Chapter 5

Data Entry and Verification

This chapter focuses on manual entry of data into Excel, while the next addresses studies in which data are taken from some other electronic source and imported into Excel. These are complementary topics, in that you will typically use one or the other in the course of creating a data set, but not both. However, there are issues that overlap between the two. For that reason, you should read both chapters so you have a basic understanding of all the strategies I discuss. Later, return to whichever chapter applies to your current research project. If the study involves preexisting electronic data, Chapter 6 will then refer you to this chapter on overlapping topics.

Beginning-level data entry in Excel is intuitive and obvious. The first row of the spreadsheet is used for the variable names. As in the SPSS Data View window, each column is dedicated to a variable, each row to a case. The first column of the sheet, column A, usually contains a case identification variable called something clever like "ID." Technically, except for a few basic Excel commands such as Save, that's all you need to know to create a data set in Excel.

OK, you say, so are we done yet? Not quite. *Optimal* use of Excel for data entry requires a good deal more expertise. There are idiosyncratic elements to Excel that require your consideration, there are navigation keys that will simplify your work, and there are several tools available for enhancing data entry and verification that clearly justify it as the data entry software of choice.

This chapter and the next cover a bunch of strategies and tools and may get a little overwhelming. Remember that in Chapter 8 I'll provide a guide to the tools I personally consider the most valuable for Excel data entry and verification. If you're going to manually enter data into Excel, I strongly recommend you read through this chapter it in its entirety before you start, to speed the process and to help avoid mistakes you'll regret later.

PREPARING FOR DATA ENTRY

The Problem with Excel

Remember that although the Excel sheet seems to demonstrate the same grid format as the SPSS Data View or the Access table, there is an important difference. Consistent with its heritage, Excel does not by default assume that values in a row or column are meaningfully related to each other. In Figure 5.1, cell E1 contains the value *S_Age* while subsequent values in that column are numbers. You as a researcher could probably figure out that S_Age is the name of a variable in this data set and the value of this variable for case 1 is 18, but Excel doesn't know that. The difference between Excel and other programs, such as Access or SPSS, can be demonstrated in the assignment of row numbers. SPSS assigns row number 1 to the data for the first case, but in Excel the row of variable names is part of the data and so is assigned row 1; the first case appears in row 2.

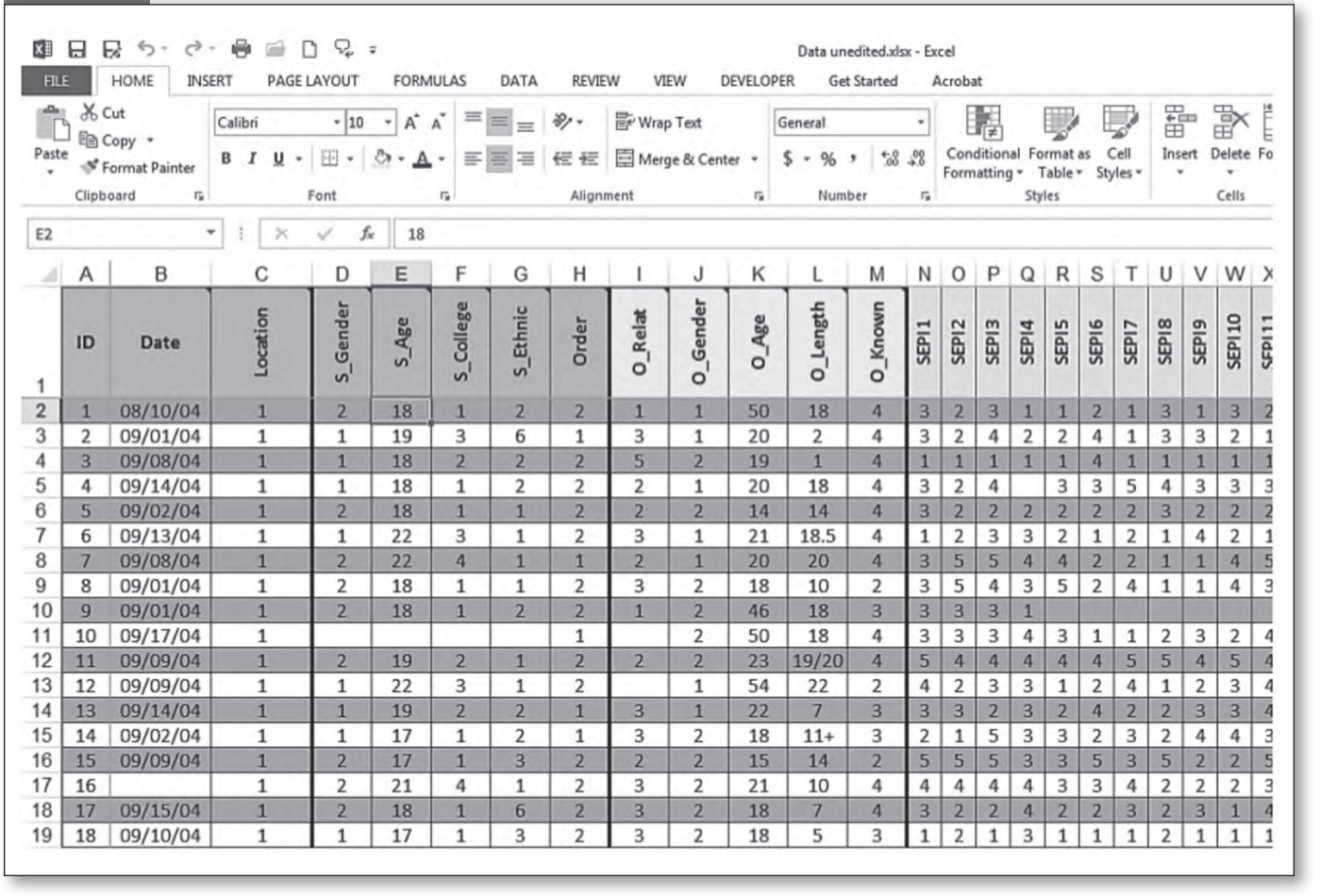
This one deficiency in Excel as a data entry system gives you the responsibility for making sure what you enter is compliant with data set structure. The data should begin in the first column and continue without empty columns until the last variable; the variable names should be in the first row; and each case should fill the following rows until the last case. I already mentioned in Chapter 3 that sorting in Excel can be a problem if there are empty columns, but empty columns and rows can also cause problems with your statistical software. Some programs have no problem with empty columns. SPSS, for example, will make up a variable name and fill the column with missing values. Other programs will choke on a column with no variable name, however. Empty rows are particularly dangerous because statistical software will generally assume they were intended, adding rows of missing values to your data file that throw off sample size counts. If you're not sure whether you have missing rows or columns, use Ctrl+Shift+8, which I mentioned earlier as a way to select an array of cells surrounded by an empty row and column. If rows or columns that contain data are omitted from the selection, you know there is an empty row or column in your data set.

Tip: No empty rows and no empty columns until all the data are entered!

Working with Multiple Sheets

There are several reasons why you may want to spread the entry of your data set across multiple sheets. It's possible that your data will not fit into a

Figure 5.1 A Sample Excel Data File



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	ID	Date	Location	S_Gender	S_Age	S_College	S_Ethnic	Order	O_Relat	O_Gender	O_Age	O_Length	O_Known	SEPI1	SEPI2	SEPI3	SEPI4	SEPI5	SEPI6	SEPI7	SEPI8	SEPI9	SEPI10	SEPI11
2	1	08/10/04	1	2	18	1	2	2	1	1	50	18	4	3	2	3	1	1	2	1	3	1	3	2
3	2	09/01/04	1	1	19	3	6	1	3	1	20	2	4	3	2	4	2	2	4	1	3	3	2	1
4	3	09/08/04	1	1	18	2	2	2	5	2	19	1	4	1	1	1	1	1	4	1	1	1	1	1
5	4	09/14/04	1	1	18	1	2	2	2	1	20	18	4	3	2	4		3	3	5	4	3	3	3
6	5	09/02/04	1	2	18	1	1	2	2	2	14	14	4	3	2	2	2	2	2	2	3	2	2	2
7	6	09/13/04	1	1	22	3	1	2	3	1	21	18.5	4	1	2	3	3	2	1	2	1	4	2	1
8	7	09/08/04	1	2	22	4	1	1	2	1	20	20	4	3	5	5	4	4	2	2	1	1	4	5
9	8	09/01/04	1	2	18	1	1	2	3	2	18	10	2	3	5	4	3	5	2	4	1	1	4	3
10	9	09/01/04	1	2	18	1	2	2	1	2	46	18	3	3	3	3	1							
11	10	09/17/04	1					1		2	50	18	4	3	3	3	4	3	1	1	2	3	2	4
12	11	09/09/04	1	2	19	2	1	2	2	2	23	19/20	4	5	4	4	4	4	4	5	5	4	5	4
13	12	09/09/04	1	1	22	3	1	2		1	54	22	2	4	2	3	3	1	2	4	1	2	3	4
14	13	09/14/04	1	1	19	2	2	1	3	1	22	7	3	3	3	2	3	2	4	2	2	3	3	4
15	14	09/02/04	1	1	17	1	2	1	3	2	18	11+	3	2	1	5	3	3	2	3	2	4	4	3
16	15	09/09/04	1	2	17	1	3	2	2	2	15	14	2	5	5	5	3	3	5	3	5	2	2	5
17	16		1	2	21	4	1	2	3	2	21	10	4	4	4	4	4	3	3	4	2	2	2	3
18	17	09/15/04	1	2	18	1	6	2	3	2	18	7	4	3	2	2	4	2	2	3	2	3	1	4
19	18	09/10/04	1	1	17	1	3	2	3	2	18	5	3	1	2	1	3	1	1	1	2	1	1	1

single sheet. The latest version of Excel allows up to 16,384 columns (variables) and 1,048,576 rows (cases, excluding one row for variable names), which would seem to be enough for any data set. However, many statistical programs have limitations on the number of rows or columns they can import from a single Excel sheet. For example, SAS can only import 255 variables at a time.

Another reason to consider dividing your data across several sheets is organization. For example, suppose your study involves multiple waves of data gathering, so that data are collected on children at age 2, 5, and 7. You may find it helpful to enter each wave into a separate sheet. If you're studying a medical intervention, you might have one sheet for pretest data, another for data at posttest, and a third for follow-up data.

A third instance that could warrant multiple sheets occurs when several people will be entering data into the computer independently. Cloud storage would allow them all to work on the same sheet, but another option would be to let each person enter his or her portion of the data set in a workbook on his or her own computer. How the data are combined later is then up to you. You can leave them in separate workbooks, use commands in your statistical software to import data from each file, and then combine them. You can combine them as separate sheets in a single workbook. You can do this by opening both workbooks, right-clicking on the tab for the sheet you would like to copy, then selecting the *Move or Copy…* option. This still requires the statistical software to access the sheets separately, but at least then all the data are located in the same file for future reference. Finally, if the total number of variables and cases does not exceed your statistical software's capacity for importing data from a single sheet, you can copy and paste the contents of one sheet into the empty region of another to combine them. I don't like this last approach myself, for fear I miscounted and the resulting sheet won't import completely, but it will work. Also, make sure not to introduce empty rows or columns this way!

You can split the data across sheets by variables, and this is the best option when using separate sheets for different waves of data collection. You can then give each sheet an informative name, such as "Age 2" and "Age 5," using the sheet tab. If you divide the data across sheets by variables, it is strongly recommended that you have some variable in common on each sheet, such as the participant ID number. This will be very helpful for matching up data later.

You can also divide data across sheets by cases, so that cases 1–100 may be entered into one sheet, 101–200 in another, and so forth. This approach usually only makes sense if the data are coming from different sources (e.g., the data are being collected at multiple locations or through multiple websites or are being entered into different worksheets by different people). Again, I recommend descriptive names for each sheet to keep them separate, even if you end

up using lame names such as "ID 1–10,000" and "ID 10,001–20,000." If you divide the data by cases, you should make sure the variable names are the same in each sheet and in the same order. This may seem obvious, but if the people doing your data entry are each manually entering the variable names, errors are possible. To make sure, I would recommend that you create the variable names in one sheet, then copy and paste them to the other sheets before distributing them to the people entering your data. The topic of combining these sheets into a single file will be addressed in Chapter 7.

Tip: There can be advantages to splitting your data set across multiple sheets. If the data are divided on the basis of variables, remember to include the participant ID variable in each sheet. If the data are divided on the basis of cases, remember to make sure variables are entered in the same order in each sheet. These are very important practices that can potentially save you a good deal of time and suffering later!

