Computer Analysis of Survey Data --
File Organization for Multi-Level Data

Chris Wolf
Ag Econ Computer Service
Michigan State University
1/3/90

Large-scale socio-economic surveys require a tremendous amount of time to set up, conduct, and analyze. The effort required for the analysis of such complex data sets is often much greater than it should be, because of the typical researcher's inexperience at this task. Perhaps the most misunderstood aspect is the management of multiple levels of data within the same survey. If the different levels of data are identified early in the project, and they are managed properly through the different stages of analysis, the researchers will be much more productive.

This discussion will focus primarily on the organization of computer data files and how it relates to survey design and analysis. The major issue is the identification of different levels of data that are to be collected and analyzed in the survey. A secondary, but strongly related, topic is the use of key variables within the data files.

Data File Concepts

A data file for survey analysis is made up of cases, variables, and values. This is most easily conceptualized as a matrix, where each row contains all the data for one entire case (or observation), and each column contains all the data for a particular variable (or question in the survey). In table format, this would look as follows:

Case Number	Variable Name VILLAGE	HH	Q1A	Q1B	Q2	Q3	Q4A	Q4B	Q4C
1	1	1	10	23	2	7	1	3	1
2	1	2	3	34	2	3	2	2	4
3	1	3	12	40	1	2	1	2	1
4	1	4	15	32	2	4	1	1	1
5	2	1	2	26	2	2	1	3	2

Each case in a data set contains information about a specific physical or logical entity. Some examples of entities that might correspond to a case include a household, a person, a town, a crop, a sale, or a parcel of land. Each of these is, in some sense, a "thing" that might come under scrutiny in a survey.

These examples seem pretty simple, but in a complex survey, as we will see, it can be tricky to make sure that a case really represents what you want it to represent or what you think it represents. One of the main goals of this discussion is to enable you to clearly and unambiguously define the meaning of a case in each of your data files, and verify for yourself that your definitions are right.

The variables in a data file contain information about the attributes or characteristics of each case. Some of the variables that might describe a case include age, quantity harvested, number of animals owned, amount borrowed, and where purchased. Obviously not every variable is appropriate to every type of entity; that is one important key to determining a proper file organization.

For computer analysis, variables must be given variable names, usually composed of letters and numbers like those shown across the top of the table above. Cases do not have names, but can be referred to in several ways, with one simple one being a sequential numbering scheme like that shown above. If we refer to a particular variable, we are referring to an entire column of the matrix, with all cases included. Similarly, if we refer to a particular case, we are referring to an entire row of the matrix, with all variables included.

If we look at the intersection of a row and column in the matrix, we find a value. For example, in the table above, variable Q3 has the value 7 for the first case shown, the value 3 for the second case, and so forth. Be careful not to confuse the variable names with the values the variables take on for each case.

Data Organization Examples -- Wrong vs. Right

There is usually only one right way to organize a particular set of data into variables and cases, while there are many wrong ways. So that you can see where we are headed, let's look at an example that demonstrates a common problem in file organization, showing both the wrong way and the right way to handle it.

Suppose you had a survey that looked in part as follows: (This is not intended to show a finished survey format, but is just a rough representation for demonstration purposes.)

Village code
Household code

1. Fields Planted in 1987

Field #	Distance from House	Size	1987 Area Planted	Plowed?	Planting Method	. . .	How Acquired	Length of Ownership
1	_____	_____	_____	_____	_____	. . .	_____	_____
2	_____	_____	_____	_____	_____	. . .	_____	_____
3	_____	_____	_____	_____	_____	. . .	_____	_____
4	_____	_____	_____	_____	_____	. . .	_____	_____
5	_____	_____	_____	_____	_____	. . .	_____	_____

2. Primary source of income _____

.
.
.

13. How long have you lived here? _____

The table of data for "Fields Planted in 1987" allows for up to five fields per household. There are eleven questions about each field, starting with "Distance from House" and ending with "Length of Ownership". (For reasons of space, these are not all shown above.)

