below.

Cumulative edit reports review. After data entry was completed for a data collection form type, a cumulative edit report for that form was computer generated by the File Update System (FUS). FUS is a SAS-based database management system we developed for the IHDP (see IHDP Data Analysis Systems Group 1988). This system performed computerized checks of the data for (1) allowable value ranges for each variable, (2) skip pattern compliance, and (3) interitem consistency (see Branagh 1988). FUS located inconsistent or invalid responses, and assigned appropriate missing value codes to variables which were legally skipped. All findings and actions taken by the FUS were documented on the cumulative edit report.

Detected errors and unclassified missing values were resolved by the data control group. The paper copy of the original data collection form was always compared with the cumulative edit report items in question to check for keypunch error. For other types of errors, the site was contacted for resolution if the site staff could be reasonably expected to obtain confirming or correcting information. (Missing information was not sought if the form had been collected more than three months earlier and the missing data were time dependent; in these instances, a special missing value code was inserted into the data file.) Corrections were generated and entered into the computer. The correction process was repeated as many times as necessary to identify and resolve all detected errors.

Initial descriptive statistics review. After all errors and unclassified missing values were corrected, the third stage of cleaning began. This involved review by data systems

staff of a standard descriptive statistics report for that data collection form. One important step in this phase of cleaning was to check outlier values (i.e., those extreme values relative to the rest of the distribution) for all quantitative variables.

Final descriptive statistics review. The last stage of cleaning consisted of a review of the descriptive statistics report by the deputy director (for the "masked" developmental assessments) or by the field operations director (for all other data collection forms). Any additional problems in the data or labeling identified during this final review were corrected, and then the descriptive statistics report was rerun and reverified. This was the final stage of review prior to data analysis.

While these cleaning steps were time consuming and tedious, such a systematic approach to data cleaning is a necessary investment in data quality.

Data Backups

There are three principal reasons for regular data backups. The first is to protect against hardware failure. In a mainframe environment, data loss due to hardware failure is not usually a problem, because data center personnel and mainframe computer manufacturers generally create elaborate safeguards. The second reason is to protect against system programmer or operator errors which can damage or destroy computer files. Finally, data backups also provide a historical trail of database development which is useful for writing documentation and conducting data audits.

The research database production and maintenance computer files were backed up from mainframe disk to tape by the NSO every week. Weekly backups were kept for two months, and monthly backups were kept as long as the NSO remained open. Additional semi-annual backup tapes were made of the database production and maintenance files and stored in an off-campus vault administered by the Stanford Data Center. Finally, all files were automatically backed up by the Stanford Data Center every week. These weekly backups were maintained for three months, while monthly copies were maintained for one year.

Database Organization

Research databases typically go through two phases: a dynamic one and a static one. The dynamic phase is associated with the data collection and data-cleaning processes. After all the data have been entered into the database and cleaned, the database becomes static. At this point the database is ready for analysis.

One general organizing principle for dynamic databases is that there should be no repetition of information (Korth and Silberschatz 1986). Repetition of information introduces unnecessary complexity to cleaning the data and may lead to

inconsistencies within the database. If an erroneous data value is corrected in one file but remains unchanged in another, the integrity of the database is compromised.

Once the database becomes static, however, redundancy of information is no longer a problem. In fact, during the static phase, there may be advantages to including the same piece of information in several locations. For example, a separate data file may be created for each analysis, containing only those variables pertinent to that analysis. If several analyses use the same data elements (variables), then there will be redundancy of information across analysis databases. This will not be a problem as long as the data have been fully cleaned before beginning the analyses.

The IHDP research database contained four types of datasets: infant inventory, cumulative, supplemental, and analysis. The first three types each went through a dynamic phase during which data were regularly cleaned and added to the datasets. The analysis datasets were always static. The structural organization among these datasets together with the data management files is illustrated in Figure 30.2.

