# The Use of Relational Databases in Large Scale, Multi-Site Research Projects: Mitigating the Impact of Data Errors

Jose M. Cortina

J. Matthew Beaubien

Robert W. Holt

George Mason University, FAA Grant Team

October 20, 1998

Technical Report #98-001

Please address all requests for reprints to:

Dr. Robert W. Holt
ARCH Lab
MSN 2E5
Department of Psychology
George Mason University
Fairfax, VA 22030-4444

**Abstract**

The scientific enterprise proceeds via three main processes: research design, statistical analysis, and the public dissemination of research findings. For the scientific process to proceed smoothly, however, a number of supporting functions must also be performed. These include data collection, data formatting/cleaning, and data warehousing. At first glance, these supporting functions may seem trivial. Nevertheless, to avoid the deleterious effects of bad data on the scientific knowledge-generating process, it is essential that these tasks be performed with meticulous attention to detail. With this in mind, the authors present a number of guidelines for generating "clean" data. These guidelines have been developed over the course of the three-year, multi-site program evaluation effort. The authors propose that adherence to these guidelines can help future research efforts proceed both smoothly and efficiently.

## Authors' Notes

This technical report has been written with the needs of both academics and applied researchers in mind. Because the material is of a complex nature, it will be difficult to satisfy the preferences of both audiences. For example, academics may feel that the material is not presented in enough detail, while applied researchers may feel quite the opposite. Therefore, to make future reports more user-friendly, the authors welcome any comments that you might have. Please direct them to the address listed on the front cover.

The authors would also like to acknowledge the support of several individuals and groups who assisted us in the generation of this technical report. Without their help, this project would not have been possible. Specifically, we extend our thanks to Dr. Thomas Longridge (FAA, AFS-230), Dr. Eleana Edens (FAA, AAR-100), Captain Bill Hamman (United Air Lines), the entire Quality Assurance department at United Air Lines, and Captain Kim Schultz (Atlantic Coast Airlines).

Funding for this report has been provided by a grant from the Federal Aviation Administration entitled "Analysis of CRM Procedures in a Regional Air Carrier". The views expressed in this report are those of the authors, and do not necessarily reflect the opinions of the Federal Aviation Administration or United Air Lines.

**Key Terms and Acronyms**
(Seamster, Boehm-Davis, Holt, & Schultz, 1998)

**Codebook** - A document which is written to facilitate the analysis of a data set. Codebooks typically include a description of each variable, the appropriate range of values for each variable, missing data codes, and instructions regarding the "reverse-coding" of variables.

**Crew Resource Management (CRM)** - The effective use of all resources (human and otherwise) on the flight deck. Such resources include air traffic control personnel, flight attendants, maintenance technicians, and dispatchers.

**Event Set (ES)** - A relatively independent segment of a simulated flight. Each event set typically includes a trigger, possible distractors, and environmental conditions.

**First Look** - An initial look at crew technical ("stick and rudder") performance in the simulator. During recurrent training, First Look performance evaluations constitute the first performance data that are collected.

**Fleet Captains** - Individuals responsible for the training and performance of an entire fleet of aircraft and their associated crews. Typically, several Standards Captains report to one Fleet Captain.

**Line Oriented Evaluation (LOE)** - An evaluation of individual and crew performance which is performed in real-time in a full-motion flight simulator. Emphasis is placed on performance evaluation, as opposed to training. Emphasis is also placed on Crew Resource Management performance, as opposed to technical maneuvers.

**Line Oriented Flight Training (LOFT)** - Much like a Line Oriented Evaluation. However, greater emphasis is placed on training, as opposed to evaluation. If a portion of the training is failed, crew members must undergo additional training until meeting corporate benchmarks of performance.

**Maneuver Training Validation (MTV)** - Much like a Line Oriented Evaluation. However, emphasis is placed on performing critical technical maneuvers, such as landing under conditions of low visibility.

**Pilot Flying (PF)** - The crewmember who is physically flying the aircraft. At any time, this may be the Captain or the First Officer.