Naming Variables

One serious problem with using Excel for data entry is the potential for creating variable names that are illegal for your statistical software. Statistical programs, like databases, tend to have certain limitations on the names of variables. If you use one of those programs to create a data set and enter an illegal name for a variable, the software will warn you. Excel will not because Excel has no idea what the rules for variable names are in your statistical software. You can even enter the same variable name in two columns without a peep from Excel. When you then import the spreadsheet into the statistical software, the software will (depending on the program) reject the attempt, or it may modify the variable name to fit its conventions, sometimes without telling you or with a warning placed somewhere you may not know to check. If it just changes the name, now you have variables with different names than you expected them to have. That is not good.

Statistical software programs have become more flexible in their variable naming conventions in recent years, but there are still limits, and these programs differ in their limitations on variable names. Here are some guidelines. They are much too restrictive for some programs, and if you know your program is more flexible than what I recommend here, then by all means get more creative with your variable names, but these should work in almost all circumstances:

1. The first character of the variable name should be a letter.
2. No spaces allowed. This is pretty standard in the major programs.

3. Early statistical software could only handle variable names up to eight characters long, but more recent versions of the major statistical programs can safely accommodate variable names of at least 32 characters. If you want to be absolutely safe, limit your variable names to no more than eight characters, but up to 32 should be OK. If you're not sure what your statistical program can handle, try creating longer variable names in the software and see what happens.

4. Use characters other than letters and numbers with care. Suppose you want to separate words in a variable name; for example, you want a variable called "Date of Birth" but find DateofBirth annoyingly compressed (especially since some programs will convert variable names to all caps). You can use the underscore character or period to separate words in the variable name (e.g., Date_of_Birth or Date.of.Birth). The underscore is generally more popular among users of SAS and SPSS, the period among R fans. Do not use underscore or period as the last character of the variable name, as this can cause problems. There are other special characters, such as @, that can be part of a variable name, but rules applying to these characters vary a great deal across programs, and they're usually unnecessary, so I avoid them.

5. If you're working with R, you should be aware that variable names are case sensitive, so the same data set can include the variables *x* and *X*. This is not an issue in SPSS, SAS, or most structural equation modeling (SEM) software.

6. Make sure to avoid variable names that match commands or keywords used in your statistical software. This is sometimes tricky. For example, you're probably not surprised to learn that certain words, such as AND, OR, and TO, are restricted terms in many programs and shouldn't be used as variable names. Others may be less obvious, though, such as LT for "less than." Because this problem may occasionally be unavoidable until you become very proficient in your statistical software, my advice to you is the following: If you get an error that seems to imply an error in the wording of a command when you were referring to the name of a variable, search for what you think is the offending variable name in the help files for your software and see whether, in fact, it has a special meaning in that program.

Finally, I repeat: All your variable names must be unique. Again, Excel will not check for duplicate variable names, but the statistical software won't allow

them. In a later section I will describe a way to automate checking for duplicate variable names. Keep in mind that if your statistical software deals with variable names of excessive length by truncating them, you can end up with duplicates even if you have checked for them, and the software will potentially both shorten the name and change it to make it acceptable.

The risk of duplicate names is particularly high when instruments are administered more than once, especially if the different waves of data are entered into different sheets. If the McGrath Make-Believe Questionnaire is administered at pretest, posttest, and follow-up, you may reflexively name the variables representing the items in all three sheets MMBQ1–MMBQ30. When you try to combine the data, statistical software will usually overwrite the data from the earlier sheets, and it may do that without even telling you (you're supposed to know not to have duplicate names). To prevent that, in your pretest data sheet you might call them MMBQPre1–MMBQPre30, for the post-test sheet MBQPst1–MMBQPst30, and at follow-up MMBQFU1–MMBQFU30.

Tip: Make sure every variable name in your dataset is unique. This may seem obvious, but duplicates occur easily, especially if the study involves multiple administrations of the same instruments or collecting similar variables. Also, make sure variable names are compliant with the name limitations of the statistical software you plan to use.

Entering Individual Items

Research with human participants often involves the administration of multi-item questionnaires whose items are summed, averaged, or otherwise combined to generate a total score. I have known researchers to score such forms by hand and enter the total score into the data set. With all my heart I say unto you: Don't do it! Always enter multi-item forms into the data set item by item. It takes more time at the computer, but there are several reasons why this is the right thing to do:

1. It is commonly expected that a statistic called *reliability* will be reported for each of the multi-item scales in any research manuscript. In the past, it was common practice in the social sciences to report the standard reliability information provided in the manual for the questionnaire. That practice, referred to as *reliability induction,* has been rejected in recent years (Vacha-Haase, Kogan, & Thompson, 2000). These days, top-tier

journals expect you to report the questionnaire's reliability for your sample, and that requires entering the data one item at a time.

2. Changes in the scoring are more easily accomplished with individual item entry. In one study, one of the items on a questionnaire did not correlate well with the other items on the questionnaire. A review of prior research with the scale revealed this was a common finding for that item. As a result, we decided to score the questionnaire both with and without that item to see whether there were any differences in the results (there weren't). This would have been a much more tedious undertaking had we not entered the data one item at a time. Item-level data entry allows for alternative ways of scoring a questionnaire.

3. In the last chapter, I introduced the possibility of missing data. Sometimes when people complete questionnaires they will skip items. There's no reason not to use those people, as long as they have completed a sufficient number of items, especially if you average the items rather than sum them so they're on the same scale no matter how many items are missing (as long as they completed at least one).[1] The definition of a sufficient number of items, however, is up to you. You may want to experiment with the impact of allowing for different numbers of missing items. On a 20-item questionnaire, how many items can an individual leave blank before the questionnaire as a whole should be considered missing? How many cases are lost from the data analysis if you only use individuals who completed all 20 items, or at least 19, or at least three quarters, or more than half? How comfortable do you feel with the accuracy of the information? With individual item entry, you can experiment with different rules for how many items can be omitted before a case should be deleted.

4. Entry of the data one item at a time increases the potential for using that data in a later study with different goals.

5. There is a substantial literature demonstrating that people are lousy at scoring multi-item questionnaires by hand (e.g., Allard & Faust, 2000). It is tedious, demanding, and error prone. Leave that kind of work to your software. Once you figure out a formula or command that will score the questionnaire, the computer can do it for you with perfect

1. Averaging the completed items is mathematically equivalent to replacing missing items with the mean of the completed items. Another solution offered by some statistical software programs is *missing values imputation*, a procedure that replaces missing values with best estimates based on all the available data.

accuracy. Of course, it's always possible the method you created has errors, and you have to be pretty meticulous. When the scoring method you developed causes errors, though, I often find the results are so bizarre (e.g., a questionnaire that should produce scores of 0–20 produces scores in the thousands) that it's immediately obvious there's a problem. This doesn't mean you're immune from errors when you use formulas or commands to score your questionnaires, but it does mean the errors are often easier to catch than when the data are scored by hand.

6. The process of entering numbers into the computer also creates the potential for error. If the data are entered one item at a time, transcription errors are likely to have at most a small effect on the total score (especially if you use the data validation methods I will describe later). For example, if total scores on the McGrath Make-Believe Questionnaire can range between 0–90, a score of 15 that is incorrectly entered as 45 could easily be missed when checking the data. In contrast, if items range from 0 to 3, no error in entering a single item score could change the results by more than 3 points, unless the people entering the data are grossly incompetent. Even if they are, I will provide methods later that will ensure near error-free data entry.

The last two points are not intended to suggest that sloppiness in data entry is acceptable just because the effect will not be that large. I suggested much earlier that accuracy in data entry is important, and perhaps becoming increasingly important as social and behavioral sciences move toward parameter estimates. Later in this chapter and the next, I will talk about procedures for minimizing error. However, a certain amount of error may be inevitable, particularly in very large or complicated data sets. If errors occur, entering questionnaire data one item at a time at least limits the impact of any one error.

This logic also applies to something called *key reversal*. Key reversal occurs when one or more items on a questionnaire are opposite in meaning to the other items. For example, imagine a self-esteem questionnaire in which respondents rate their agreement with 20 statements on a scale from 1 (*strongly disagree*) to 5 (*strongly agree*). For an item such as "I am a person of worth," a higher score implies more self-esteem, while a higher score on an item such as "I do not think much of my abilities" implies less self-esteem. Before combining scores across items, the score on this second item must be reversed, so that a score of 5 becomes a 1, 4 becomes a 2, and so forth. If a person responds 4 to an

item that needs to be key-reversed, you could do it manually and enter it into the computer as a 2. Resist that temptation! If you think people are error-prone when adding values, that's nothing compared to how bad they are at key reversal. Keeping track of which items need to be reversed will slow down your data entry and increase the likelihood of error. This is the kind of mathematical manipulation computers are better at than we are; enter the score as a 4 and let the computer do the flipping for you later.

Tip: If there is a consistent rule that will be used to change your data, never manipulate data before entering it into your data set. Do not reverse items; do not score questionnaires. The more you enter your data exactly the way it appears on the forms on which the data were collected, the less opportunity there is for error.

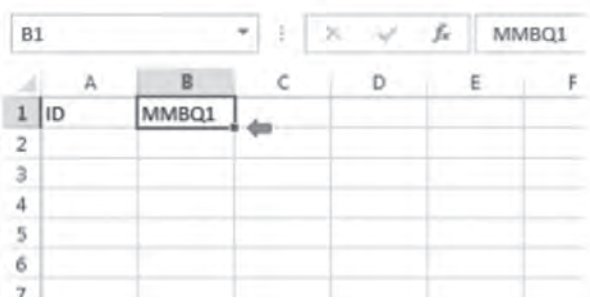
I hope I've convinced you that when your study includes multi-item questionnaires, you want to enter each item as a separate variable. Excel has a very clever tool for creating the variable names in this case. Assuming your ID variable is in column A, MMBQ1 would go into cell B1 (see Figure 5.2a). If you click on B1, it is highlighted by a thicker border. In the last chapter I mentioned the fill handle, which is the little black square in the lower right corner of the border. If you drag the fill handle to the right, you will see Excel start to suggest content for subsequent cells in the sequence MMBQ2, MMBQ3, and so on. Continue dragging until you have gotten to the cell for MMBQ30, release the mouse button, and the 30 variable names are entered into the cells.

What Excel does when you drag the fill handle is use the content of the cell that was selected to make a best guess about a sequence of cell contents. The sequence generation tool is pretty clever. For example, in Figure 5.2b, two cells are selected. If you then drag the fill handle to the right, Excel understands you are trying to use both cells to convey information about the sequence and so fills subsequent cells with MA2, MB2, MA3, MB3, and so forth. If it can't figure out a sequence, Excel will just copy the contents of the original cell to all the cells.

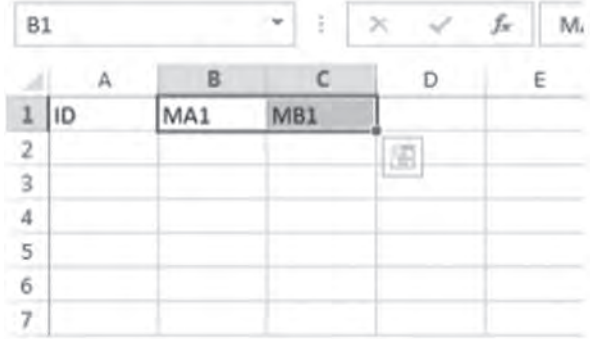
In some studies, ID numbers for different cases reflect some sequence, typically just increasing the number by one for each new case. If that is true for your study, you can use the same tool to fill in the ID numbers in the sheet. In cell A1, enter the variable name *ID,* or whatever you would like to call your ID variable. In cell A2, enter the number 1; in cell A3, enter the number 2. Now select cells A2 and A3 and drag the fill handle until you have entered a unique ID number for each participant.

Figure 5.2a and b Creating Variable Sequences

a



b



Note: The arrow in (a) points to the fill handle.

Tip: If there is information to be inserted that follows some sort of sequence, such as ID numbers or variable names that are distinguished by a sequence of numbers, the fill handle is a great tool. This is one of many advantages Excel has over data entry in Access or SPSS, where variable names or variable values must be entered individually.