The infant inventory dataset contained basic data, such as birth weight, gestational age, treatment group, and disposition code for each study infant. These data were transmitted during randomization by the sites to the NSO by electronic mail, and were later verified against duplicate data recorded on the early data forms. To adhere to the principle of nonredundancy, we did not include these basic data in any cumulative or supplemental dataset. Corrections and updates to this dataset were made as necessary.

The data from the 61 data collection forms were stored in cumulative datasets. To facilitate data cleaning, each cumulative dataset corresponded to one data collection form, and, except for a few instances, contained one record for each infant. Documented missed assessments for any infant were represented by the missing value code Q ("Missed Assessment") appearing in the dataset for every variable for that infant. In addition, a small number of supplemental datasets were created to meet special processing needs. Each supplemental dataset is described in detail in the dataset history and review summaries, below.

A primary analysis dataset was created from the database to be used for the primary study analysis. Similarly, ancillary analysis datasets were created from the database for ancillary studies. All analysis datasets were maintained as SAS datasets only. Each was created according to variable and sample subset specifications provided by the principal investigator for that study.

Data Analyses

Upon completion of the cleaning and freezing of a cumulative dataset, the data became available for statistical analysis.

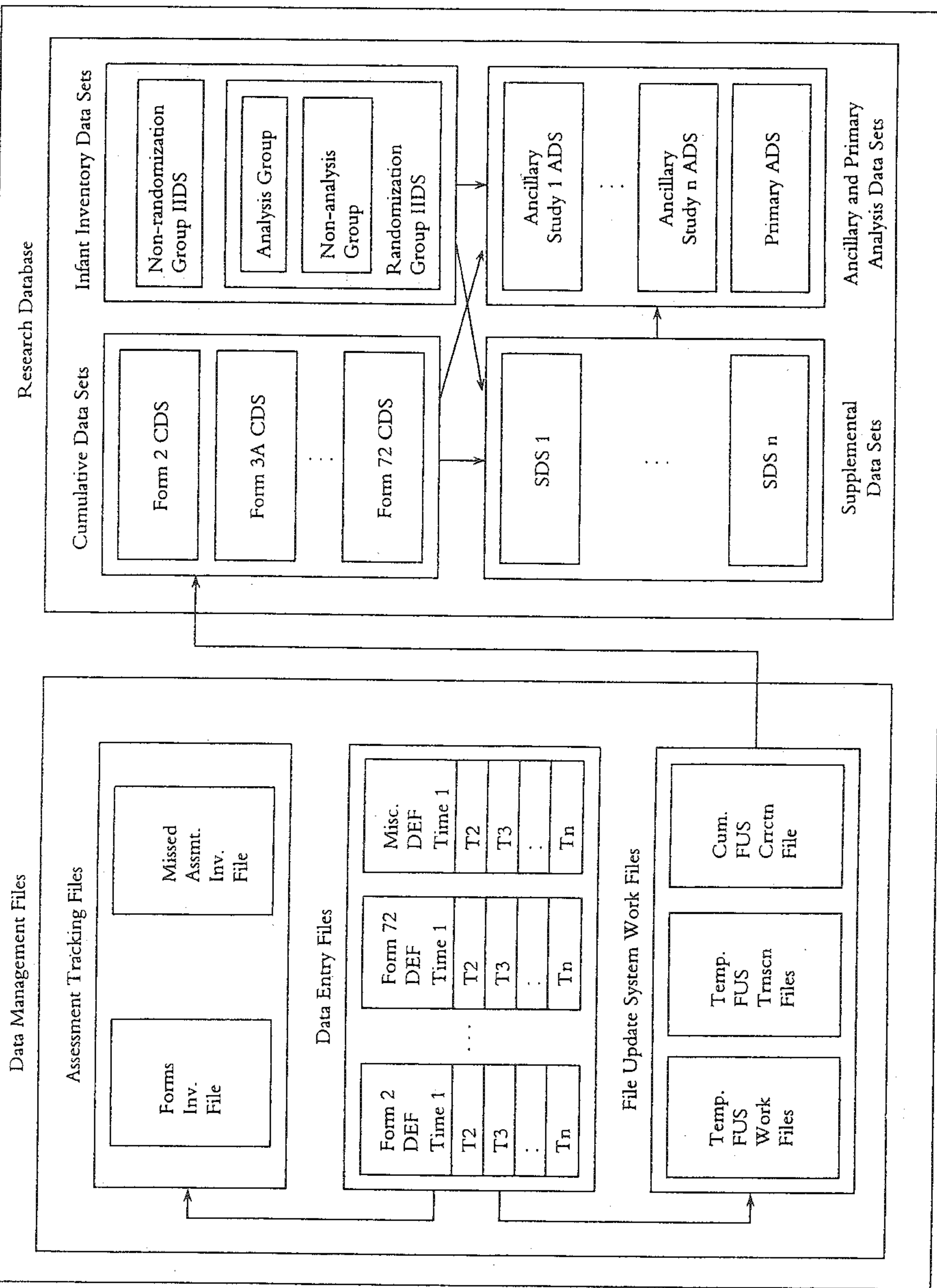


Figure 30.2. Structure of IHDP databases

Primary Analyses

Analyses for the main IHDP journal article (see Chapter 11) were considered the primary analyses for the study. A primary analysis dataset was created, consisting of a variety of outcome, initial status, and other variables for each primary-analysis-group infant. These analyses were implemented under the direction of the chair of the RSC and the data analysis systems director, with reviews from the national study director, RSC members, site directors, and the NSO senior research staff.

Ancillary Analyses

The NSO also provided analysis support for more than 80 ancillary studies. Upon proposal review and approval by the national study director, the analyses for an ancillary study were assigned a lead data analyst and scheduled according to data availability and competing priorities. After review of the written proposal for the ancillary study, the lead data analyst contacted the investigator and/or statistician for the study to resolve problems and obtain final working specifications. An ancillary analysis dataset then was created according to the established schedule, and, in most cases, the analyses were performed at the NSO. The dataset and, when applicable, the analysis results were formally and independently verified by a second analyst and reviewed by the data analysis systems director before distribution. Upon completion of a prepublication draft manuscript, the tables and figures were reviewed by the lead analyst, and further analyses were run, if needed, to resolve inconsistencies or implement minor modifications.