**Pilot Not Flying (PNF)** - The crewmember who is not physically flying the airplane.

**Pilot Performance Database (PPDB)** - A computerized database which tracks pilot and crew performance data over time. Information is typically identified by means of unique individual and

crew identification numbers.

**Standards Captain (SC)** - The individual responsible for conducting training, as well as providing the assessment of such training. Standards Captains are often referred to as a Pilot Instructor/Evaluators (I/E's) or Check Airmen.

## Introduction

Any data collection effort can be fraught with problems that make the data difficult or impossible to use. This is especially true for large, multi-site efforts, in which the data are collected by one team and analyzed by another. However, many of these problems can be avoided with meticulous attention to detail during the research design and data collection phases. Therefore, it is no understatement to suggest that the effort to produce clean, usable data must be marked by constant vigilance.

Data quality errors can occur at any time, and they are far more common than one might suspect. Therefore, the authors present a list of potential problems (below), along with guidelines for avoiding and/or mitigating these problems. At first glance, many of these guidelines may seem like a waste of time. However, the reader should bear one critical point in mind: it is much less time consuming to avoid or correct errors at the initial stages of data collection, than to pore over computer output looking for subtle causes of bizarre results.

Before we begin, it must also be noted that this document is not meant to be the definitive source for addressing all data collection errors. Interested readers should consult Kerlinger (1986), Pedhazur and Pedhazur-Schmelkin (1991), and Smith et al. (1986) for more information. These sources can be found at most university research libraries. Additional information can be obtained in a recent data management guide published by the Air Transport Association (1998).

## Research Design

Many errors in scientific data sets occur because of improper planning prior to actual data collection efforts. With a few moments thought to the future, a number of these problems can be

avoided altogether. At this stage, decisions must be made with regard to pilot testing of items, changes in questionnaire/item format, identifying necessary (but unknown) information, and the integration of data from multiple sources. Each of these areas will be discussed in turn.

1. Ensure that scale items are meaningful to respondents.

Whenever a new scale is developed, it should be "pilot tested" before it is put into official use. In other words, the scale should be administered to a group of people who are similar to those who will eventually complete the scales. For example, if a new Line Oriented Evaluation (LOE) scale for the 737 fleet has been developed, then it should be administered to a group of 737 Standards Captains prior to implementation. The reason for "pilot" testing is quite simple: the Standards Captains may interpret the scale item stems quite differently than originally intended by the scale developers. Suppose an item in a "Takeoff to Cruise" event set is written as follows:

> *"The PF considers using the autopilot to reduce workload and avoid task saturation. Autopilot engagement is clearly communicated."*

Typical response options for this item may include "Not Performed", "Partially Performed", and "Performed". The scale developer may have a very clear notion of what is meant by its components. For example, the developer may interpret the phrase "considers using the autopilot" as clearly meaning "mentions out loud the possibility of using the autopilot". However, the Standards Captain, who must actually use the rating instrument, might interpret any number of nonverbal behaviors as indication that the PF considered using the autopilot. Likewise, the scale developer might assume that the phrase "clearly communicated" entails acknowledgment from the PNF; but since this is not stated in the item stem, the Standards Captain might not require acknowledgment.

Such potential problems should be eliminated before the scale is put into operational use. This

can only be done through some form of pilot testing. Once those who will use the instrument appear to interpret the items in the scale as intended by the scale developer, then and only then can data collection proceed.

2. Exercise caution when altering previously used items.

Over time, items on some data collection instruments become obsolete, and must be altered or discarded. This is only natural; as procedures and technologies develop, the questions that are used to evaluate performance must also change. Nevertheless, coordinators of large, ongoing data collection efforts must carefully consider the costs associated with modifying existing data collection instruments. For example, suppose that the item mentioned above (the autopilot question) had been in use for a couple of years. After its ambiguity became a concern, the item was rewritten as:

> *"The PF verbalizes the possibility of using the autopilot to reduce workload and avoid task saturation. Autopilot engagement is clearly communicated and acknowledgment is received."*

This new item may indeed be less ambiguous. However, there is also a cost associated with modification; data collected using the old item cannot be directly compared with the data collected using the new item.

