

# **Responsible Data Management: Avoiding Data Disasters**

Shelley M. MacDermid  
College of Consumer and Family Sciences  
February 13, 2007

# Data Disasters

- Data full of unexplained errors
- Inability to reproduce the results of earlier analyses
- Data so poorly documented that no one can figure out how to use them
- Data that contain identifying information about research participants
- Management processes that fail to protect data integrity
- Management processes that fail to prevent duplication of effort
- Data set up in such a way that they can't be rearranged to address new research questions

# Four Key Principles

Minimize error

Maximize integrity

Minimize cost

Maximize flexibility

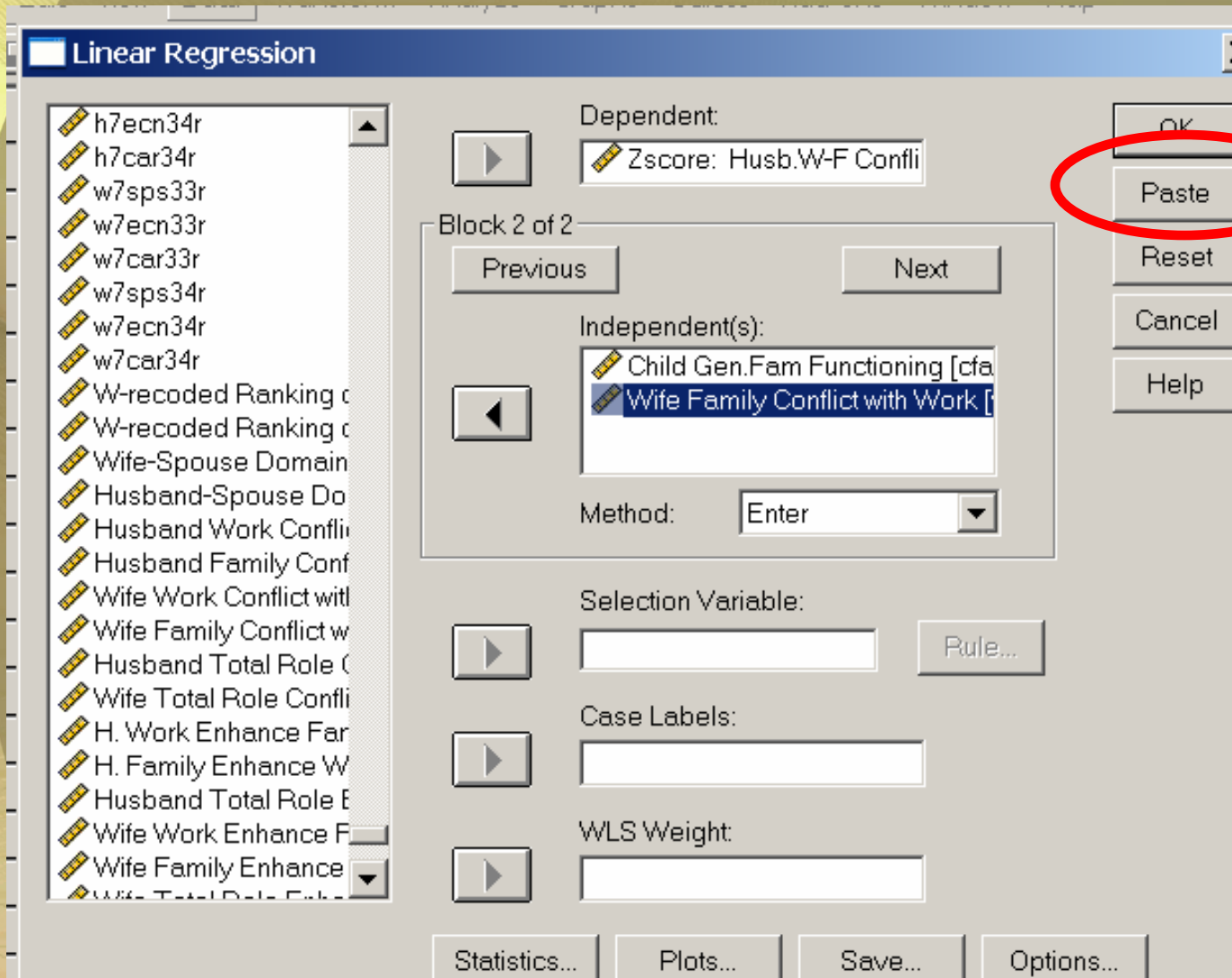
# Four Key Principles

Minimize error

# Avoiding Data Disasters

- Never rely on memory
  - Create written records of all data manipulations -- Use syntax, not 'point and click'
- Protect data from unintentional or malicious changes
  - Password access for reading and writing
  - CDs for distribution
  - Limited write access to 'master' files