The Wrong Organization

One possible way to organize the data from this survey would be to store it all in a single file with variables defined as follows (for space reasons again, some variable names are written in the locations where the values would be on the actual questionnaire):

Village code VILL

Household code HH

1. Fields Planted in 1987

Field #	Distance from House	Size	1987 Area Planted	Plowed?	Planting Method	. . .	How Acquired	Length of Ownership
1	H1	H2	H3	H4	H5	. . .	H10	H11
2	H12	H13	H14	H15	H16	. . .	H21	H22
3	H23	H24	H25	H26	H27	. . .	H32	H33
4	H34	H35	H36	H37	H38	. . .	H43	H44
5	H45	H46	H47	H48	H49	. . .	H54	H55

2. Primary source of income H56

.

.

.

13. How long have you lived here? H67

This type of organization gives us a total of 69 variables, with one case per household. Because many households will have fewer than five fields, many of the variables will have missing values for some cases.

This is probably the simplest and most obvious way of organizing the data, but, unfortunately, it is also the wrong way. The main drawback is that it will make the analysis much more awkward and error-prone than it needs to be, as we will see shortly.

The Right Organization - Fields-Planted Data

A better structure to use for this survey would be to divide it into two files, with different parts of the data in each. The first part, the fields-planted information, should be entered into a file of its own, with 14 variables as shown below. (This division of the data might also make it desirable, particularly for data-entry purposes, to divide the questionnaire into two separate parts. To simplify this example, however, I will use the original questionnaire layout and numbering, showing only the portion that goes into the file under discussion.)

Village code VILL

Household code HH

1. Fields Planted in 1987

Field # FIELD	Distance from House F1	Size F2	1987 Area Planted F3	Plowed? F4	Planting Method F5	. . .	How Acquired F10	Length of Ownership F11
1	___	___	___	___	___	. . .	___	___
2	___	___	___	___	___	. . .	___	___
3	___	___	___	___	___	. . .	___	___
4	___	___	___	___	___	. . .	___	___
5	___	___	___	___	___	. . .	___	___

Notice first that we have removed all of the questions that do not apply to field planting. Also, the variable names are now placed at the top of the fields-planted table to indicate that each entire column now requires only a single variable. This reduction in the number of variables occurs because now each row in the table will become a separate case in our data file, rather than being combined with other rows into a single case. Thus, any household that farmed more than one field will be represented by more than one case in the fields-planted data file.

This is a very significant change, and one which must be understood fully in order to deal with multi-level data properly. To state it another way, different households may now have different numbers of cases, depending on the number of fields they planted. For example, a household with one field would have only one case, while a household with four fields would have four cases. The change in case structure also necessitated one other change in the form, the conversion of the Field column from being just a helpful notation on the form to being an actual variable called FIELD. As you will see later, this is to enable us to distinguish one case (a field) from another.

You can see that what we have done, in effect, is to trade variables for cases. The first proposed structure had 55 variables for the fields-planted data, all occupying a single case and representing a household. The improved data structure has fewer variables, but has more cases, each representing a field.

It's also important to understand that, although the questionnaire layout doesn't show it explicitly, each case must include the VILL and HH variables in addition to FIELD through F11. Compare the following matrix layout of the data file with the form at the top of the page.

Case #	VILL	HH	FIELD	F1	F2	F3	. . .	F10	F11
1	1	1	1	23	2	2	. . .	3	1
2	1	1	2	34	3	2	. . .	2	4
3	1	2	1	40	1	1	. . .	2	1
4	1	2	2	32	2	2	. . .	1	1
5	2	2	3	26	4	2	. . .	3	2

Notice that the farm designated as village 1, household 1 appears twice with identical entries for the VILL and HH variables, but with different values for the FIELD variable.

The Right Organization - Household Data

The remainder of the data would be stored in a second file, also containing (purely by coincidence) fourteen variables, as shown below. This file, unlike the fields-planted file, would contain only one case per household. It would basically be identical to the file we described above as the wrong organization, but with the fields-planted data removed.

	Village code	VILL
	Household code	HH
2.	Primary source of income	H1
	.	
	.	
	.	
13.	How long have you lived here?	H12

The division of the data into two files with different variable/case structures is desirable because the fields-planted questions represent a different level from the household questions that follow them. Don't be surprised if it's not immediately apparent to you why this is so. The remainder of this paper will attempt to explain the concepts behind this, to show you how to recognize this kind of situation in your own data, and to demonstrate how to handle it when you do.