Suppose that, during two years of usage, the average score on the old item was approximately 3.5 (on a four-point scale). Suppose further that the average score during the first year of data collection with the modified item was approximately 3.0 (again, using a four point scale). Does this suggest that performance has decreased in the past year? Or is this difference simply a function of the change in wording? Quite simply, there is no way of knowing which of these explanations is correct. Scientifically, this is known as "confounding". Because changes in both the item wording

and the average scores occurred during the same time period, it is impossible to disentangle their effects.

There are four potential solutions to this quandary. The first option is to leave the item in its original form. This leaves the data analyst with an inferior item, but one that provides consistent data over time. The second option is to conduct interviews with Standards Captains, in order to obtain their opinions on the likely impact of such modifications. If they suggest that such changes are likely to make the item more difficult, then the observed lower scores might be expected. The third option is to include both items in the LOE. If scores on the old item remain the same, while scores on the modified item drop, the difference can be attributed to the item modifications. The final option is to follow-up the data collection effort on the new item with a data collection effort using the old item. If this yields the same drop in scores over time, then the drop would appear to be legitimate and not simply a function of item wording.

In summary, scale developers must carefully balance the costs and benefits of making changes to the data collection procedure. Many times, however, the benefits of consistency over time outweigh obtaining a marginally superior item. Nevertheless, the decision to make such changes must be made on an item-by-item basis after considerable deliberation. Lastly, it should also be noted that preliminary testing is just as important for modified items or scales as it is for new ones. As suggested earlier, even small changes in wording can result in large differences in interpretation.

3. Clearly document all changes to items and scales.

If items or scales are modified, it is imperative that the alterations be meticulously

documented. This documentation should include: 1) the previous form of the item or scale, 2) specific changes to the item or scale, 3) detailed justifications for the alterations, and 4) the exact date when the new item was put into operational use. Without these essential pieces of information, it may be impossible for the data analyst to track changes in performance (in the database) over time.

4. Identifying necessary, but unknown information.

Many times, substantial research questions are identified "post hoc" (after the fact). Suppose that during the collection of LOE performance ratings, various Standards Captains notice that relevant personality characteristics, such as conscientiousness, play an important role in how well pilots perform CRM-based behaviors. Acting on such an assumption, the Fleet Captain mails a personality questionnaire to all of the pilots in his/her fleet. After tabulating the pilots' personality scores, the Fleet Captain then compares personality scores with CRM performance ratings. After identifying a positive relationship between these two scores, the Fleet Captain asserts that conscientiousness scores are in fact causing higher CRM performance.

Although intuitively appealing, this assertion would be scientifically questionable. Such an assertion violates one of the basic principles of the philosophy of science. Specifically, for one variable to cause another, it must precede it in time. Even for traits such as personality characteristics, which are hypothesized to be stable over time, it is quite possible that the pilots' performance evaluations on the LOE could have influenced their self-reports on the personality questionnaire.

To address such theoretical (and practical) questions, fleet personnel and applied researchers must engage in strategic scientific planning. Specifically, they must anticipate likely research

questions, and attempt to collect such information as soon as possible, in order to test such hypotheses in the future. The lesson here is that to answer substantive research questions, one must have the appropriate data to do so. Therefore, it is imperative that applied researchers plan for the long-term data collection efforts well in advance. Collaboration between applied researchers and academicians could be of great value in making this process more manageable.

5. Ensure the ability to integrate different sources of information.

Relational databases, such as Microsoft Access®, provide data analysts and applied researchers with a variety of advanced features, such as the ability to quickly and easily link information from multiple databases. Given that carrier personnel may have already accumulated a vast array of data (pilot background/experience databases, pilot performance databases), it may be possible to use such data to answer substantive research questions. For example, one goal may involve comparing pilot background/training data with "First Look" maneuver ratings or Line Oriented Evaluation CRM performance ratings. Among other things, such information could provide valuable information about the efficacy of the recurrent training program, and/or possible gaps in the curriculum.