# The 'Paste' Button is Your Friend



**REGRESSION**

**/MISSING LISTWISE**

**/STATISTICS COEFF OUTS R ANOVA**

**/CRITERIA=PIN(.05) POUT(.10)**

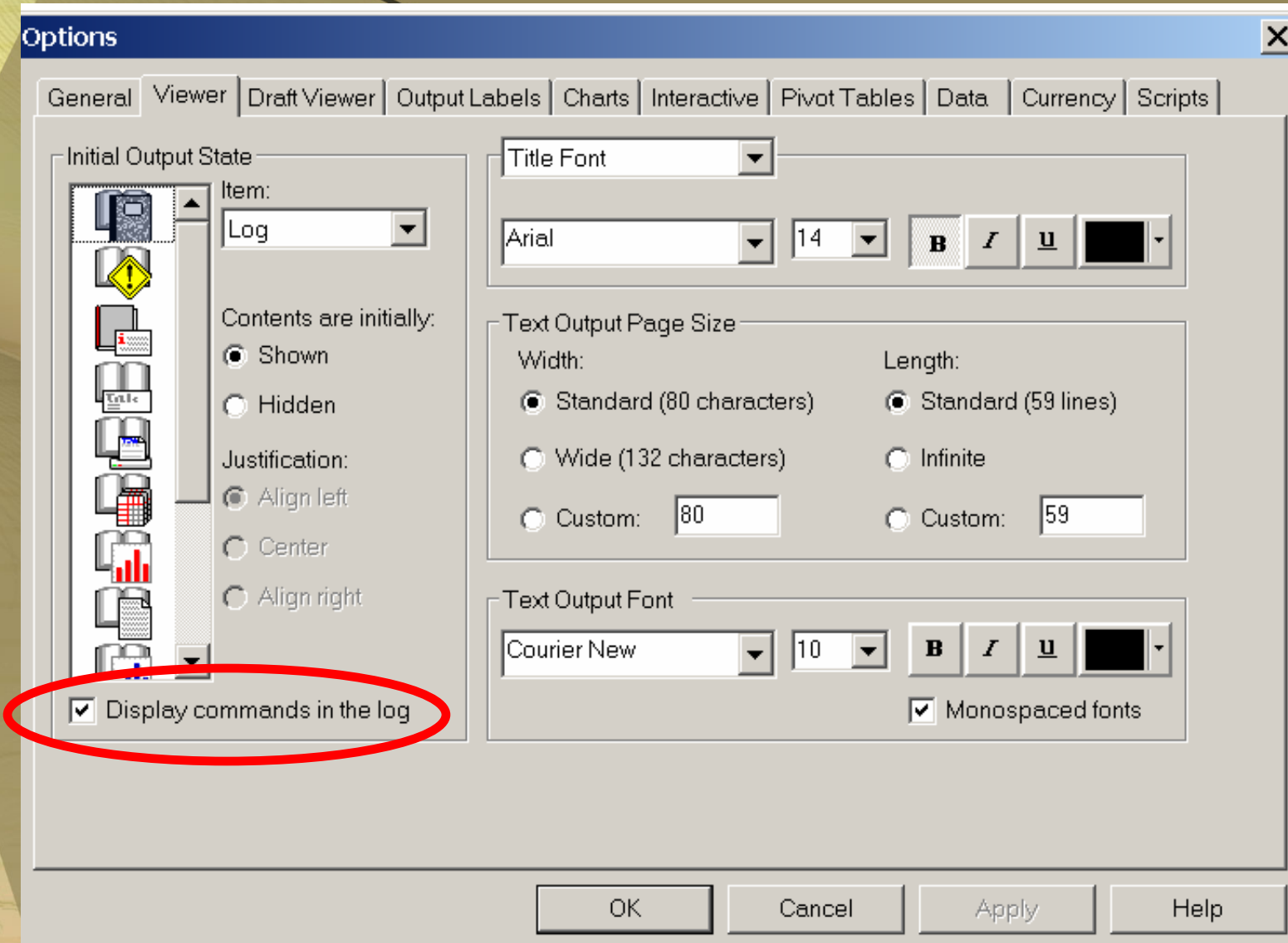
**/NOORIGIN**

**/DEPENDENT Zhpwfneg**

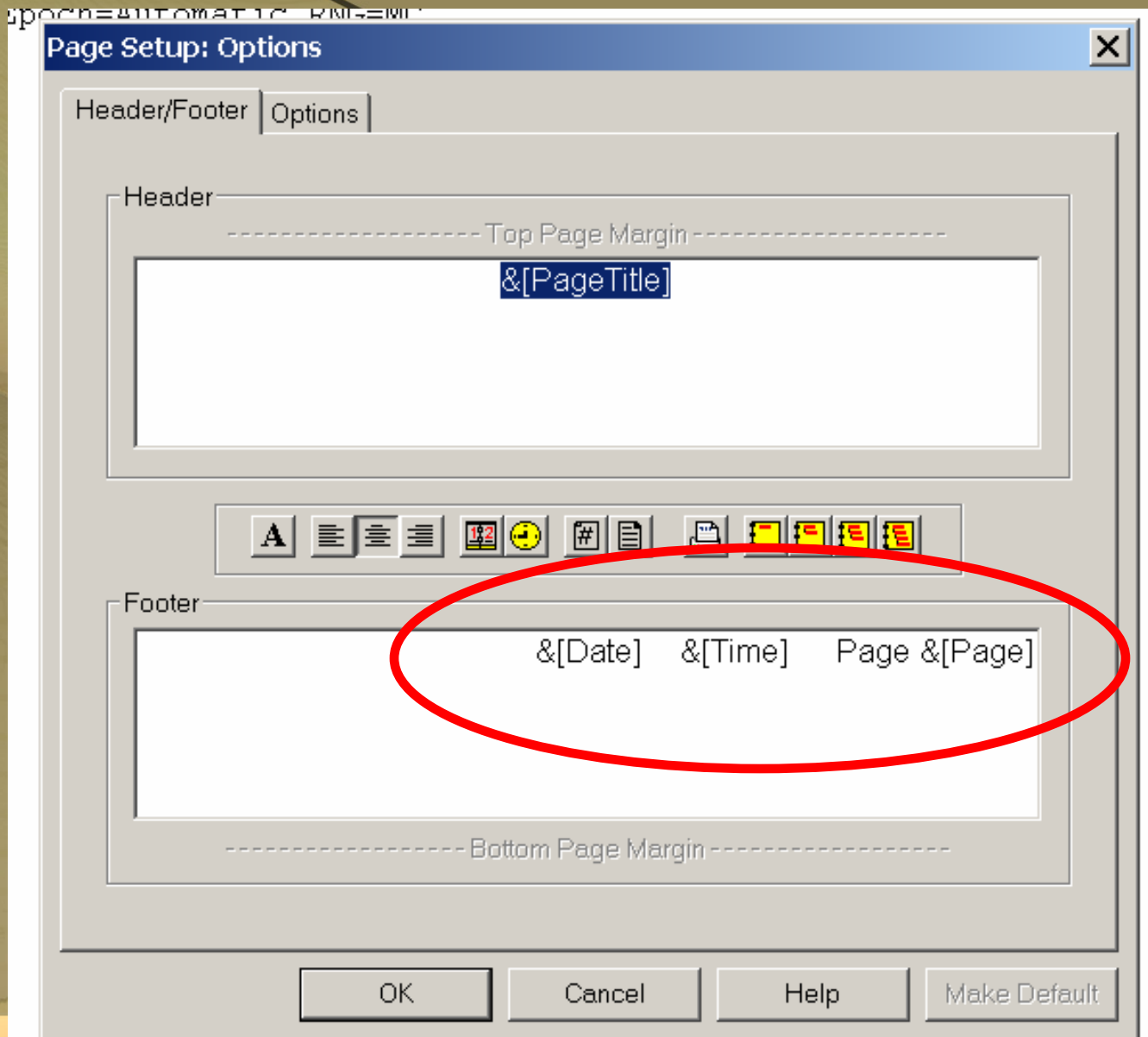
**/METHOD=ENTER w9renhnce momaccpt**

**/METHOD=ENTER cfamfunc w9famcwrk .**

# There Are Easy Ways to Help Yourself



# There Are Easy Ways to Help Yourself



# Four Key Principles

Maximize integrity

# Avoiding Data Disasters

- Monitor quality of BOTH processes and data
  - Create a culture of transparency in research processes that discourages privacy in data work – for anyone
    - Data work is public – syntax files written to “public” (i.e., research team) standards, stored in “public” (i.e., to team) locations
  - Prepare for the possibility of malice. What if someone walked into your lab off the street (or through cyberspace) – what damage could they do?
    - Could they alter electronic data files?
    - Could they destroy data?
    - Could they find identifying information about participants?
  - Assume people on your team might be careless from time to time, or that someone might be malicious – make sure regular processes will find that
    - Send interviewers out together – check for ‘drift’
    - Have interviewers review one another’s work for errors or lack of clarity
    - Spot check