The Data Dictionary

Now that you have entered all the variable names into your spreadsheet, you're ready to start to create your data dictionary. The data dictionary is an extremely important part of your data set. At its most thorough, a data dictionary includes a description of each variable in the data set; the type of variable

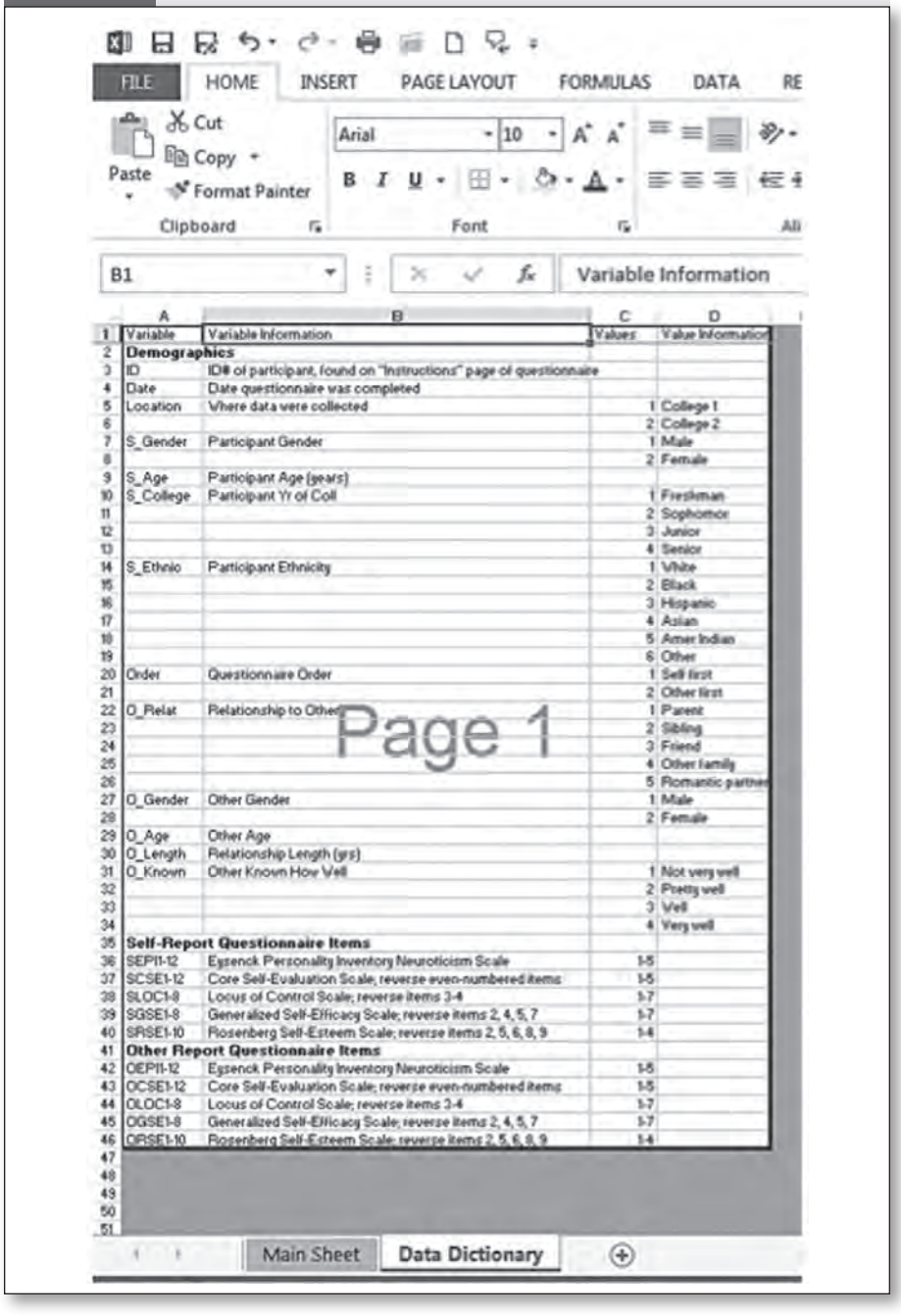
it is (text, numeric, etc.); and, when the meaning of each variable value is not obvious, information about those values. This last information tends to be the most important. For many variables, this information may be obvious (e.g., a variable called *Gender* with values M and F is pretty straightforward, so you may not feel the need to write down this information). In contrast, for a variable such as *Marital_Status* with values 1–5, it's very important to keep a record somewhere indicating whether 1 means married and 2 means single or vice versa.

In Excel, the data dictionary can be created in one of two ways. One is to create a separate sheet that contains the data dictionary. Figure 5.3 provides the Data Dictionary sheet to the data set in Figure 5.1. Elements of a thorough data dictionary include a column for variable names, a column for variable information, a column of possible values if each value has a specific meaning, and a column indicating what each value means. You could also have a column for variable type (text, numeric) or just include the type in the information column when the type is not obvious. Including the data dictionary in a separate sheet of the data entry workbook means the people doing your data entry and verification can review it at any time and can print it out for reference. This will also come in very handy during the data analysis.

Another approach to storing information about variables, one that can be complementary to the data dictionary, is to use comments. In Figure 5.4, two comments are visible. Notice that some cells in the first row have a little triangle in the upper right corner of the cell. These triangles let you know there is a comment associated with that cell. For example, there is also a comment attached to cell F1 (S_College) in Figure 5.4, though it is not currently visible.

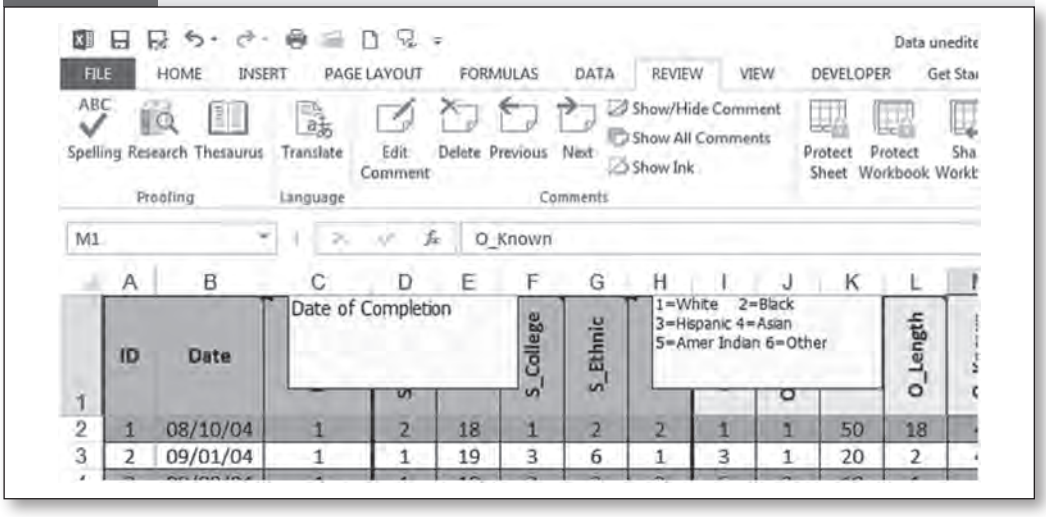
Comments enable you to reference information about the variables directly on the data entry sheet, especially if you have set the sheet so row 1 with the variable names remains visible at all times (see "Freeze Panes" below). Of course, that can be a nuisance if it gets in the way of where you want to enter or edit data. The easiest way to create, edit, delete, or show/hide the comment is to use the context-sensitive menu that appears when one cell is right-clicked (in Windows, or Ctrl+clicked on a Mac). Once a comment is created, you can also move it around the screen by dragging it to a new position (move the cursor around the comment and the cursor will change to a four-headed arrow; then the comment can be dragged), or resize it by dragging the border of the currently selected comment (the cursor will change to a two-headed arrow when the comment can be resized). As shown in Figure 5.4, the Review ribbon also has options for editing, moving between, and showing or hiding some or all comments. A word of warning: Showing all the comments in a data set when comments have been created for many of the variable names can result in a

Figure 5.3 A Data Dictionary Sheet



	A	B	C	D
1	Variable	Variable Information	Values	Value Information
2	**Demographics**			
3	ID	ID# of participant, found on "Instructions" page of questionnaire		
4	Date	Date questionnaire was completed		
5	Location	Where data were collected	1	College 1
6			2	College 2
7	S_Gender	Participant Gender	1	Male
8			2	Female
9	S_Age	Participant Age (years)		
10	S_College	Participant Yr of Coll	1	Freshman
11			2	Sophomor
12			3	Junior
13			4	Senior
14	S_Ethnic	Participant Ethnicity	1	White
15			2	Black
16			3	Hispanic
17			4	Asian
18			5	Amer Indian
19			6	Other
20	Order	Questionnaire Order	1	Self first
21			2	Other first
22	O_Relat	Relationship to Other	1	Parent
23			2	Sibling
24			3	Friend
25			4	Other family
26			5	Romantic partner
27	O_Gender	Other Gender	1	Male
28			2	Female
29	O_Age	Other Age		
30	O_Length	Relationship Length (yrs)		
31	O_Known	Other Known How Well	1	Not very well
32			2	Pretty well
33			3	Well
34			4	Very well
35	**Self-Report Questionnaire Items**			
36	SEPI1-12	Eysenck Personality Inventory Neuroticism Scale	1-5	
37	SCSE1-12	Core Self-Evaluation Scale; reverse even-numbered items	1-5	
38	SLOC1-8	Locus of Control Scale; reverse items 3-4	1-7	
39	SGSE1-8	Generalized Self-Efficacy Scale; reverse items 2, 4, 5, 7	1-7	
40	SRSE1-10	Rosenberg Self-Esteem Scale; reverse items 2, 5, 6, 8, 9	1-4	
41	**Other Report Questionnaire Items**			
42	OEPI1-12	Eysenck Personality Inventory Neuroticism Scale	1-5	
43	OCSE1-12	Core Self-Evaluation Scale; reverse even-numbered items	1-5	
44	OLOC1-8	Locus of Control Scale; reverse items 3-4	1-7	
45	OGSE1-8	Generalized Self-Efficacy Scale; reverse items 2, 4, 5, 7	1-7	
46	ORSE1-10	Rosenberg Self-Esteem Scale; reverse items 2, 5, 6, 8, 9	1-4	
47				
48				
49				
50				
51				

Figure 5.4 Excel Comments



mess. *Show All Comments* is really best when there are only a couple of comments. If some comments are visible and you want to make them all disappear, the easiest way to do so is by hitting Show All Comments (to make them all appear), then hit it again (to make them all disappear).

You can also print the comments. Click on the Page Layout tab, then on the arrow in the lower right corner of the Page Setup section of the ribbon to open the Page Setup dialog box. Click on the Sheet tab, then on the dropdown box for Comments. Choose *At End of Sheet*. Now if you print the sheet, the comments will be printed at the end. Because a worksheet can sometimes take hundreds of pages to print, I don't recommend actually printing all the data. Just limit your printing to the pages at the end that contain the comments.

In terms of a data dictionary, I recommend the following: First, create the data dictionary as a separate sheet in the workbook. If you want a printed reference of the data dictionary, print out that sheet. However, if you find you frequently have to reference the data dictionary for certain variables (perhaps because there are many options), you may find a comment helpful.

One final point I will make about the data dictionary is this: You should build the dictionary before you ever start entering data. However, don't be hesitant to add to it as your data set is built and you have new ideas about how to make the data dictionary more accurate or useful. Make sure you distribute these changes to all the people doing data entry.

Tip: Any information about your variables that could be lost should be incorporated into a data dictionary. Creating a separate sheet for the dictionary is strongly recommended, though comments can also be helpful at times. Be thorough; any information that would be dangerous to lose about your variables should be documented. Once it's done, print out a copy as a reference.

Decisions about Missing Data

Another issue to consider before beginning data entry is how you will handle missing data. Missing data are an inevitable issue in research, and before you begin entering data you should have a plan for how you will represent missing data. For most statistical programs, leaving a cell blank is an acceptable means of identifying a missing value. This can occasionally cause problems with some programs that require converting the data set to a text file first, however. Many researchers prefer to enter something into every cell, even if some things indicate nothing.

SAS and SPSS, by default, assume a cell containing a period indicates a missing value for a numeric variable. R instead uses NA (not available). All three recognize empty cells as missing values, and they also will let you define other indicators of missing values. For example, if the acceptable responses to an item are in the range 1–5, you could use a 9 to indicate the item was omitted. You can even use different values to represent data that are missing for different reasons. For example, you could use 8 when data are missing because the item was not administered to some participants and 9 to indicate that the item was administered, but the participant didn't respond. You should become familiar with the missing value facilities of the statistical software you will be using if you want to get all fancy like that. In particular, programs can be quite different in terms of the options they permit for missing text values. This knowledge will also help you decide whether you are safe leaving the cell blank when data are missing or whether you should use some placeholder that indicates missing values.

Tip: Before you begin entering data, decide how you will treat missing values. This will require some understanding of the best options for your statistical software.