To link such databases, however, both the First Look and LOE data must be in formats that are amenable to merging. At a minimum, it is essential that unique crew identification numbers be used consistently throughout the data collection process. This may require assigning each crew a unique identification number when they enter the training program, and using that number on all rating forms throughout the training process. This crew identification number would then be used when entering performance evaluations in the pilot performance database. Or, it may simply necessitate entering entire "packets" of data (MTV, LOFT, and LOE performance ratings) into the

database at the same time. Regardless of the technique used, it will only be possible to answer such research questions by consistently using unique crew identification numbers throughout the training program.

Along a similar line of reasoning, it is necessary to ensure that the items being compared are in fact, logically comparable. For example, if a researcher wishes to examine performance on specific "First Look" maneuvers, and to compare such performance ratings to specific items on the Line Oriented Evaluation, both sets of items must have some logical basis for being grouped together. Items may be grouped together because they require similar skills, occur during similar phases of flight, and so forth.

Nevertheless, it is highly probable that the scale developers who create the LOFT (Line Oriented Flight Training) items are not necessarily the same scale developers who create the Line Oriented Evaluation items. Therefore, it is essential that these professionals coordinate their efforts by engaging in strategic planning, so as to ensure comparability. Such coordination projects may necessitate including other carrier personnel, such as Fleet Captains and Quality Assurance managers. Regardless of who is included, the goal is the same: to achieve "fit" between the various data collection efforts, so as to answer theoretically-meaningful questions which are important to the carrier as a whole.

**Data Collection, Data Entry, and Data Cleaning**

Much like the research design phase, the processes of data collection, entry, and cleaning present unique challenges. These challenges include ensuring the efficient distribution of labor,

following-up on problematic responses, and double-checking computer data with source documents. Each of these issues will be discussed in turn.

6. Distribution of tasks: Who does what

Data analysis is a very complex and specialized field. For example, large data sets may be collected by one group of professionals, and analyzed by another group of professionals. In other words, the individuals analyzing the data may have little or no involvement in the development and implementation of the measures. This is a mistake for two reasons. First, a data analyst who is completely removed from the design and distribution of measures may not recognize a problematic value or set of values when he/she sees it. For example, suppose two consecutive items on the demographic portion of a measure relating to the 737 fleet are:

1. *Please circle option 'a' if you are a 737 Captain and option 'b' if you are a 737 First Officer.*

2. *Please estimate the number of years that you have been in the 737 fleet.*

Suppose that a given person responds "b" on the first question and "20" on the second. This particular combination is not especially likely, but a data analyst who lacks knowledge of the measure and its implementation might not perceive the need to follow-up. Although this may be an exaggerated example, the point is as follows: data analysts who are not involved in the development of measures and the day-to-day data collection process are much less likely to recognize problematic data points than those who are included in the process.

The second reason for keeping a data analyst involved in the data collection effort is somewhat more subtle. Every large data collection effort has certain nuances and idiosyncrasies that only someone intimately involved with the development and collection process could understand.

Consider once again the “autopilot” question.

Suppose that the Standards Captains providing these ratings don’t insist that the pilots being evaluated actually verbalize the possibility of engaging the autopilot. Further, suppose that they don’t grade down if acknowledgment of autopilot engagement is not received. Thus, a pilot can receive a top score on this item simply by engaging the autopilot and letting the crew know that this has happened. If this were the case, then anything less than a top score would suggest a fairly serious problem. If the data analyst were involved in the data collection effort in some capacity, then he/she might appreciate the importance of maintaining a high cutoff for the autopilot question. Instead, if the analyst had no dealings with those collecting the data, less-than-perfect scores on this item might not be recognized as a potential problem (and thus not emphasized in their report to fleet personnel).