# Four Key Principles

Minimize cost

# Avoiding Data Disasters

- Avoid missing data
  - Data collection processes should minimize missing data
  - No data should be missing without a recorded reason – respondent refused, instrument failure, technician error, etc. Codes added at the time will save expensive detective work later on
- Estimate costs thoroughly
  - Consider the entire data cycle so that data management does not get short shrift
  - Thoughtful pilot-testing of data collection, data management AND data analyses

# Write Self-Documenting Syntax

\*\*\*\*\*

```
*
*                               MASTER SCALES SYNTAX
*   USE THIS SYNTAX FILE FOR RE-RUNNING AGAINST A NEW,UPDATED MASTER DATA FILE!
*   Running syntax on Master 36 Dual Families dated 9-23-04.
*   This syntax reflects the scales and statistics that Rona had previously run plus new ones created for NCFR 2004.
*   This syntax also reflects new scales that Tracy computed from our documentation that have been checked by Rona.
*
*   The file has been reorganized by section. Each scale lists the husband variables first and then the wife.
*
*   Any recoding is done at the beginning of a section. All recodes are renamed with an ' r ' at the end of the variable
*   name.
*
*   This syntax was created to make scale scores for husbands, wives, and children. It is not formatted for the person
*   level data file.
*   To use on the person level data file, a rename or find and replace function must be ran for h and w scale names.
*
*   PLEASE READ THE COMMENTS AT THE BEGINNING OF EVERY SCALE!
```

\*\*\*\*\*

```
** note-alphas are too low to use as a scale.
RELIABILITY
/VARIABLES= hpr152a hpr152c hpr152f hpr152h
/FORMAT=NOLABELS
/SCALE(ALPHA)=ALL/MODEL=ALPHA
/STATISTICS=DESCRIPTIVE SCALE CORR
/SUMMARY=TOTAL MEANS VARIANCE CORR .
```

# Four Key Principles

**Maximize flexibility**

# Avoiding Data Disasters

- Never do any data manipulation you can't UNdo
  - Never recode a variable over itself – always create a new variable with the recoded values
  - Always keep copies of data files in their intermediate forms when manipulating data structure so you can retrace your steps
  - Include thorough indexing information on every record (e.g., case number, date, data collector, location, etc.) – makes it easier to rotate cases and variables in the future

# Case Study Part A

Dr. Edwards is a clinical psychologist and researcher. Several years ago, looking at the shelves and drawers filled with patient files, he came up with the idea of developing a multimedia, fully integrated database that would allow physicians and researchers to store, analyze and query patient/subject information quickly and easily.

The MEDUSA database allows researchers to record and track all aspects of an experiment including patient samples and records (everything from name and address to CAT scans), experimental reagents, protocols, raw data and primary analysis.

When Edwards began developing MEDUSA, he did not inform his patients or ask their permission to be included. He believes that storing the data in MEDUSA is equivalent (if not superior) to storing it in folders in filing cabinets and is simply the best way for him to provide care to his patients. The database is located in Edwards' lab on only one computer, which is accessible over the web.

Edwards would like to market his database tool, but it needs beta-testing with a range of data types and formats not commonly encountered in psychology (e.g., DNA sequence data or results from animal breeding experiments). Edwards gives presentations in several other labs on campus to advertise and recruit beta-testers. During demonstrations, all patient names are encrypted and family relationships obscured. Edwards navigates through MEDUSA, showing off the ease with which one can toggle between a patient's blood chemistries and the results of behavioral tests.

- 1. Is there a substantive difference between paper records and MEDUSA?*
- 2. Is Edwards justified in using patient information for database development and promotion? Why or why not?*

# Case Study Part B

Edwards convinces three labs with large ongoing projects to import their data into MEDUSA, helping him work out bugs and continue to develop the design and marketability of his database.

In order to devise ways to import their data, the staff of the other labs must learn the data structures and file formats in MEDUSA, as well as how to manipulate the encryption utility. To do this, some individuals must have full access to the database, which means they also have full access to Edwards' data.

Each beta-lab programmer is provided with the encryption key. Amy is one of these programmers. Amy now has access to complete patient files and experimental data in the database. Periodically, her supervisor asks her to update the rest of the lab on her progress. During her lab presentations, it is easier to demonstrate MEDUSA's functionality without the encryption in place. Although Edwards is working on it, the key currently must be entered each time a query is submitted, which is cumbersome and slow for demonstration purposes.

- 1. If Edwards had sought his patients' informed consent for use in the database, what risks and benefits would have to be disclosed?***
- 2. Given this additional information, do you feel differently about Edwards' use of patient information in the development and promotion of MEDUSA? Why or why not?***

**FALL 2007**

**CDFS685D:  
Data Management in Social Science Research**

- [shelley@purdue.edu](mailto:shelley@purdue.edu)