Generic Analysis Programs

Four generic SAS macro programs were developed and maintained for use in commonly required analyses. PIGCOR was developed to compute pooled intra-group correlation coefficients and provide testing for group homogeneity (Constantine and Shing 1988; Shing and Constantine 1988). BIFAC was a SAS program developed for implementing analyses of variance and accommodating the potential effects of site differences and interactions on main effects, based on principles and procedures described by Fleiss (1986). PIGLIN was developed to apply related strategies for site differences to multiple linear regression analyses. IRANREG was developed to implement an individual random regression growth model. This was used specifically to analyze individual infants' growth patterns over time, rather than the more common approach of analyzing aggregate group statistics (see, as an example, Casey et al. 1990). These programs were used by several different investiga-

tors within the IHDP, and they all proved to be useful and relevant to the research questions posed.

Database Documentation

Documenting the database can be one of the most dreaded tasks faced by a programmer/analyst. However, good database documentation is important to any research project, and it is essential for large-scale longitudinal studies. To avoid recall errors and problems resulting from staff departures, documentation should not be left until the end of the project; instead it should be done concurrently with design and construction of the database and analyses.

Codebooks

The codebooks for the cumulative datasets provided variable-by-variable listings for each dataset, including the variable names, variable labels, and item locations on the form. In addition, value ranges were provided for continuous variables, and value lists and definitions were provided for categorical variables. These defined the range checks which were performed by the computer data management system.

Each codebook was supplemented by a variable documentation report showing technical computer characteristics of the dataset and its variables. These reports were produced by a SAS procedure.

Dataset Status Tracking

Cumulative datasets were tracked through six distinct status stages: (1) construction and cleaning, (2) descriptive statistics cleaning, (3) descriptive statistics verification, (4) temporary freeze, (5) analysis freeze, and (6) final freeze. For the three other types of datasets (infant inventory, analysis, and supplemental), only the last two stages were formally tracked.