This is not to say that a single person should be responsible for all measure development, data collection, and analysis procedures. For large data sets, this would be impossible. Rather, it is important that the data analyst maintain enough familiarity with the ongoing data collection processes to recognize issues that aren’t necessarily obvious from the numbers themselves. It is worth noting, however, that involvement is a two-way street. Just as data analysts should be involved in collection, those responsible for the data collection should be involved in their analysis, to ensure accurate interpretation of the results. As expected, this presents special problems in large, multi-site data collection efforts.

7. Follow-up on problematic responses.

Occasionally, parts of the data collection measures will be filled out inappropriately. For example, responses to different questions may be inconsistent with one another, such as in the

"Captain vs. First Officer" and "Fleet Experience" questions mentioned previously. More commonly, certain items will be omitted, either because they were misunderstood or inadvertently skipped. Whatever the case, these problematic responses and omissions must be followed-up as soon as possible. Quite simply, the person who completed the measure may forget why the item(s) in question was treated as it was, or what the correct answer should have been in that particular instance. Furthermore, the process of following-up might reveal an ambiguous item; an item that is likely to be systematically omitted, if not modified. Obviously, the sooner this follow-up occurs, the better.

8. Double check the computer file

Data entry can be very tedious. As a result, there is a tendency for organizations to streamline the data entry process as much as possible. Unfortunately, this often leads to transcription errors. Even under ideal circumstances, the most meticulous data-entry person is bound to make occasional mistakes. Indeed, even data sheets that are computer scored can produce mistakes due to the inability of the scanner to correctly interpret illegible marks on the sheet. Therefore, it is extremely rare that large amounts of data will be hand-entered without mistakes.

Thus, all data entered into computer files must be double-checked against the original data collection forms. At a minimum, database entry forms should contain validation rules which call attention to incorrect keystrokes, such as "out of range" responses. For example, if the data analyst is entering LOE performance data (which is based on a four point scale), and actually types in the number "7", instead of "4", the computer would provide an audible alert, and prevent further data entry personnel until the error is corrected. As stated earlier, data entry errors can lead to problems

during the data analysis phase; problems that may be extremely difficult to detect and correct.

**Data Analysis**

9. Coding missing data

Often, it is advisable to assign a unique identification number to missing data instead of simply leaving a field blank in a data set. For example, consider the following contrived data for seven people on three variables:

| ID Number | Variable #1 | Variable #2 | Variable #3 |
|---|---|---|---|
| 1 | 3 | 2 | 5 |
| 2 | 4 | **9** | 5 |
| 3 | 9 | 4 | **9** |
| 4 | 4 | 4 | 5 |
| 5 | 3 | 5 | **9** |
| 6 | 4 | 3 | 4 |
| 7 | 3 | 2 | 4 |

Respondents 2, 3, and 5 have not provided answers to all items. However, instead of leaving these spaces blank, they have been assigned the value of "9". Note that the value "9" is entirely arbitrary; it is used only to indicate the absence of data point on a given variable. While "9" is most commonly used within the social science community, any number that lies clearly outside the range of possible responses to a given item can be used.

Since the choice of values for such purposes is entirely arbitrary, there is no reason to use more than one value to indicate missing data. In a large data set, however, there may be many variables that all contain some amount of missing data. To minimize confusion, it is advisable to be as consistent as possible. To alleviate misunderstandings between the data entry specialist and the data

analyst, missing data values should be included in the data code books, so that the data analyst can structure his/her analyses properly.

Missing data coding schemes provide a number of advantages. Unfortunately, a discussion of these advantages is extremely technical, and lies outside the scope of this document. For example, this would allow the data analyst to perform statistical tests to search for systematic patterns of missing data in a given evaluation (e.g., LOE, LOFT).

10. Reverse code immediately.

Quite often, questionnaire items are "negatively worded". That is, they are phrased in a negative fashion in order to minimize certain response biases. For example, suppose a given form has three items designed to measure the quality of a pilot's technical performance. Two of the items might be worded positively (Pilot interpreted data appropriately) while the third might be worded negatively (Pilot failed to perform operations in the appropriate order). All three items are scored on a scale that ranges from "1" (Strongly Disagree) to "5" (Strongly Agree). These items are all designed to measure the same thing: technical performance.