Using Cell Formats

Excel offers a number of tools for formatting the look of cells so it's easy to distinguish between them. Figure 5.1 illustrates some of the options: Sets of item variable names from different questionnaires are shaded using different levels of grey and thicker vertical bars between sets. Variable names have been turned 90 degrees to fit more on a single screen. There are many other options: You can fill the cells using different colors rather than shades of grey; change the color of the text; change borders, fonts, positioning of text within the cell, and so on. All of these tools are to be found on the Home ribbon, particularly under *Format, Format Cells*. Clicking the dropdown arrow to the *Number* subgroup of the Home ribbon will bring you to the same place. You can also select a column, row, or group of cells, then click on Format Cells from the context-sensitive right-click/Ctrl+click menu.

Notice that in Figure 5.1, even-numbered rows are shaded while odd-numbered rows are not. This formatting can help you read across a row, and I find it pretty useful when working with data. You could achieve this effect by manually highlighting each second row and shading, but that would be a tedious process. Also, if you reorder the rows for some reason you would lose the effect.

There are two better options. The first is conditional formatting: formatting that varies depending on conditions you define. This option is more complicated than the second I'll discuss below, but it gives you more control over the outcome. In this case, here are the steps I followed:

1. I clicked the *Select All* button.
2. I clicked on *Home, Conditional Formatting, New Rule*.
3. In the *New Formatting Rule* box, I clicked on *Use a Formula to Determine Which Cells to Format*.
4. In the box titled *Format Values Where this Formula is True,* I entered the formula

   =ROW()/2=INT(ROW()/2)

5. I clicked on the Format button, then the Fill tab, and selected the shade of grey I wanted to fill cells in the even-numbered rows.
6. I clicked on OK until I was back at the spreadsheet. Poof, even-numbered rows are grey while odd-numbered rows remain white.

Why did this work? The conditional formatting instructed Excel to shade any cells where the function was true. The ROW function returns the row number of the current cell (another function that, like RAND, requires the parentheses even though there is no argument), and INT returns the integer. For cells in row 3, dividing the row number by 2 returns a value of 1.5. INT then truncates this value to 1.0, so the formula is false. For cells in row 4, the result of INT(ROW()/2) and ROW()/2 are both 2.0, so the formula is true (and is true for every cell in an even-numbered row) and the shading is applied. What's particularly neat is that reordering the rows or inserting a row has no effect because the row numbers adjust and so does the shading. Conditional formatting can take a little practice to master. As you can see from the New Formatting Rule box, there are all sorts of conditions you can set to control cell formatting. Once you become proficient, though, it's a nice way to set different formats.

The second option is to define the data set as an Excel table (*Select All,* then *Home, Format as Table* or *Insert, Table*). A variety of display options are available, some of which include differential shading of alternate rows. You can then format cells further to add borders, shade sets of cells differently, or whatever. Excel tables is a good way to speed up cell formatting.

Tip: I find shading alternate rows, either through conditional formatting or defining the data as a table, really helps when manually entering data. Otherwise, I'm not a big fan of spending the time formatting the look of cells as a technique for improving data entry, but I've had some students who disagreed with me about this. You can decide for yourself.

Data Validation

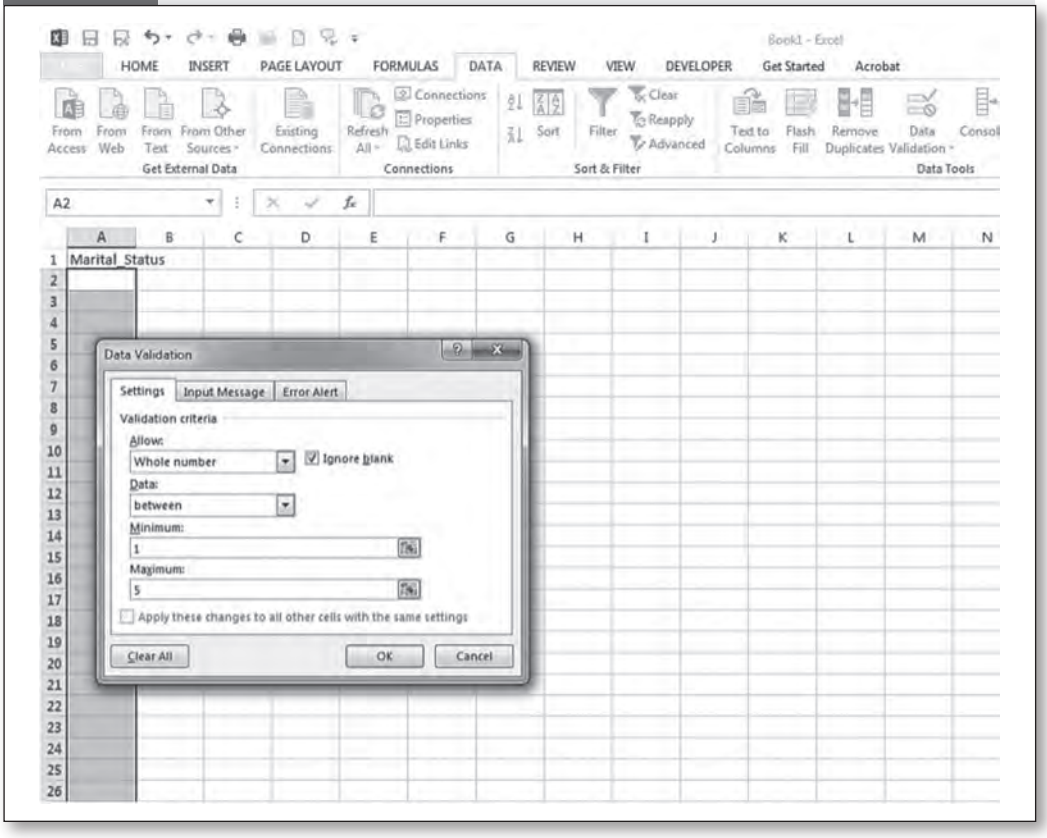
When variables are created in SAS or SPSS, you must define the variable type as numeric, date-time, and so on. This is an important feature of data entry in those programs because it keeps you from accidentally entering numeric data for a text variable and vice versa. I've already discussed the problem that Excel doesn't assume entries represent data of a specific type, and so by default allows entry of any type of information into any cell. Data Validation was developed to address this problem, and it turns out to be far more powerful as a method of restricting input than what SAS or SPSS offers.

Suppose the data dictionary indicates data for the variable Marital_Status should be restricted to the integers 1–5 to indicate Single, Married, Divorced,

Separated, or Widowed, respectively. Data Validation can be applied to all the data cells in the column to make sure no values are entered except the permissible ones.

Figure 5.5 shows you the Data Validation dialog box. To reach this point, I entered the variable name in the first row. I then selected the entire column and clicked on *Data, Data Validation*. The Data Validation box has a number of options, most of which I've found useful at one time or another. On the Settings tab, the Allow dropdown box offers a series of options. The default is Any Value, meaning Excel places no restrictions on the content of the cell. A second option, illustrated in Figure 5.5, is Whole Number. Choosing this option allows you to restrict data in a variety of ways. In the figure, entries are limited to whole numbers between 1 and 5. Other options for the Data box include Not Equal To and Greater Than. Several other options for the Allow box—Decimal, Date, Time, and Text Length—use the same options for the Data box.

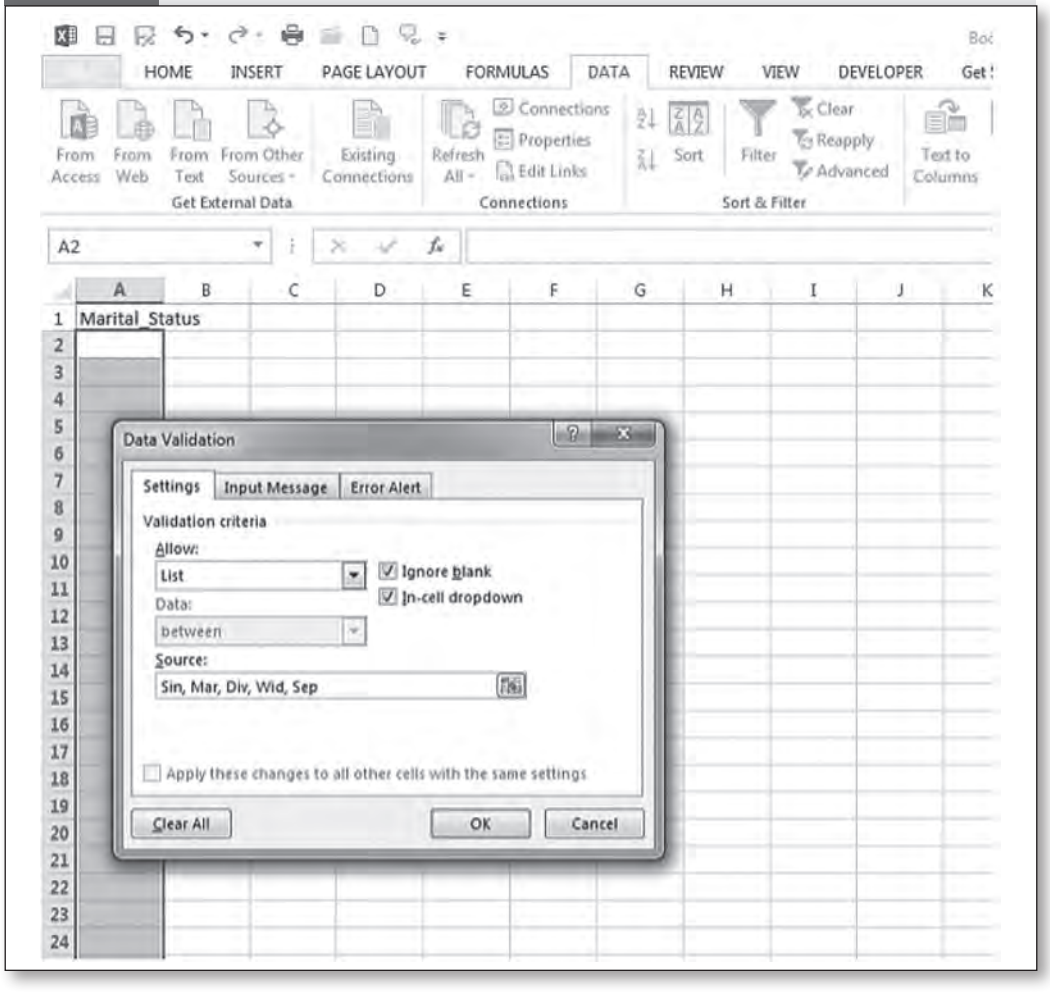
Figure 5.5 Data Validation



If you click on OK in Figure 5.5, no cell in column A can accept any entry except the integers 1–5. "But wait," you say, "the variable name in cell A1 isn't an integer from 1–5." That's not a problem; Data Validation only affects subsequent data entry. For that reason, you should enter the variable names before you implement data validation. In the next chapter, I'll cover *Validation Circles,* which looks for existing instances of incorrect data entry.

I want to pay special attention to the List option in the Allow box. This is a particularly powerful method for controlling input. If you choose List, a text box called *Source* becomes available (see Figure 5.6). You can list the acceptable

Figure 5.6 Data Validation Using a List



options for the cell in this box, separated by commas. For example, rather than using 1–5 for Marital Status, you might elect to use the options "Sin, Mar, Div, Wid, Sep" instead. The Source box in Figure 5.6 restricts the choices to those five options. List is also helpful if you want to use the 1–5 variable, but you also want to use a nonadjoining integer such as 9 to represent missing values: In that case, you would enter 1, 2, 3, 4, 5, 9 in the Source box. If you do decide on a value that represents missing data, remember to uncheck the "Ignore Blank" box so no cell can be left empty.

The List option also reveals a new checkbox labeled *In-Cell Dropdown*. Leave it checked. Now when you click on a cell in column A, a dropdown arrow will be available that displays the acceptable options from which you can select. You can still type in the data as well.

Instead of typing the permitted entries into the Source box, you could put the five marital choices into five adjoining cells somewhere else in the workbook, maybe in the data dictionary sheet or in a sheet called *Tables* where you also have lookup tables, then enter the range of addresses for the cells in the Source box. You can type the cell addresses into the Source box, or click on the Selection button at the right end of the Source box, navigate to the cells containing the options, select them, and click on the Selection button to return to the dialog box. This is often a much better method than typing the options directly into the Source box, especially if there is a lengthy list of options.