Programming Effort Required -- Wrong vs. Right Organization

For a vivid demonstration of why the separate-file type of organization is preferred, let's look at a particular kind of calculation that might be performed on this data. Suppose we wanted to calculate two additional numbers for each household -- the average size of all the fields that they plowed and the average size of those fields they did not plow.

With the all-in-one file organization, this calculation would require the following commands in SPSS/PC+ (Statistical Package for the Social Sciences, a commonly used statistical analysis program):

```
COMPUTE NUM_PLOW=0.
COMPUTE TOT_PLOW=0.
IF(H4 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H4 = 1) TOT_PLOW=TOT_PLOW+H2.
IF(H15 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H15 = 1) TOT_PLOW=TOT_PLOW+H13.
IF(H26 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H26 = 1) TOT_PLOW=TOT_PLOW+H24.
IF(H37 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H37 = 1) TOT_PLOW=TOT_PLOW+H35.
```

```
IF(H48 = 1) NUM_PLOW=NUM_PLOW+1.
IF(H48 = 1) TOT_PLOW=TOT_PLOW+H46.
COMPUTE AV_SIZP=TOT_PLOW/NUM_PLOW.
COMPUTE NUM_NPLO=0.
COMPUTE TOT_NPLO=0.
IF(H4 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H4 = 0) TOT_NPLO=TOT_NPLO+H2.
IF(H15 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H15 = 0) TOT_NPLO=TOT_NPLO+H13.
IF(H26 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H26 = 0) TOT_NPLO=TOT_NPLO+H24.
IF(H37 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H37 = 0) TOT_NPLO=TOT_NPLO+H35.
IF(H48 = 0) NUM_NPLO=NUM_NPLO+1.
IF(H48 = 0) TOT_NPLO=TOT_NPLO+H46.
COMPUTE AV_SIZNP=TOT_NPLO/NUM_NPLO.
```

Pretty imposing, isn't it? And worse yet, if the questionnaire had allowed for up to ten fields rather than five, that would double the number of IF statements required! Not only is it tedious to enter all these statements, but there is tremendous potential for errors (in the variable names, for example) which might never be caught.

In contrast, the separate Fields-Planted file would require only a single statement to calculate the same variables:

```
AGGREGATE OUTFILE='AGGFIELD.SYS'
        /BREAK VILL HH F4
        /C_AVESIZ=MEAN(F2).
```

Even better news is that this statement would still do the complete job no matter how many fields there were! To be completely fair, I must point out that this single statement would not be completely equivalent to the 26 statements required for the other method. To get a comparable file structure for further analysis would require a total of five commands -- two PROCESS IFs, two AGGREGATEs, and a JOIN MATCH. Nevertheless, the savings in effort is substantial, and, possibly more importantly, the likelihood of errors is lower.

A further advantage of the division into separate files is that less disk space will be required because there is less waste. Remember that the single-file structure required 55 variables for the fields-planted data -- variables for which space must be allocated whether a household has five fields or only one. The separate fields-planted file has only fourteen variables, and only as many cases as are required for each household. Unless most households have the maximum number of fields, the separate file will require less space.

Key Variables

Before we talk about organizing different levels of data, we need to introduce the concept of key variables. This is a basic concept of data management that arises commonly in dealing with database programs, but is also applicable to survey analysis.

The word key is used here not in the sense of important or fundamental (although the key variables certainly are important), but rather in an analogy to the key for a lock or the key for deciphering a secret code. This analogy is used because key variables are an important instrument in helping you to extract the information you need from your data files.

Our earlier definition of variables is general enough to apply to key variables as well, but there's more to be said about this special class of variables. In a sense, key variables act more as identifiers than they do as attributes. Let me use an example to illustrate.

Suppose we performed a survey of the members of the Ag Econ Department at MSU. For such a survey, a case would obviously be a person. Now suppose I filled out this survey and answered some of the questions, as follows:

Employee No: 793278634
Name: Chris Wolf
Years employed: 16
Job classification: 7

Years employed and Job classification are good examples of the normal kind of variable we discussed in the section on Data File Concepts. They are attributes that describe me, the entity represented by this particular case. However, while they do describe me, they do not identify me, as distinct from someone else in the survey. It is quite possible, even likely, that there will be others in the survey who also have been employed at MSU for 16 years, as well as others with job classification 7.