Each status stage determined the level of access to the data for statistical analysis and archiving. Cumulative datasets in one of the first three stages were not available for analysis. These three stages correspond to the three levels of review described above in the Data Cleaning section. The fourth stage (temporary freeze) signified that the data had been thoroughly and systematically cleaned. Informal preliminary analyses were sometimes conducted using datasets in the temporary freeze stage. At the fifth stage (analysis freeze), records containing missing value codes for missing observations were created and added to the dataset so that the dataset contained one record for every infant. *Only after a dataset was classified at the analysis freeze level was*

it available for use in ancillary or primary study analyses. Datasets were granted final freeze status after any data anomalies noted by analysts or investigators were reviewed, and corrected as necessary, and after each cumulative dataset was fully reconstructed and verified from the original keypunch files. Final freeze status was required of the analysis dataset for publication of a study, and for all datasets before distribution and archiving.

Dataset History and Review Summaries

The dataset history and review summaries provided historical reference information about each infant inventory, cumulative, supplemental, and primary analysis dataset. Each consisted of up to four sections: dataset description, dataset management, special circumstances, and dropped variables. These sections are described fully in the introduction to the dataset history and review summaries in the public-use database tapes documentation.

In place of the dataset history and review summaries, historical documentation for each ancillary analysis dataset consisted of the original dataset specification form and copies of all subsequent correspondence with the investigator regarding changes or clarifications to the specifications.

Other Documentation

The SAS code used to create the dataset (as referenced in the dataset history and review summaries) was part of the documentation maintained for every dataset. This included the exact SAS code for all value checks, skip pattern processing, and generic recodes. For the cumulative and infant inventory datasets, paper files were retained of every correction or modification made subsequent to keypunching the original data.

Data Archives

IHDP-Use Database Tapes and Documentation

These IHDP-use database tapes were distributed to all site directors, RSC members, and NSO senior research staff in late 1989 so that IHDP investigators could continue ancillary research after the close of the NSO. All datasets were provided in SAS format.

The documentation to accompany the tapes included the codebooks and dataset history and review summaries described above, as well as a user's guide explaining the use of and technical details about the tapes.

Public-Use Database Tapes and Documentation

The public-use database tapes and documentation are archived at the Inter-University Consortium for Political and Social Research (ICPSR) in Ann Arbor, Michigan. All datasets are provided in both SAS and ASCII format. These are subsets of the datasets in the IHDP-use database containing only those variables which cannot be used to identify the research subjects.

NSO Archives

All completed data collection forms are archived in a Stanford-area long-term storage facility which specializes in preserving confidential record material under temperature- and humidity-controlled conditions. Access to these forms is supervised by the deputy director of the IHDP.

Conclusions

The IHDP research data management systems were a success. Over a period of six years, more than 60,000 complex data collection forms were received and processed. An integrated research database was created, thoroughly cleaned, and validated, and is supported by comprehensive descriptive and historical documentation. Statistical analyses were conducted and verified in support of the primary study together with more than 80 separate ancillary studies.

In looking back over the six years of operation of the IHDP data systems, it is instructive to consider the essential qualities of successful research data systems in relation to the IHDP experience. We hope this will help designers of future systems avoid some of our mistakes as well as profit from our positive experiences and attainment of our goals. In this chapter we have described in some detail our experiences and our resulting systems. In closing, we note and briefly discuss six key components of successful research data management systems which we encountered head on, for better or for worse: (1) planning and pilot testing, (2) total quality control, (3) feedback loops, (4) integrated systems, (5) data systems personnel, and (6) commitment of resources.

Planning and Pilot Testing

"Plan to throw one away; you will, anyhow" (Brooks 1975, as quoted in Helms 1978). The IHDP systems suffered early inefficiencies and increased time and resource pressures by omitting advance planning and pilot testing. We were fortunate, however, to lose no data and to release no analyses based on our untested early

systems. True to Brooks's warning, these early systems became our unplanned pilot tests as we scrambled to redesign systems on the basis of the lessons and experiences of our early operations.