The final option in the Allow box, Custom, is also very powerful, enabling you to use a variety of formulas to limit cell entry. Once you have become proficient at formulas, you can use Custom data validation for all sorts of purposes (e.g., to make sure you don't duplicate a number in a column or to make sure the total for a column does not exceed some value). Basically, if you have a data restriction you want to implement that cannot be accomplished through any of the preceding options, I'm willing to bet you can find some way of doing it with a Custom data validation. I have not found this level of power necessary in statistical data entry, so I won't discuss the Custom option further, but you can find many examples of its use online.

To make Data Validation even more helpful, notice the other two tabs in the Data Validation dialog boxes in Figures 5.5 and 5.6. The more useful is *Input Message,* which includes text boxes for Title and Input Message. You might enter "Marital Status" as the title and

1 = Single
2 = Married
3 = Divorced
4 = Separated
5 = Widowed

for the input message. Then when you select a cell in that column, a popup box appears with that title and message so that the person entering the data knows exactly what values are allowed and what they mean. The Error Alert offers you the opportunity to present a message when a forbidden value is entered. I don't find this option particularly useful, although having a message like "What's wrong with you?" pop up when you make a mistake is amusing for about a minute. If, after seeing an input message that gives the name of the variable, the acceptable values, and what each value means, the person doing the data entry still inserts an incorrect value, the situation is pretty much hopeless.

Data Validation is an incredibly useful tool for reducing the likelihood of entering incorrect information. I strongly recommend you set data validation features for every cell in your sheet before beginning data entry. Because the individual items from a questionnaire almost always are completed on the same scale, you can often apply a single data validation protocol for large regions of the sheet.

There is one minor down side to Data Validation worth noting: You may not always be able to anticipate all the acceptable values. For a 1–5 item, the boundaries of an acceptable response may seem clear, but then you may find someone who writes on the questionnaire "I'm right between 2 and 3. I'm a 2.5." You can arbitrarily choose to enter this result as a 2 or 3, or you can lose the data by treating it as missing, or you can enter the response as a 2.5. This last option is the most accurate, and generally it would have no negative effect on the overall score, so why not? The fact you didn't anticipate this value when you set up the Data Validation for that cell is not a serious problem, though, because you can simply turn off or change the Data Validation setting for the cell.

In summary, SAS, SPSS, and database software originally had an advantage over Excel as a data entry method in that they required setting the variable type for each variable, making it harder to enter incorrect information into the file. The Data Validation tool now available in Excel not only serves some of the same purpose; it actually gives you even greater control over what can be entered than those other programs. It's a good one.

Tip: I consider Data Validation one of the best Excel tools ever developed for manual data entry. Except for the Custom option, which you'll probably never need, it's very easy to use. Use it!

Though I'm not a big fan of taking the time to format cells so they look different, with the exception of shading every second row, I'm a big fan of formatting the content of the cells to make it easier to read and/or to improve the data entry. The tools I discuss in this section all fall in the Number tab of the Format Cells dialog box. You can reach that dialog box via *Home, Format, Format Cells;* via Home, then the dropdown arrow for the Number subgroup; or by right-clicking/Ctrl+clicking the selected cells.

Working with Text Variables

Data Validation doesn't do much to control what you enter for text variables. You can use the Source box to restrict text variable entries to a small set of options. There is another option in Allow called *Text Length* that can control the number of characters you can input, though this applies to numeric variables as well as text variables. You can build more sophisticated methods of controlling text with the Custom option, but that requires a good understanding of formulas.

There are two issues to be aware of when entering text variables into Excel. The first has to do with text variables represented by numbers, the classic example of which is zip codes. When you enter a zip code, Excel by default assumes what you have entered is a number. When you enter a zip code such as 01234, Excel by default assumes this is the number 1234 and drops the zero at the beginning. If you used Data Validation to set the Text Length to 5, Excel adds insult to injury by indicating you made an error because the entry is not five characters long.

There are two ways to address this problem. The first is to remember to precede each area code with a single quote before you enter it (e.g., '01234) so Excel treats it as text. This isn't my favorite solution, as you have to remember it each time you enter a zip code. Also, don't think you can add the quote only for those zip codes that begin with a zero. You just end up with a column in which some of the data are treated as text and some as numbers, a state of affairs that will wreak havoc when you try to transfer your data to statistical software or run statistical analyses. A much better and simpler solution involves selecting the column and choosing Text in the Number tab of the Format Cells dialog box. All entries into cells of that column will then be treated as text, no matter whether they consist of letters, numbers, or dates. As text, zip codes will retain beginning zeroes.

One way to tell whether digits are being treated as numbers or text is their position in the cell. By default, numbers are right-justified in Excel; text is left-justified. If zip codes or social security numbers are lining up on the right side of the cell, you know Excel is treating them as numbers. Once you format them as text, they will move to the left side of the cell.

The second issue to know about text variables is that capitalization matters when dealing with text variables in statistical software. To understand what this means, consider what happens when you search for a word such as "fat" in a word processor. You should find each appearance of the word, whether it appears as "fat," "Fat," or "FAT," because the search facility is case neutral by default. In contrast, text variables in statistical software are case sensitive. If you have one person entering data who uses the values "F" and "M" for Gender while another uses "f" and "m," a frequency table for that variable in SPSS or SAS will indicate you are dealing with four genders. It is for this reason that even categorical variables such as gender or marital status are often entered using numbers, even though it means you have to keep track of which number represents which gender or marital status. Make sure you build that data dictionary!

Tip: If you're working with a categorical variable with a small set of values, such as marital status, you're probably better off representing the values with whole numbers. If you need to use a text variable (e.g., for addresses or telephone numbers), make sure to format the cells as text before you start. This won't keep you from entering numeric values by mistake, but it will correctly format numbers that are really text.

Working with Date/Time Variables

As with text variables, when working with date or time variables I recommend you restrict the column to date or time values first in the Number tab of the Format Cells dialog box. You can choose the entire column and apply the date or time format even though the first cell in the column contains the variable name, and you can even enter the variable name after you've formatted the column. If the format for a cell isn't appropriate for the content of the cell, Excel just ignores the format.

You can choose from a variety of formats for displaying dates or times. Because they're stored as numeric information, Excel can switch from one format to another with no problem. You can enter years as two digits or four digits (2012 versus 12), and you can display years as two digits or four digits.

It's obviously faster to *enter* the year as two digits, and that's fine to do, but I recommend using a format for *displaying* dates that presents all four digits of the year rather than just the last two digits, especially if the date variable will potentially include values outside the 21st century. If you don't enter the century, Excel must guess which century you mean, and if the year is being displayed as two digits, you don't necessarily know whether Excel has made a mistake. Displaying all four digits will help prevent errors. The standard formats for times either include AM/PM or use a 24-hour clock (e.g., 18:25 for 6:25 PM) to prevent confusion.

Tip: Columns that will contain date or time variables should be formatted as such at Home, Format, Format Cells. Make sure to use a format that displays four-digit years or AM/PM information for times.

Working with Numeric Variables

Excel has various defaults for the presentation of numbers: Whole numbers are displayed without decimals and with leading zeroes removed, 0 is added before the decimal point for values less than 1, and so on. All of these options can be modified—big surprise here—with the Number tab in the Format Cells dialog box. If a variable allows both decimal and whole values, I will generally set the format for the column to present all numbers to two decimal places, so they will line up for easier comparison. You can still enter integers without decimals; Excel will reformat the data for you. Again, when setting the format to Number, don't bother omitting the first cell in the column containing the variable name; Excel will just ignore the format for that cell.

Tip: If numeric variables will include both integer and decimal values, you may want to format the column so all values are displayed to the same number of decimals.

VIEWING MULTIPLE SHEET REGIONS

Freeze Panes

Like Data Validation, Freeze Panes is a tool that was developed to address some of the inherent limitations of the free-form spreadsheet as a data entry

system, and it ends up being better than what's available from software that was developed specifically for data entry.

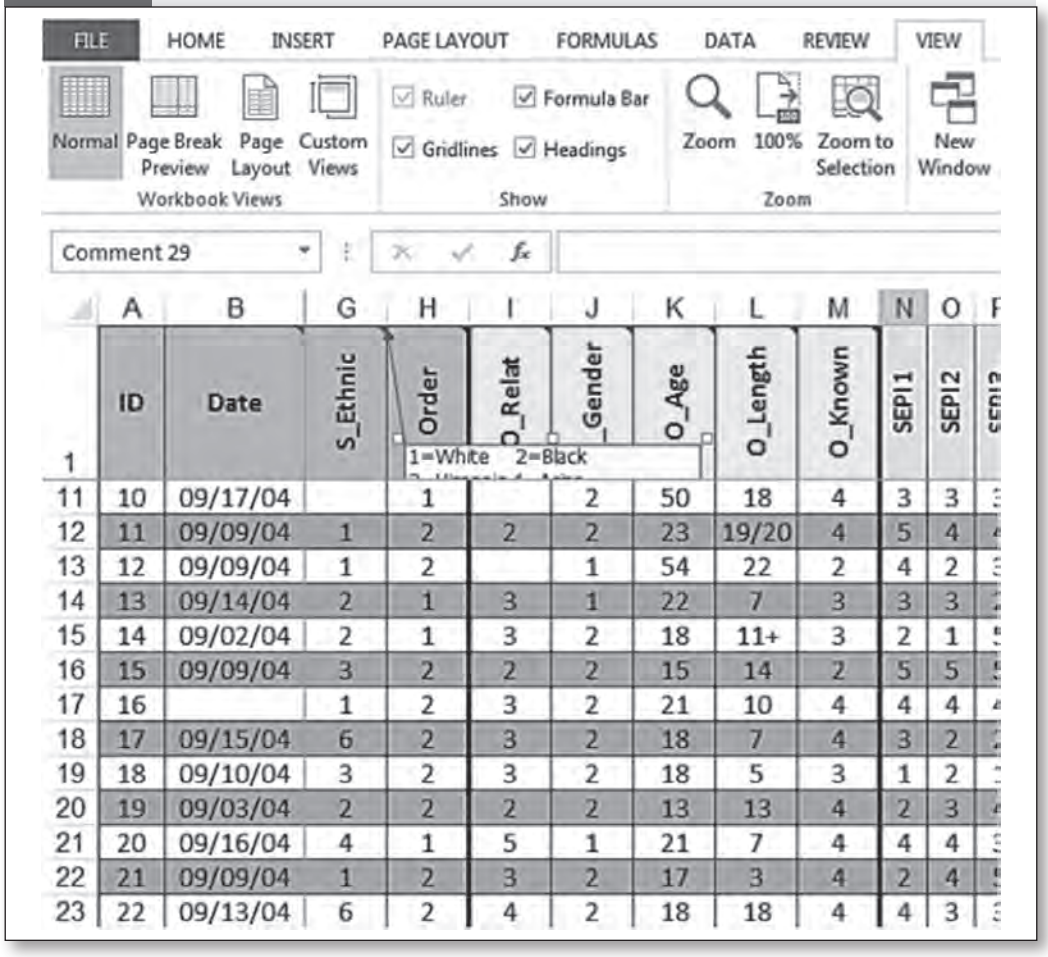
I've already noted one advantage of statistical programs over spreadsheets for creating data sets is that the former distinguish the variable name from the variable values in a column. One consequence of that distinction is evident when you scroll down through the data set window in SAS or SPSS, either to examine the data or to add new values. Those programs always keep the variable names at the top of each column and the case number at the beginning of each row for reference. This is a very valuable tool for keeping yourself oriented as you enter and move through your data. In contrast, because Excel does not consider the values in the first row anything special, by default the variable names scroll out of the window as you move through the data.

Freeze Panes is the solution for this problem, available by clicking on *View, Freeze Panes*. This brings up a menu of three options. Freeze Top Row will fix row 1, so the variable names always remain in the window. Freeze First Column does the same for column A, so the ID numbers always appear. The third option, Freeze Panes, is more flexible, allowing you to freeze all the columns to the left of the currently selected cell and all rows above it. This could be used to freeze both the first row and the first column, by selecting Freeze Panes when cell B2 is selected. Freeze Panes also enables you to freeze more than one column and/or row. In Figure 5.7, the cell C2 was selected when I clicked on Freeze Panes so that the first row and the first two columns are frozen in place even when columns and rows start scrolling out of the window. This can be helpful if, for some reason, you have two columns at the left end of the sheet or two rows at the top of the sheet that provide information you want to access at all times, for example, if you're identifying individuals by first and last name (though this is not recommended because of confidentiality issues, unless absolutely necessary).

Freezing the first row provides the same functionality as always having the variable names on screen in statistical software data grids. Freezing the first column is actually better than what you can do in statistical software because it enables you to keep the participant ID number on the screen rather than the case number. In fact, I consider Freeze Panes another one of the tools that make Excel the best choice for data entry and verification.