However, a rating of "5" on the positively worded items indicates high technical quality, while a rating of "5" on the negatively worded item indicates low technical quality. This difference is dealt with in the analysis stage by "reverse coding" the negatively worded item. That is, the values for these item are re-coded so that a "5" on this item also indicates high technical quality. Thus, the item is receded so that a value of "5" on the negatively worded item is re-coded to a value of "1", a value of "4" is re-coded to a value of "2", and so on.

Large data sets typically involve several negatively worded items, all of which need to be re-

coded. In situations where reverse coding is necessary, it is wise to perform the reverse coding immediately. If the reverse coding is not performed immediately, this critical part of the data analysis phase may be forgotten. The end result would be statistical output that (for certain items) doesn't make logical sense. Provided that the data analyst catches the error, these nonsensical analyses will be investigated, and the unrecoded item will usually be identified as the cause of the error.

Nevertheless, it is much easier, and altogether less time consuming, to ensure that the reverse coding happens immediately upon receipt of the data set by the data analyst. As stated earlier, items that need to be reverse-coded should be clearly indicated in the codebook.

11. Frequency analysis

Once the data have been entered into the computer, frequency analyses should be conducted for all numerical variables. In a frequency analysis, tallies are made of the number of times that each value occurs for a given variable. Consider the negatively worded item mentioned above. A frequency analysis of responses to this item might look like the following:

| **Value label** | **Value** | **Frequency** |
|---|---|---|
| Strongly Disagree | 1 | 115 |
| Disagree | 2 | 130 |
| Neither Agree Nor Disagree | 3 | 40 |
| Agree | 4 | 10 |
| Strongly Agree | 5 | 3 |
| Missing | 0 | 2 |
| **Total Number of Responses** | | 300 |

Frequency analysis can be useful for various analytic functions, but at this initial stage, it is

most useful for identifying "out of range" values. The item in question has five possible response options: 1, 2, 3, 4, and 5. However, two people responded with zeroes. These two people may have misread the instructions, or they may have simply been careless. Regardless of the reason, these values of zero are out of range. They have no meaning on the 5-point scale in question, and must be either replaced with appropriate values or discarded. Otherwise, they may inappropriately influence the results of any analyses involving this item.

12. Get to know the sample from which the data came.

It is important to become familiar enough with the sample to be able to detect mistakes in other descriptive statistics. For example, suppose that initial analyses of a sample of professional pilots reveals an average age of 55.17. There is nothing particularly strange about this value in the abstract. In a given instance, however, it might be clearly incorrect to a person who knows that all pilots are less than 60 years old, due to FAA regulations. A data analyst who is entirely removed from the sample would have no way of knowing that the value yielded by the analysis is clearly mistaken. Only a person who is familiar with the pilot group from which the data came could detect the problem, and report such a problem to carrier personnel.

13. Scrutinize descriptive statistics carefully.

There are a variety of problems in a data set that cannot be detected with a simple frequency analysis. Consider the age question mentioned earlier; suppose an initial analysis reveals the mean for this item to be 28. This is clearly within the range of normal values for the variable, so why should this concern us? The problem is that the likelihood of any variable having a mean that is an integer (a number containing no decimal points) is extremely small. It is far more likely that the data field has

been formatted erroneously; in essence, all values that follow decimal places have been discarded by the computer. Thus, instead of 28.25, the computer simply yields a value of 28. This is just one example, but there are many other examples of potential problems that can be detected by a careful examination of descriptive statistics.

**Data Archiving/Warehousing**

14. Electronic Storage

Quite often, data files are provided to outside researchers for analysis and/or re-analysis. Unfortunately, because these individuals were not intimately involved in the research design and data collection stages, it may be difficult for them to "get up to speed". Therefore, when providing data files to outside researchers, it is essential to provide them with adequate information for use in performing their analyses. At a minimum, four files must be provided to the researcher. These include a copy of the actual data file, a data codebook, electronic copies of all evaluation forms, and a "readme" file.