Total Quality Control: No Error Is Too Small to Warrant Detection and Remediation

Total quality control requires that every aspect of data quality and data integrity be examined, checked, and double-checked. Something not quite right in the database should make the data managers lose sleep at night, as should some aspect of the database that could be checked but hasn't been. As with the IHDP systems, quality control procedures should pervade every level of every system: at the data collection, data entry, data management, and analysis levels.

Feedback Loops: What the Data Collectors Don't Know Will Hurt Them

Feedback loops were a central component of the IHDP systems. In a longitudinal study, merely identifying and correcting errors is not likely to be as effective as communicating performance back to those responsible for the errors, usually data collection or data entry staff. Feedback should be immediate, regular, constructive, and public. It should be reinforced with additional training or support as appropriate.

Integrated Systems: The Pieces Must Work Together

Integrated systems work together in a manner which is more effective than the sum total of the parts. Each component of the IHDP systems was designed to function effectively together with every other component by building on and supporting what the other pieces did. Just as data checking and feedback loop components must work smoothly together, so too should all other pieces of any integrated system.

Data Systems Personnel: Technical Skills Are Not Sufficient

"Quality management has more to do with people than with machines. People are the most variable, least predictable, and most important part of any system" (Kerridge 1991). Technical proficiency and continuing education are critical qualifications for data management personnel. These are not sufficient, however. No person who is not an obsessive-compulsive type should even consider a career in

data management (see Lutz 1977). Data systems personnel should be service oriented and derive satisfaction from providing the research staff with the services which enable them to meet the objectives of the project successfully. Furthermore, it is helpful to have on the data systems team at least some staff who have mixed backgrounds in both the technical expertise required for the systems and the research methods and substantive content area of the investigation. Intimacy with the data and a commitment to their validity are promoted by sharing the investigator's role.

Commitment of Resources

Don't underestimate everyone's natural tendency to underestimate the resources needed to do the job. It is not unusual for investigators and funding agencies, not to mention system designers, to underestimate substantially the resources required to develop and operate successful research data management systems. With the IHDP we were fortunate to acquire funding over time which was commensurate with the importance of the data and the research questions being addressed.

A key component of our success was investment in competent staff. It is interesting to review job postings and advertisements for research data managers requiring a long list of diverse skills and experiences, but offering salaries better suited to an entry-level research assistant. This type of savings in the end is as poor an investment as skipping the pilot test.

Another important resource area is computer time. The IHDP systems were designed during 1984–85. At that time, for a database as large as we planned, PC-based systems were not practical. Given the pressures for immediate startup, the state-of-the-art but expensive Stanford mainframe was our only alternative. This resulted in charges for computer use time in excess of \$150,000 per year. These costs for computer use were comparable to those of other large collaborative clinical trials of the past (see, e.g., Kronmal et al. 1978). Today, of course, the situation is different. A system as complex and large as the IHDP's could be accommodated on a 486 fileserver with a 600-megabyte hard disk and tape backup. The cost, based on mid-1995 prices, would be under \$3,000. This solution would require some additional staff time for system maintenance and network administration. The savings, nevertheless, would be tremendous.

Maintaining the Cohort

*Mark Swanson, Debbie Flood, Jackie Hickman,
Pat Lee, Judy C. Bernbaum*

For a successful clinical trial, effort must be made by investigators to keep the intervention and comparison groups intact. Attrition often involves that part of the sample that is of most interest to the research questions (i.e., the sickest, most fragile, least stable), thus threatening internal and external validity and limiting subgroup analysis opportunities (see Chapter 10) (Meinert 1986; Mosteller et al. 1980; Pocock 1983). Further, any analyses addressing the effectiveness of the intervention must, by design, include all enrolled study subjects, regardless of their degree of participation. When study results fail to demonstrate that the intervention is effective, but attrition was high, the intervention has not been fairly tested: The failure to demonstrate the intervention's effectiveness may have been due to the dropouts not having been fully exposed to the intervention.