Clearly this is a very useful tool, but it does change Excel's behavior in two ways. First, in the note to Table 3.1 that discussed navigation tools, I defined the data region as "a rectangular matrix containing all the actively used cells of the spreadsheet except frozen rows and columns." If you hit the Home key with frozen columns, instead of moving to column A you will move to the first not-frozen column. The left arrow key will still let you move into the

Figure 5.7 Freeze Panes



	A	B	G	H	I	J	K	L	M	N	O
1	ID	Date	S_Ethnic	Order	O_Relat	_Gender	O_Age	O_Length	O_Known	SEPI1	SEPI2
11	10	09/17/04		1		2	50	18	4	3	3
12	11	09/09/04	1	2	2	2	23	19/20	4	5	4
13	12	09/09/04	1	2		1	54	22	2	4	2
14	13	09/14/04	2	1	3	1	22	7	3	3	3
15	14	09/02/04	2	1	3	2	18	11+	3	2	1
16	15	09/09/04	3	2	2	2	15	14	2	5	5
17	16		1	2	3	2	21	10	4	4	4
18	17	09/15/04	6	2	3	2	18	7	4	3	2
19	18	09/10/04	3	2	3	2	18	5	3	1	2
20	19	09/03/04	2	2	2	2	13	13	4	2	3
21	20	09/16/04	4	1	5	1	21	7	4	4	4
22	21	09/09/04	1	2	3	2	17	3	4	2	4
23	22	09/13/04	6	2	4	2	18	18	4	4	3

Note: In this case, the first row and the first two columns are frozen. Comments that are not completely contained in the frozen region will be cut off.

frozen columns, but those columns are excluded from the data region. This is no big deal.

Second, Freeze Panes and Comments do not play nicely together. If a comment is attached to a cell in the frozen region, but portions of the comment extend outside the frozen region, the portion inside the frozen region will stay on screen but the portion outside will scroll off the screen. The result is a comment that gets cut in half or that disappears completely, as demonstrated in

Figure 5.7. If you've made the comment visible, I suspect it contains information you want to have available to you, so this is usually not desirable behavior. To keep the comment completely visible with frozen panes, you must adjust the comment so it appears completely in the frozen rows or columns. There are two tools you can use to make it fit. You can reshape the comment by dragging the comment border, as noted earlier. You can also make the frozen row(s) taller or column(s) wider to accommodate larger comments. However, if you have commented extensively, it can be very difficult to fit all your comments into the frozen region. These are minor inconveniences and trivial in comparison to the value of being able to freeze panes. It argues for relying primarily on a data dictionary that's printed out and/or contained in a separate worksheet rather than comments.

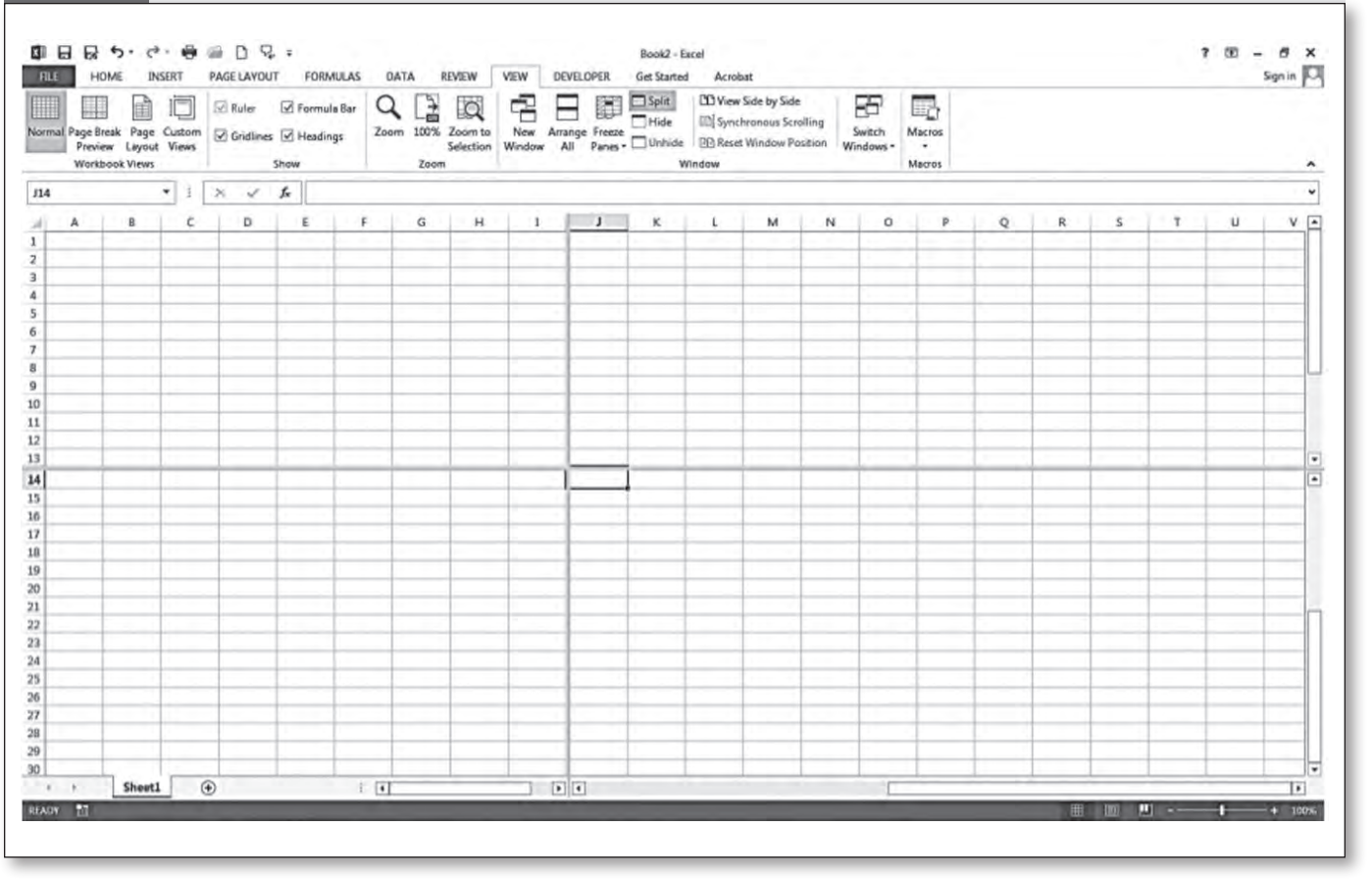
Once you have frozen the ID variable and the variable names on the screen, the column letter and row number headings are superfluous. In fact, the row numbers can be confusing because they differ from the case number by one digit. You may choose at this point to turn off the headings. Simply go to the View ribbon and unclick Headings in the Show subgroup.

Tip: Before you begin entering data, freeze (at least) the first row with your variable names and the first column where you will enter the ID variable. That way you are always oriented to the variable and case.

Split Screen

The remaining tools discussed in this section have to do with observing two portions of a data set at the same time. The first involves splitting the screen. In Figure 5.8, I selected cell J14, then clicked on *View, Split*. This split the sheet at the upper corner of the selected cell. Notice there are now scroll bars for each of the four sections of the screen. I can scroll each of the four sections on a different part of the sheet, though the two sections in a row will always show the same rows and the two sections in a column will always show the same columns. You can drag the horizontal or vertical border that splits the sheet to resize the sections, you can reduce the number of sections to two by double-clicking on one border, or you can eliminate the split screen by clicking on Split again or by double-clicking on the juncture of the vertical and horizontal borders.

Figure 5.8 Split Windows



Hide/Unhide

I will describe Hide in terms of rows, but what follows applies to columns as well. Suppose you would like to compare the data in rows 23 and 100. Select the array of rows 24:99 and choose Hide either from the context-sensitive menu or from the View ribbon. To reveal the intervening rows again, select rows 23 and 100, the rows surrounding the hidden row, and choose Unhide on the context-sensitive menu or the View ribbon. You can also unhide all hidden rows in the sheet by clicking on Select All, then on *Home, Format, Hide & Unhide, Unhide Rows*.

Unfortunately, Hide is a little annoying if you want to unhide the first row because you cannot select a row prior to the hidden row. In this situation, you have two options. The first is to use the procedure just described for unhiding all rows. If you really want to unhide only the first row, type 1:1 or an address in the first row into the Name bar. Now the selection is in row 1 even though the row is hidden, so Unhide Rows will only affect the first row. This method can also be used to unhide a single row or subset of rows in an array of hidden rows.

Be careful about using Hide. You can easily forget that columns or rows are hidden. This is especially dangerous when doing data entry, when you're mechanically entering values one after another without realizing you're putting the data in the wrong columns or have left hidden rows empty.

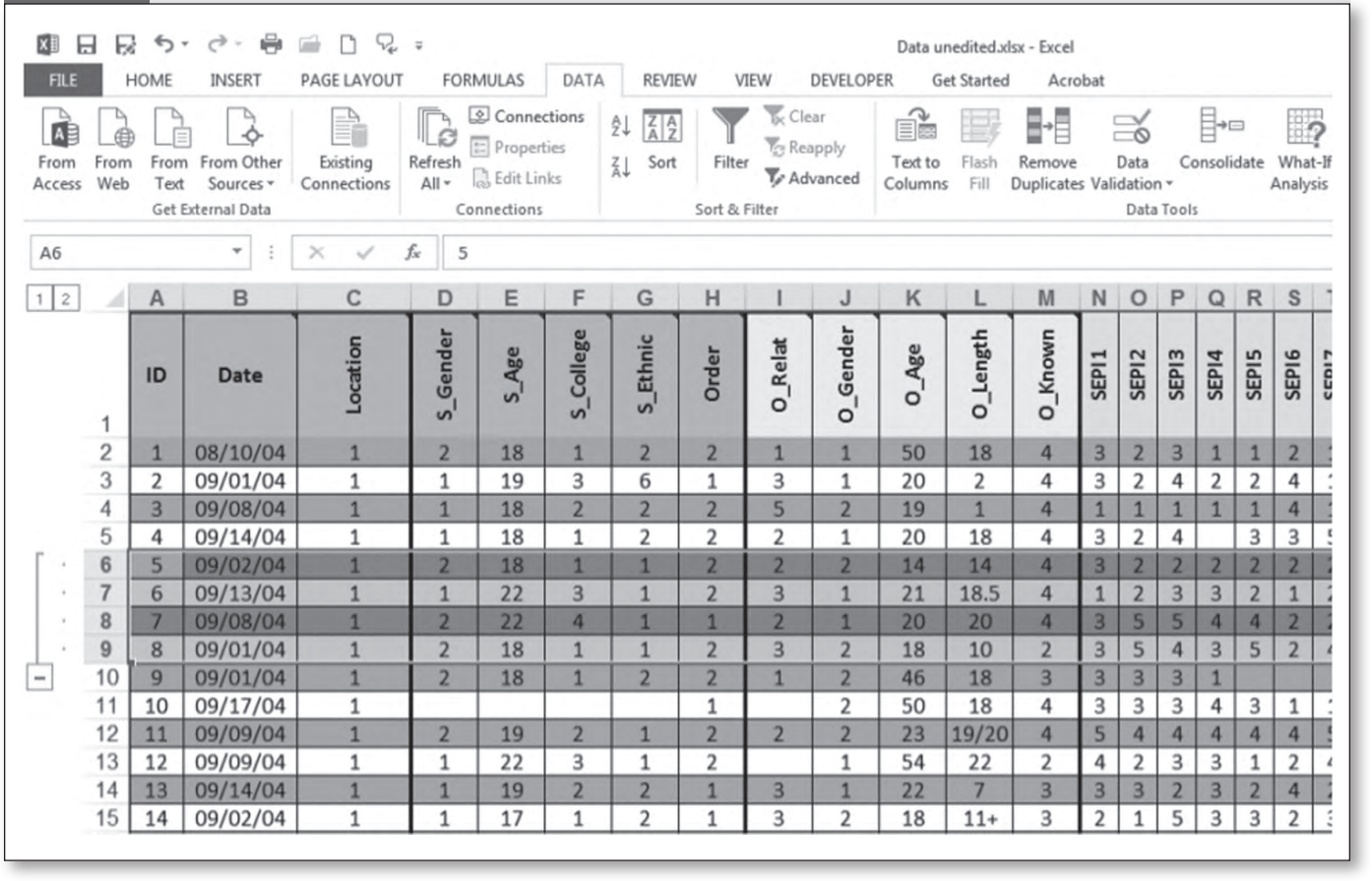
Group

An alternative to hiding is offered by grouping rows or columns. Once you've selected the rows you want to hide, click on *Data, Group* (see Figure 5.9). You can now click on the minus sign associated with the group to hide the rows, and once they are hidden, click on the plus sign to reveal them again. *Ungroup* removes the grouping.