The first variable, Employee number, is rather different. There is no other employee at MSU whose employee number is the same as mine, so that particular number is unique to me among all the people who might be included in the survey. This uniqueness is a very important characteristic of a variable, one which makes the variable very useful to us. This uniqueness lets the variable serve as a key variable.

Also note one other interesting characteristic of this variable -- unlike my length of employment and job classification, which tell an observer something about me, my employee number does not really describe me in any useful way. You will often find that key variables are not descriptive in the way that other variables are, although this is not a prerequisite for a key variable.

And what about the remaining variable, Name? It might be tempting to call it a key variable as well, because a person's name is fairly unique. There are two problems with this. First, a name is not unique; it is possible, although not likely, that there would be two employees with the same name. For the purpose of assigning key variables, uniqueness must be guaranteed by the design of the survey. If the possibility exists that the value of a variable will be duplicated on two unrelated cases, that variable must not be used as a key variable. Second, you should never use an alphabetic variable, such as a name, as a key variable in a survey data file. In fact, most statistical packages don't handle variables with alphabetic values very well, and you should usually avoid them for non-key variables as well.

So the requirement for a key variable is fairly simple -- it must take on unique values for each of the cases in the data file. In the hypothetical employee survey described above, there was a natural, pre-existing variable that was an obvious candidate for a key variable, but in most surveys this is not the case. You will usually have to create a key variable to suit your own data. This is done by developing a coding scheme that assigns ID numbers to each of the important entities in your survey.

A typical example of this would be a household questionnaire where you were going to survey 45 households. The simplest key variable to use here would be one whose values ranged from 1 to 45, perhaps called HH. With a key variable of this type, it is best to assign the values to the cases arbitrarily. You might assign them by the order they were chosen for the survey, or by the order in which they were interviewed, or any method that is essentially random.

There are many times when it's advantageous to use more than one key variable. Suppose those 45 households in the previous example were located in three different villages, with fifteen households per village, and that the study was going to look at, among other things, differences between villages. In this case it would be important to include a variable (VILL) identifying the village. Then the normal procedure would be to number VILL from 1 to 3, and to number HH from 1 to 15 within each village. Then VILL and HH together become the key variables for the household data file, because neither by itself is enough to identify a unique case. In surveys of any complexity, multiple keys such as this are more common than single keys.

Levels of Data

With the introduction to key variables out of the way, we can now return to the subject of data organization. One of the most significant characteristics of a file of survey data is its unit of observation or level. Any moderately complex survey will involve several different levels of data, and thus will require several different files, with different variable/case structures, to manage all the data. It's very common for researchers to fail to recognize this, as in the opening example, and to try to treat all the data as if it were at the same level.

So, the question is -- how can you recognize the different units of observation in a survey in order to organize the data properly? Throughout this discussion, I'll be using the terms unit of observation and level without defining them precisely, but their meanings should become clear as we progress.

Examples of Multi-Level Data

Every survey will have one base unit of observation which is of central importance to the project and on which the other levels of data are based. For surveys done by agricultural economists, one of the most common units of observation is a farm household, so we will use this in the remainder of the discussion as our base unit of observation.

If the household is our basic level of data, it seems obvious that we would be collecting a certain amount of information at the household level. This might include such variables as income, religion, length of residence in the current location, outstanding debt, and others.

Given a survey that's centered around the household, what might our other levels of data be? Well, to start with, we would probably want to know something about the different members of the household. This could include many different characteristics, such as age, gender, relationship to the head of household, years of education, and more. This information would comprise the household-member level of data, and, like the other levels we are about to see, it would have to be kept in a separate file from the household level data.

Another level of data in our hypothetical survey might be the household-crop level. For each crop the household raises, we might want to know the number of acres planted, amount of fertilizer used, production, sales, consumption, etc.