A data codebook is an electronic file that contains a description of each variable within the database file, the appropriate range of values for each variable, as well as missing data codes. Electronic copies of all evaluation forms should also be included. Quite simply, the researcher needs to know how to structure the appropriate queries, in order to convert the data from reduced normal form (e.g., Access, FoxPro, etc.) to statistical normal form (e.g., SPSS, SAS, etc.). Typically, data database files will be transferred electronically, such as via File Transfer Protocol (FTP), or as electronic mail attachments. Therefore, by including an electronic copy of the evaluation form with the actual database, the researcher is assured of obtaining the necessary information for performing

his/her analyses. For example, the researcher can print copies of the evaluation forms, and use these forms to identify the number of event sets (and their titles), and the number of topic ratings per event set (and their titles). Lastly, after generating the appropriate queries, the researcher can then double-check that the query results match that which would be expected from the grade sheets.

Nevertheless, it is understood that the individuals who typically create the evaluation forms are not the same as those who create the databases. Furthermore, the original evaluation forms may be stored on a variety of platforms (Windows, Macintosh, etc.) or file formats (Word, WordPerfect, PageMaker, etc.). Note that most modern word processing programs, such as Word97, can import virtually any file format. If it is impossible to locate a disk-based copy of the evaluation form, it may be necessary to optically scan the form (using a scanner with OCR software) manually. Another alternative would be to include a document which contains hypertext linkages to the evaluation form on the carrier's web site.

The last file in the archive should be a "readme" file. Typically, this file is stored in ASCII format. ASCII is a text file format that can be read by virtually any computer. This file, typically called "readme.txt", includes the names of all files in the archive, a description of each file's format, as well as contact information, such as the database programmers' phone number (should problems arise). Typically, once an archive is un-compressed, the "readme" file is the first one read by the recipient.

15. Keeping hard copies of evaluation forms.

After the data have been entered into a computer file, it is tempting to discard hard copies of the data. After all, hard copies can take up a lot of space, and the computerized version can be

backed up ad infinitum. Unfortunately, errors in large data sets often go unnoticed until well into the analysis stage. For example, no matter how often a data set is checked for errors, transcription problems still arise. If such problems are detected, they must be checked against the original evaluation forms. If these data have been retained, then the problem can be easily resolved. If not, then the data point in question may have to be deleted without replacement.

Of course, there is usually limited space available for storage of paper copies of documents. Thus, it essential to develop a system for retaining original data sheets. If data are typically analyzed within three months of input, then it might make sense to keep all hard copies for four to six months. Then, if no problems have arisen, one can discard the hard copies with a clear conscience.

**Summary and Conclusions**

The analysis of large-scale databases, especially within multi-site research efforts, can be fraught with problems. Therefore, it is essential to mitigate their effects by means of careful planning, meticulous attention to detail, and verbose documentation procedures. A number of "best practice" guidelines were addressed in this document. These guidelines were developed collaboratively during a three-year training evaluation project between academic researchers and a major air carrier. It is believed that adhering to such guidelines can help future research efforts to proceed much more smoothly and efficiently.

References

Air Transportation Association. (1998). Data management guide. Unpublished technical report. Data Management Focus Group.

Kerlinger, F. (1986). Foundations of behavioral research (2nd Edition). Orlando, FL: Holt, Rinehart, & Winston.

Pedhazur, E. J.,& Pedhazur-Schmelkin, L. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Earlbaum Associates.

Seamster, T. L., Boehm-Davis, D. A., Holt, R. W., & Schultz, K. (1998). Developing advanced crew resource management (ACRM) training: A training manual. Federal Aviation Administration, Office of the Chief Scientific and Technical Advisor for Human Factors, AAR-100; Washington, DC.

Smith, P. C., Budzeika, K. A., Edwards, N. A., Johnson, S. M., & Bearse, L. N. (1986). Guidelines for clean data: Detection of common mistakes. Journal of Applied Psychology, 71, 457-460.