Comparing the Options

Freeze Panes is a great tool, and it probably should always be used during data entry. The other tools covered in this section are more relevant when you want to, for example, compare what you've entered in Column F with what you've entered in column AX. Split, Hide, and Group all let you compare any two portions of the dataset, an option not available in any statistical software. So which should you use?

Figure 5.9 Grouped Rows



	ID	Date	Location	S_Gender	S_Age	S_College	S_Ethnic	Order	O_Relat	O_Gender	O_Age	O_Length	O_Known	SEPI1	SEPI2	SEPI3	SEPI4	SEPI5	SEPI6
2	1	08/10/04	1	2	18	1	2	2	1	1	50	18	4	3	2	3	1	1	2
3	2	09/01/04	1	1	19	3	6	1	3	1	20	2	4	3	2	4	2	2	4
4	3	09/08/04	1	1	18	2	2	2	5	2	19	1	4	1	1	1	1	1	4
5	4	09/14/04	1	1	18	1	2	2	2	1	20	18	4	3	2	4		3	3
6	5	09/02/04	1	2	18	1	1	2	2	2	14	14	4	3	2	2	2	2	2
7	6	09/13/04	1	1	22	3	1	2	3	1	21	18.5	4	1	2	3	3	2	1
8	7	09/08/04	1	2	22	4	1	1	2	1	20	20	4	3	5	5	4	4	2
9	8	09/01/04	1	2	18	1	1	2	3	2	18	10	2	3	5	4	3	5	2
10	9	09/01/04	1	2	18	1	2	2	1	2	46	18	3	3	3	3	1		
11	10	09/17/04	1					1		2	50	18	4	3	3	3	4	3	1
12	11	09/09/04	1	2	19	2	1	2	2	2	23	19/20	4	5	4	4	4	4	4
13	12	09/09/04	1	1	22	3	1	2		1	54	22	2	4	2	3	3	1	2
14	13	09/14/04	1	1	19	2	2	1	3	1	22	7	3	3	3	2	3	2	4
15	14	09/02/04	1	1	17	1	2	1	3	2	18	11+	3	2	1	5	3	3	2

Note: Rows 6–9 have been grouped. Clicking on the minus sign will hide them. The minus sign then converts to a plus sign that, if clicked, makes the rows visible again.

The easy answer is that any of them will do the job in most instances. If you're looking for the perfect solution in your situation, there are differences worth knowing among the three options. Hide and Group differ from Split in three ways. First, they allow you to review columns or rows from three or more regions of the sheet simultaneously because you can hide multiple sets of rows or columns. Second, portions hidden with Group or Hide will not print, but the entire sheet still prints when Split is in use. Third, Split allows you to scroll around to compare different sections.

Group has one disadvantage compared to Hide: Group takes three steps to hide rows or columns (select the array, click Group, and then click on the minus sign) to Hide's two (select the array, and click Hide). Believe me, you will notice the difference, at least until the process becomes automatic to you. On the other hand, the presence of the plus sign is a constant reminder that rows or columns have been grouped. Also, once rows or columns are grouped, it's easier to toggle between hiding and showing the array than with Hide. If you think you will want to hide and unhide the array on a regular basis, Group is the better option.

Tip: When you want to compare different portions of the data, Split, Group, and Hide will all do in most situations. Grouping is the best option if you think you will have a reason to toggle back and forth between hiding and showing the same array of rows or columns. If this is a one-time thing and you only want to compare two points in the data, Split is probably your best choice. If it's a one-time thing and you want to compare more than two points in the data, Hide is the easiest option. However, remember to Unhide the array again before you do any more data entry.

ENTERING DATA

Finally, your preparations are complete, and you can begin to enter data. I know this has seemed like a great deal of preparation, but if you use the tips I've introduced so far, I think you will find the chance of errors is substantially reduced. There are a few issues during the data entry to consider as well.

Speak Cells

Speak Cells actually refers to a set of Excel commands that can be used to get Excel to read the contents of cells out loud, assuming your computer has

speakers or you use headphones. One of the Speak Cell commands, *Speak Cell on Enter,* can be particularly useful when you're entering data. By default, Speak Cell on Enter is in the Quick Access Toolbar. If you look at the Quick Access Toolbar in any of the figures in this book, it looks like a little word balloon from a cartoon. Clicking on Speak Cell on Enter toggles the feature on and off. Once it's on, when you enter data into a cell and hit Enter, Excel will tell you the contents of the cell. Another option, *Speak Cells,* will sequentially read each cell in the sheet out loud. It can also be added to the Quick Access toolbar, but it cannot be accessed through the Ribbon.

Tip: Some people find Speak Cells on Enter very helpful to prevent data entry errors. Other people find it extremely annoying (e.g., me). You decide.

Autocomplete and Related Tools

When you are manually entering data that involves text variables, you will undoubtedly see Autocomplete in action because this feature is turned on by default in Excel. When you enter text, Autocomplete attempts to find previous entries in the same column that match what you are typing and suggests them as possible entries.

To demonstrate Autocomplete, imagine one of your variables contains first names, and these are the first four cases whose first name begins with a "D":

Case 12: Daniel
Case 23: Darlene
Case 42: Daniel
Case 53: Dan

When you type the D from "Daniel" for case 12, Autocomplete looks at the previous entries in that column, finds none that begins with D, and so gives up. You finish typing "Daniel" and move on. Now you come to case 23 and type the D. Autocomplete again looks for earlier entries that match and this time suggests "Daniel" as the value. That happens to be wrong, so you keep typing, and once you get to the "r," Excel recognizes there is no prior match and Autocomplete gives up. At case 42, you type in "Da." At this point, Excel sees two relevant preceding entries, so Autocomplete waits. When you type in the "n," though, it now has only one prior suggestion, so offers "Daniel" as the value. To accept it, you hit the Enter, Tab, or an arrow key and Excel completes

the entry. Finally, at case 53 you type in "Dan." Once you type the "n," Excel again suggests "Daniel" as an option, but instead you hit Delete to stop Autocomplete. To summarize, when Autocomplete suggests an entry you can continue typing, you can hit Enter to accept the suggestion, or you can hit Delete to tell Autocomplete to mind its own business.

There are several issues to know about Autocomplete. First, Autocomplete scans the entire column for matches, not just those cells above the current cell, but because data entry usually proceeds from top to bottom in the sheet, this may not matter in practice. Second, Excel terminates searching for a match if it comes upon a completely empty row, so if Autocomplete doesn't work, it probably means you have an empty row in your data set. Third, Autocomplete searches only in the same column; information entered in column C will not be considered for completing entries in column D. Finally, you can turn off Autocomplete under *File, Options, Advanced*. However, Autocomplete only reduces the speed of data entry when you use the Delete key more often than Enter, so I've never known a situation where it was worth turning it off.

There are also several keyboard shortcuts that provide similar functions to Autocomplete. Unlike Autocomplete, these work with numbers, dates, and times as well as text:

1. To fill a cell with the contents of the cell immediately above it, press CTRL+D (because it's down from the source).
2. To fill a cell with the contents of the cell to its left, press CTRL+R (to the right of the source).

There are also options for filling a cell with the adjacent cell below it or to the right, but those tend not to be useful in data entry.

One final feature that attempts to streamline what you're entering is called *Autocorrect*. This tool attempts to improve the accuracy of your entry. However, sometimes what Excel thinks needs fixing is what you intended. For example, Excel (and Word) automatically converts the entry (c) into the copyright symbol ©. This is a nice shortcut when you're dealing with copyright issues, but if your goal is to enter (c) it's a little annoying when your software decides it knows what's best for you. You can use Undo to remove the autocorrect. If you will frequently be entering some data that Excel keeps changing, you can control Autocorrect through *File, Options, Proofing*.

Tip: Autocomplete can be very useful with text variables. If Autocorrect gets annoying, just turn it off.

AFTER DATA ENTRY

Once your data are completely entered, there are a few more tricks you can use to identify errors. The first issue to address is the unfortunate truth that sometimes a case or variable name gets entered more than once. This can sometimes happen when data sets are combined, for example, or if data entry is divided among several people. I will first discuss identifying duplicate cases, then identifying duplicate variable names.

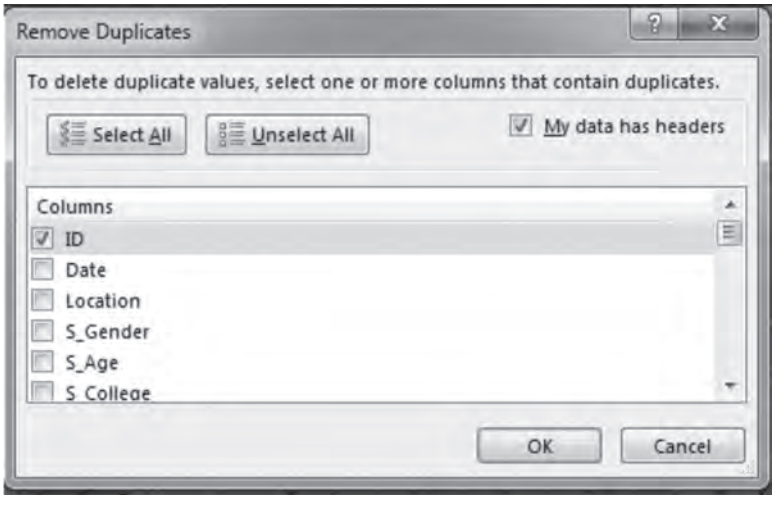
Identifying Duplicate Cases

This is one place where allocating a unique ID to each case from the very beginning is really helpful, because it eases the task of identifying duplicates. In the following, it is assumed the ID variable was entered into column A.

The simplest way to deal with duplicates is to select *Remove Duplicates* on the Data ribbon, which reveals the Remove Duplicates dialog box (see Figure 5.10). Excel tries to determine whether there are headers (variable names) in the first row, but you may need to check *My Data Has Headers*. By default, all columns are checked, meaning that a case is not considered a duplicate unless it matches a previous case on every variable. If it's possible there are errors in data entry so the cases don't match perfectly, you might want to change that. In this case, I clicked on Unselect All, then checked the ID column. When you click on OK, Excel will compare ID values across all cases, and any case where the ID value matches the value in a previous row will be deleted.

This method is very efficient, and it will inform you how many records were deleted when done. I don't like how automatic it is, though: I prefer to review data before I delete it forever. Maybe participant 24 was entered twice by mistake, but the second entry was more accurate than the first. If so, you're deleting the more accurate version of the data, but you'll never know if you use Remove Duplicates. Also, there have been instances when I've been curious whether it's actually the same case or whether an ID number was assigned twice by mistake.

Figure 5.10 The Remove Duplicates Dialog Box



For that reason, I prefer to identify rather than remove duplicates, then decide what needs to be done to fix the data. Figure 5.11 illustrates one way to identify duplicates without automatically removing them.