And finally, to add a more complex level of data, suppose there was a multiple-visit component to it, where we collect labor information each week, showing how many hours each household member worked on each crop that week. What would the level of observation for this data be? You can pretty much just summarize it from the preceding sentence, although it's still quite a mouthful -- the household-member-crop-week level.

In each of the preceding three examples where we talked about the level of the data, what we did (although we didn't express it that way) was to decide what it is about each case that makes it unique from all other cases in the survey. In looking at your own data, you will find that this is the easiest and most reliable method of identifying the different levels involved. Let's look at how this works for the three examples from above.

In the household-member file, we know that each member of each household will never be represented by more than a single case. If one member were included twice, we would know we had made a mistake. This uniqueness tells us that the unit of observation for this data is the household member.

Similarly, in the case of the household crop file, we can see that the uniqueness criterion is a combination of the household and the crop. Each household may raise several crops, resulting in one or more observations in our data file for each household. This tells us that the data is not at the household level. If, however, there were more than one case for the same crop for a particular household, it would be a mistake. Therefore the data in this file is at the household-crop level.

As the data organization becomes more complex it is easy to overlook a factor that contributes to the uniqueness of the cases. The labor file described above offers an example of this. The first two factors -- the household in question and the person who performed the labor -- are fairly obvious. But, of course, a person who worked on more than a single crop will be represented by more than one case, so crop is another identifying variable.

It would be easy to stop at this point in describing the file, but that would ignore the final uniqueness factor. Since the survey was done over a period of time, there can be many cases for the same household, the same person, and the same crop, but representing different weeks. Thus, because these particular four pieces of information are required to identify a unique case, we see that the unit of observation for this data is the household-member-crop-week.

This last example shows that multiple-visit data is not treated substantially differently than single-visit data. With either type, the basic approach to data organization is the same, starting with the identification of the factors that distinguish the cases from each other. A multiple-visit survey simply adds one more such factor that accounts for the time or date aspect of the data.

Key Variables and Levels of Data

You may have noticed by now the link between key variables and units of observation, which is simply that both of them deal with the uniqueness of cases in a data file. If you can correctly determine which are the key variables in a data file, you will also know its unit of observation, or level. The converse is true as well, of course; determining the unit of observation will determine what is needed for key variables.

You will find another connection between units of observation and key variables when you use SPSS to analyze survey data. Two commands you will use repeatedly when working with different levels of data

are AGGREGATE and JOIN MATCH. It's impossible to use these commands properly without knowing what your key variables are. JOIN MATCH requires you to use the BY sub-command to specify the key variables from the files you are joining. AGGREGATE requires you to use the BREAK sub-command to specify "break variables", which you will discover to be a subset of the key variables in a particular file.

The Hierarchy of Levels

The use of the word level in describing the organization of data implies that there is a sort of ranking or hierarchy involved. This can be diagramed in a way that may help to understand the relationships involved.

The levels of data are traditionally thought of as being arranged from the most aggregated data at the top to the least aggregated data at the bottom. So, in a typical survey of a single village we might have a hierarchy like the following (You'll recognize some of these levels from the example given previously, but we've added a few as well.):

Highest rank		Household	
Middle rank	Household-Crop	Household-Member	Household-Equipment
Lowest rank	Household-Crop-Parcel	Household-Crop-Transaction	

In one sense, the items that are grouped together on the middle rank have little in common. A file representing the Household-Crop level can't actually be combined with or used together with a file at the Household-Member level. Why then is it appropriate to think of them as being in the same rank in the hierarchy? What do they have in common?

Here, once again, we see the connection between key variables and levels of data. The different levels that are ranked together in the diagram have the same number of key variables, and thus may be thought of as having a similar level of aggregation. We could have labelled the diagram, perhaps more appropriately, as:

One key		Household	
Two keys	Household-Crop	Household-Member	Household-Equipment
Three keys	Household-Crop-Parcel	Household-Crop-Transaction	

So, it's important to remember that the hierarchy represented here is strictly a convenient pictorial way of showing the relationships among data files. Files that are shown on the same rank on the chart do not necessarily have identical units of observation. They are, however, similar in their level of aggregation, in that they have the same number of key variables.