Strategies for maintaining participant interest in clinical trials have been described by Meinert (1986). The most important factor cited is a positive attitude shown by the clinic staff, defined as "treating patients with courtesy and dignity and [taking] an interest in meeting their needs." Specific helpful techniques include: a convenient clinic location for those arriving by public or private transportation, reimbursement of travel and parking fees, reimbursement of clinic registration fees, scheduled but flexible appointment times, establishing for the patients a positive identity with the program through such methods as certificates and ID cards, regular telephone or mail contacts between visits, and remembering patients with cards on special days (holidays, birthdays, etc.). These approaches have been endorsed by

Helping
Low Birth Weight,
Premature Babies

*The Infant Health and
Development Program*

E D I T E D B Y

*Ruth T. Gross, Donna Spiker,
and Christine W. Haynes*

STANFORD UNIVERSITY PRESS

Stanford, California 1997

Contents

	Abbreviations	xxxiii
	Foreword	xxxv
PART I	The IHDP Clinical Trial: Rationale and Program Description	
1	The LBW, Premature Infant	3
2	The Intervention Model	17
3	Home Visiting	27
4	The Child Development Centers	42
5	The Research Plan	59
6	Rationale for Selection of Measures: Cognitive Development	67
7	Rationale for Selection of Measures: Behavioral Competence	84
8	Rationale for Selection of Measures: Health Status	93
9	Random Assignment in Clinical Trials: Issues in Planning	106
PART II	The Study Results	
10	Recruitment and Retention	125
11	The Primary Child Outcomes	139
12	Possible Confounding Issues Concerning the Primary Child Outcomes	154
13	Enhancing the Cognitive Outcomes of LBW, Premature Infants: For Whom Is the Intervention Most Effective?	181
14	Participation in the Intervention and Its Effect on the Cognitive Outcome	190

15	Changes in Cognition and Behavior from 12 to 36 Months	203
16	Effects of the Intervention on Different Domains of Cognitive Functioning	218
17	Use of Health Services	228
18	Quality of the Home Environment	242
19	Mother-Child Interaction	257
20	Maternal Problem Solving	276
21	Maternal Attitudes and Knowledge About Child Development	290
PART III Studies of Growth and Development		
22	Growth Studies	307
23	Neurologic Status at 36 Months of Age	324
24	Social Competence: The Adaptive Social Behavior Inventory (ASBI)	335
25	The <i>Neonatal Health Index</i>	341
PART IV Operational Issues		
26	The National Study Office: Structure and Function	361
27	The Program Development Office: Structure and Function	370
28	Operational Issues in Implementing the Evaluation	381
29	Considerations in Implementing the Masked Assessments	394
30	The Data Management System	408
31	Maintaining the Cohort	425
32	Serving Children with Special Needs	432
33	Health and Safety in the CDCs	448
34	Staffing and Interdisciplinary Teamwork	460
PART V Cost Analysis of the IHDP		
35	The Cost of Implementing the Intervention	479
	Appendixes	505
	Journals Cited	557
	References	563
	Index	627

Figures

CHAPTER 1

Fig. 1.1	U.S. annual rates of neonatal mortality and LBW births, 1955-85	4
Fig. 1.2	Evolution of developmental dysfunction in LBW, premature children	8

CHAPTER 2

Fig. 2.1	Biosocial systems model	19
----------	-------------------------	----

CHAPTER 3

Fig. 3.1	STOP card	35
Fig. 3.2	THINK card	36
Fig. 3.3	PLAN card	37
Fig. 3.4	PLAN sheet	38

CHAPTER 4

Fig. 4.1	Number of <i>Partners</i> activities introduced and used in the CDCs	51
----------	--	----

CHAPTER 5

Fig. 5.1	Research schematic	60
----------	--------------------	----

CHAPTER 9

Fig. 9.1	Sample accrual graphs in the heavier and lighter groups over the weeks of recruitment	119
Fig. 9.2	Example of monitoring graphs	120