1. Move to the first blank column to the right of your data and enter the following formula in row 3 (the second row that contains data):

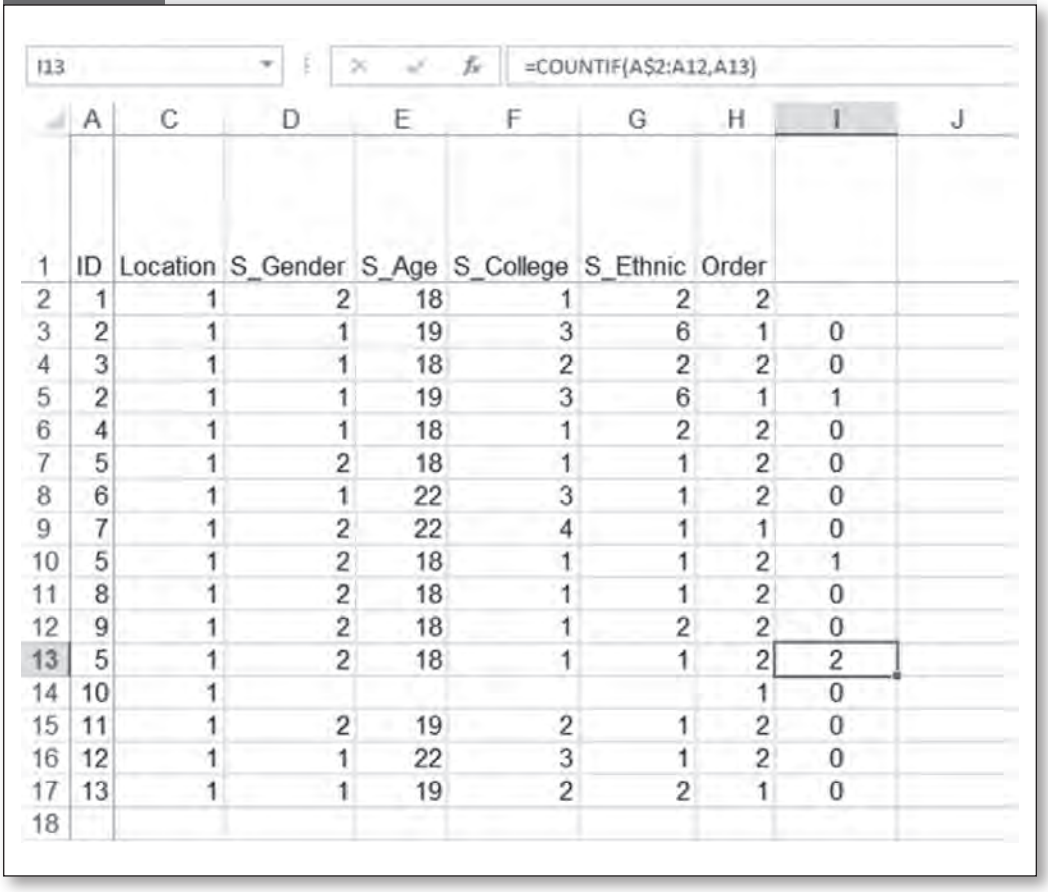
   =COUNTIF(A$2:A2,A3)

2. Copy this formula to the remaining rows in the data set. Row numbers will update when copied, except for the initial A$2, so that, for example, in row 13 the formula will read:

   =COUNTIF(A$2:A12,A13)

In row 3, what this formula does is count the number of cells in the array A2:A2 that equal the value in cell A3. The formula returns 0, indicating there are no earlier values in column A with the same ID number as that in row 3. In row 13, it counts the number of cells in the array A2:A12 that equal the value in cell A13. The formula returns 2, indicating two previous cases with an ID of 5. Now you can directly compare the rows with the same ID—in a large data set this is

Figure 5.11 Checking for Duplicate IDs



	A	C	D	E	F	G	H	I	J
1	ID	Location	S_Gender	S_Age	S_College	S_Ethnic	Order		
2	1	1	2	18	1	2	2		
3	2	1	1	19	3	6	1	0	
4	3	1	1	18	2	2	2	0	
5	2	1	1	19	3	6	1	1	
6	4	1	1	18	1	2	2	0	
7	5	1	2	18	1	1	2	0	
8	6	1	1	22	3	1	2	0	
9	7	1	2	22	4	1	1	0	
10	5	1	2	18	1	1	2	1	
11	8	1	2	18	1	1	2	0	
12	9	1	2	18	1	2	2	0	
13	5	1	2	18	1	1	2	2	
14	10	1					1	0	
15	11	1	2	19	2	1	2	0	
16	12	1	1	22	3	1	2	0	
17	13	1	1	19	2	2	1	0	
18									

a point when Group or Hide will come in handy to hide the intervening rows—and decide which to retain or whether this was an error in entering the ID. Eyeballing the results for this formula, you hope you see nothing but 0s, but other values indicate data deserving a closer look. In a very large data sets, where I didn't trust I would catch a non-zero value, I've sometimes used conditional formatting to shade cells where the formula produces a result greater than zero. Don't forget to remove the column of formulas later, so your statistical software doesn't think that column is part of your data!

If your cases are spread across multiple sheets, neither of the methods described will compare all IDs. However, you could copy IDs from all sheets to a column in a new sheet, so the entire set of IDs from all sheets are collected in one column, and then use the second method to look for duplicates across all your data.

Identifying Duplicate Variable Names

This strategy can be modified to determine whether variable names are duplicated. Copy and paste the variable names into a new sheet. If your variables are spread across multiple sheets, first create a new sheet and copy all the variable names from each sheet into row 1 of the new sheet, so the entire set of variable names across all sheets are collected together. Now in cell B2, enter the formula

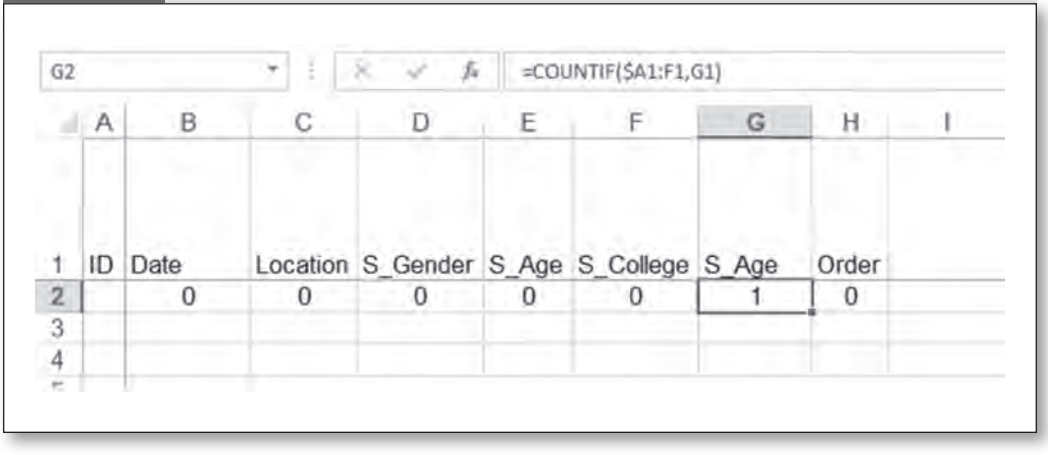
=COUNTIF($A1:A1,B1)

Copy and paste this into each column with a variable name. In Figure 5.12, the value 1 in cell G2 indicates a previous variable exists called S_Age.

Double Entry

Data Validation is a remarkably useful tool for making sure no unexpected values appear in the data set, and everything discussed in this chapter contributes to more accurate data entry. However, what I've described so far is not a complete solution to the problem of error. Even Data Validation can't detect errors that are within the permissible range. For example, Data Validation will make sure a 6 is not entered for a variable that should only be 1–5, but it will do nothing to keep the person entering the data from typing in a 3 when the

Figure 5.12 Checking for Duplicate Variable Names



correct value is a 4. As I noted earlier, such errors will usually have little impact on the results, but you want to avoid all error if possible.

One option is to double-check every entry. This is a labor-intensive task, especially for very large data sets, particularly because the traditional way to check data is to have one researcher read the Excel values out loud while another looks at the original data entry form, or vice versa. This can be made more efficient by using Speak Cells, but it still requires a great deal of time. Also, double-checking is a pretty tedious process and is itself susceptible to a high error rate. Ah, Tahiti!

Double entry offers a better option. It can also be labor-intensive, but it's not quite as bad as reading hundreds of previously entered values. It also has the advantage that it reduces the likelihood of errors surviving the process nearly to zero.[2]

For this example, consider a workbook with several sheets of data. One called *Pre* contains pretest data in columns A:X (with ID numbers in column A) and rows 1:365 (with variable names in row 1). Follow these steps:

1. Once all data are entered, create a completely new workbook with one sheet for each data entry sheet in the original workbook. Give these sheets names that parallel the names in the original file so it's clear which sheet in the new workbook corresponds to which sheet from the old (e.g., the sheet intended to correspond with the sheet *Pre* could be called *Pre2*). To be really meticulous, the people responsible for entering the data a second time should have no access to the original workbook.

2. For each sheet in the new file, copy the entire set of variable names from the original corresponding sheet to the first row and ID numbers to the first column.

3. Decide how much of the data set to double-enter. If possible, double enter the entire data set. If the original data set is very large, a subset of cases can be selected at random for double entry, though the minimum would be about 20% of cases, perhaps by choosing every fifth case. Because the ID numbers have been copied directly from Pre in the original order, double entering a subset of cases means some rows in the new data set will be left blank. More important for the process, double-entered cases will appear in the same row as in the original sheet.

2. Double entry can be accomplished with statistical software instead of with Excel, but Excel is a much more efficient tool for the job. For one thing, double entry in SPSS or SAS would require new names for all reentered variables.

4. Once the data have been entered a second time, copy the new sheets to the original workbook. As mentioned earlier, this involves right-clicking the sheet tab in the new workbook and selecting Move or Copy. Now you will have one workbook with pairs of sheets, one from the original data entry with all cases entered and one with repeated entry of at least 20% of cases.

5. For each pair of sheets, create a third sheet, again using a name to connect them, such as *Pre3*. Again copy the variable names across the first row and ID numbers down the first column.

6. In cell B2 of sheet Pre3, enter the following formula

```
=IF(COUNTA(Pre2!$B2:$X2)=0,"",IF(Pre!B2=Pre2!B2,"","X"))
```

7. Copy this cell, highlight all cells in Pre3 from B2 to X365, then Paste.

Consider what this formula does in cell B2. COUNTA counts the number of non-empty cells in the array B2:X2. If that sum is 0 it means no data were entered into Pre2 for that case, so it must be one of the cases not being double-entered. It therefore leaves cell Pre3!B2 empty (that's the first double quotes). If the sum is greater than 0, it means data were entered for that case in Pre2, so those data should be compared with the original data in Pre. That's the purpose of the second embedded IF function. This function checks whether the values in Pre!B2 cells and Pre2!B2 are equal. If they are, then Pre3!B2 is left empty. If they are unequal, which would suggest an error either in Pre or Pre2 or both, then an X appears in Pre3!B2. The $ before the B and X is important because it enables you to copy the formula from column B to other columns but keep the references for the COUNTA function correct.

The result is that the array of cells B2:X365 in Pre3 will only show an X if the case was reentered and if the corresponding cells from Pre and Pre2 have different values. Unless exactly the same error was entered into both Pre and Pre3, these Xs will flag every error in the two sheets.

There are all sorts of ways you could enhance the double entry process. Here are three examples:

1. You could set conditional formatting for the cells in Pre3 using the formula above, so that mismatches between Pre and Pre2 will also result in a cell that looks different (e.g., it's shaded bright red).

2. If you select cases in rows 5, 10, and so on, for double entry, instead of flagging each row double-entered, you could modify the formula as follows:

   =IF(Row()/5=INT(ROW()/5,IF(Pre!B2=Pre2!B2,"","X"),"")

   I'll just put it out there and let you work through it.

3. You can enter the following formula into an empty cell, let's say Y1, in Pre3:

   =COUNTIF(B2:X365,"X")

   The result will be the number of cells in Pre3 that contain Xs, which is the number of entries that need to be checked.

Double entry also offers a particularly neat example of using Excel's automatic formula updating. Suppose an X appears in cell C24. Once you correct the mistake that caused the problem, the X in Pre3!C24 will disappear, and the counter in cell Pre3!Y1 will drop by 1. The value in Y1 will, therefore, give you a countdown of the number of cells still needing to be checked. For this reason, I recommend correcting errors even if they appear in Pre2: You want to reach the point where that counter in Y1 equals 0, indicating there are no disagreements left. It's a very reinforcing experience to receive immediate visual feedback as each error is eliminated from your data set. By the time you're done, you have efficiently identified and eliminated all instances of inaccurate data entry, at least in those rows that were double entered. Once all the Xs are gone from Pre3, you might want to delete the double entry and comparison worksheets such as Pre2 and Pre3 so that what remains is the complete data set you'll use for your statistical analyses.

Though I suggested you could double-enter a subset of cases, doing so can create a dilemma. If the errors are rare, you probably will feel OK about only checking 20%, but what if the errors are frequent? If so, you probably should go back and double-enter the entire data set to correct all errors. Unfortunately, there is no standard for rare versus frequent errors, but if the majority of the cases double-entered have at least one error, you should be seriously considering double-entering the entire data set.

Tip: If there is one tip that summarizes important things to know from this chapter, it's this one. There are four tools that make Excel the best choice for manual data entry: Freeze Panes, Data Validation, setting the format for variables using the Number tab in the Format Cells dialog box, and double entry. If you use all four, you will make your data entry efficient and, more important, extremely accurate.

CHECK YOUR UNDERSTANDING

Some of these problems require that you download the file Chap5.pdf from http://www.sagepub.com/mcgrath.

C5-1. What is Excel's biggest drawback as a data entry platform?

C5-2. Name three possible reasons for spreading data entry across multiple sheets.

C5-3. Create a workbook for the data contained in Chap5.pdf. In creating the workbook, you should address the following issues. Make some mistakes first time around to make the double entry more interesting.

1. Variable names
2. The data dictionary
3. Missing data
4. Using an Excel table
5. Data validation
6. Number formatting
7. Freeze panes
8. Double entry