Terminology of Levels

You will often find that people who have worked with survey data for a while develop an abbreviated terminology for referring to levels of data. In a technical sense, this terminology is sloppy and inaccurate, but since it is so common you should be familiar with it. Then, at least, if you start using it yourself you will be aware of what you are doing.

These abbreviations usually come into play with surveys that have a primary level of data, say the household level, which has several other levels of data below it. It's quite common, for example, to have a set of data containing information on each of the crops grown by each household, which would be correctly described as being at the household-crop level. Because it gets tiresome to use the phrase household-crop level over and over, most people soon start referring to this as the crop-level file.

As you can see, this is technically incorrect, because if the file were at the crop level it would have no more than one case for each distinct crop. If, for example, fifteen major crops were grown in that region, it would have no more than fifteen cases. Since the file actually has data on each crop grown by each household, it might have hundreds or thousands of cases.

It's quite possible to have data files from the same survey that represent both the crop level and the household-crop level, which makes it especially important to use the right terminology to avoid confusion. The SPSS AGGREGATE command can be used to take data from the household-crop-level file and create a new crop-level file , by summing or otherwise combining the cases for all households for each crop. This new file will have one case per crop, and so will truly be at the crop level. You can see that this file is radically different from the one that you may earlier have been referring to (incorrectly) as the crop-level file.

So the shorthand terminology is fine as long as you know what you are really describing. When there is any chance for ambiguity, however, you should be careful to use the correct terms.

Checking Your Data Organization

Once you have devised a tentative file organization for your data, there are some tests that you can apply to it to see if the organization is correct.

1. Each particular type of variable should appear only once in a given file.

The first sample survey used above, with the Fields Planted questions, demonstrates this potential error. As you recall, the following was shown as an incorrect data organization:

Village code VILL

Household code HH

1. Fields Planted in 1987

Field #	Distance from House	Size	Area Planted	1987 Plowed?	Planting Method	. . .	How Acquired	Length of Ownership
1	H1	H2	H3	H4	H5	. . .	H10	H11
2	H12	H13	H14	H15	H16	. . .	H21	H22
3	H23	H24	H25	H26	H27	. . .	H32	H33
4	H34	H35	H36	H37	H38	. . .	H43	H44
5	H45	H46	H47	H48	H49	. . .	H54	H55

2. Primary source of income H56

.

.

.

13. How long have you lived here? H67

In this file organization there are many instances of variables of the same type. For example, variables H1, H12, H23, H34, and H45 all contain the Distance from House data. They might be described, respectively, as Distance of Field 1 from House, Distance of Field 2 from House, and so forth. This kind of variable structure is a sure sign that something is wrong with the data organization.

The solution, of course, is to remove the offending variables from the file they are in and put them in another file as single variables, just as we did in the corrected version of that first example, shown again here:

Village code VILL

Household code HH

1. Fields Planted in 1987

Field #	Distance from House	Size	Area Planted	1987 Plowed?	Planting Method	. . .	How Acquired	Length of Ownership
FIELD	F1	F2	F3	F4	F5	. . .	F10	F11
1	_____	_____	_____	_____	_____	. . .	_____	_____
2	_____	_____	_____	_____	_____	. . .	_____	_____
3	_____	_____	_____	_____	_____	. . .	_____	_____
4	_____	_____	_____	_____	_____	. . .	_____	_____
5	_____	_____	_____	_____	_____	. . .	_____	_____

This new file will be at a lower level of analysis than the original file. The original file will remain at the same level it was, but will now contain only the household variables.

Notice that three things happen when you correct this kind of error. The new file at the lower level will:

a) have fewer variables,
b) have more cases, and
c) require an additional key variable.

The last point is worth some elaboration, because it was not discussed when we originally presented that first example. Notice that the set of variables defined for the first (incorrect) data organization doesn't include the field number as a variable. The field number column is included on the form, because it helps to make the table clearer, but the numbers there don't need to be recorded in the data file. The file doesn't really need a field-number variable because the variables themselves embody the information about field number, with H1 through H11 referring to field 1, H12 through H22 for field 2, and so forth. There is only one case per household, so the village and household IDs serve as the key variables, without any need for the FIELD variable.

However, when we change to the preferred lower-level organization, each field becomes a separate case, so the resulting file has multiple cases per household. This necessitates another key variable to distinguish the cases, thus we have to add the FIELD variable.

When we say that each particular type of variable should appear only once, the use of the word type is purposely vague, leaving a lot to your own determination. The similarity of type is not based solely on the words used in the question, but on the intent in asking the question and the way the data might be used.

For example, a survey that did not include detailed household-member information might still ask for the age of the head of household. There might also be a question asking the age of the house itself. In this case, these would both be appropriate variables for the household-level file. The presence of the word age in both questions is not, by itself, enough to classify these two variables as the same type.

2. All variables in a file must depend on the complete set of key variables for that file.

We've already discussed two possible ways of organizing our sample data, but there is at least one more way that might occur to some people. It is also wrong, but for a different reason than what we've discussed so far.

Many people have some recognition of the importance of different levels of data, but are still not comfortable with the idea of dividing their data into many separate files They sometimes try at all costs to fit their data into as few files as possible. This approach could lead to a file organization like the following:

Village code VILL

Household code HH

1. Fields Planted in 1987

Field # FIELD	Distance from House F1	Size F2	Area Planted F3	1987 Plowed? F4	Planting Method F5	. . .	How Acquired F10	Length of Ownership F11
1	_____	_____	_____	_____	_____	. . .	_____	_____
2	_____	_____	_____	_____	_____	. . .	_____	_____
3	_____	_____	_____	_____	_____	. . .	_____	_____
4	_____	_____	_____	_____	_____	. . .	_____	_____
5	_____	_____	_____	_____	_____	. . .	_____	_____

2. Primary source of income F12

.

.

.

13. How long have you lived here? F23

This is sort of a hybrid of the correct and incorrect organizations presented above. It could be thought of as basically the same as the fields-planted file from above, with one case per field, but with the addition of the household-level variables (questions 2 through 13).

Since there can be many fields (cases) per household, a decision must be made as to which case will contain the household data. It could be stored with each and every field (case) that belongs to a particular household, but this would require entering the data multiple times. The other choice is to store it as part of only one of the cases for each household, with the most likely candidate being the first field. This lack of a clear and obvious place to include the household data is a clue that something is wrong.

The key variables in this file are VILL, HH, and FIELD, because all three are required to identify a unique case. This creates an unusual situation, because the household-level variables, F12 through F23, have no relationship to the FIELD variable. For example, the question How long have you lived here? clearly is not an attribute of a particular field. Although we may find it convenient to assign them to the case containing field number 1, there is no real relationship between that data and that particular field number.

This shows that this file organization fails the second test, which says that all variables in a file must depend on the complete set of key variables. The solution to the problem is to move the offending variables to a file of their own, and the resulting file structure is identical to the correct one we presented earlier.

When correcting this kind of violation, one of the two new files to be created will be at a level above that of the original file, and will show two main changes. The file at the higher level will have:

a) fewer cases, and
b) fewer key variables.

Unlike our correction to the first wrong example, there is no overall change in the number of variables. This is because the nature of the error is different.

Defining the Meaning of a Case

If you look at these two tests more closely, you can see that they are both attempting to do the same thing -- to enforce a clear and consistent definition of a case within each data file.

In the example that failed the first test above, a case ostensibly represented a household. Closer examination revealed that each case for a given household could contain data for multiple fields being farmed. This can be thought of as allowing a case to contain multiple "sub-cases" that are inconsistent with the definition of the case itself. This is, as we have seen, an undesirable situation.

In the example that failed the second test above, a case was designed to represent a field. Here we saw that each case could also contain data that did not apply to that particular field, but rather to the household as a whole. This is also inconsistent with the definition of the case, and thus undesirable.

Understanding key variables and levels of data should give you a much clearer idea of the meaning of a case than you had before. When a file organization is correct, you should be able to clearly and unambiguously recognize just what a case represents in that file. When a file organization is wrong, you should be able to tell what is wrong with it, and how to move part of it to another file to correct the problem. Your ability to do this will always improve with experience, but you now have the basic principles required as a starting point.

NOTE: There are a number of other computer-related topics that need to be covered as extensions to this material. These include questionnaire design, variable naming, additional file-handling questions, and documentation issues.