CHAPTER 2

Data Coding, Entry, and Checking

This chapter begins with a very brief overview of the initial steps in a research project. After this introduction, the chapter will focus on: 1) getting your data ready to enter into SPSS or a spreadsheet, 2) defining and labeling variables, 3) entering the data appropriately, and 4) checking to be sure that data entry was done correctly without errors.

Plan the Study, Pilot Test, and Collect Data

Plan the study. As discussed in Chapter 1, the research starts with identification of a research problem and research questions or hypotheses. It is also necessary to plan the research design before you select the data collection instrument(s) and begin to collect data. Most research methods books discuss this part of the research process extensively (e.g., see Gliner and Morgan, 2000).

Select or develop the instrument(s). If there is an appropriate instrument available and it has been used with a population similar to yours, it is usually desirable to use it. However, sometimes it is necessary to modify an existing instrument or develop your own. For this chapter we have developed a short questionnaire to be given to students at the end of a course. Remember that questionnaires or surveys are only one way to collect quantitative data. You could also use structured interviews, observations, tests, standardized inventories, or some other type of data collection method. Research methods and measurement books have one or more chapters devoted to the selection and development of data collection instruments. A useful book on the development of questionnaires is Salant and Dillman (1994).

Pilot test and refine instruments. It is always desirable to try out your instrument and directions with, at the very least, a few colleagues or friends. When possible, you also should conduct a **pilot study** with a sample similar to the one you plan to use later. This is especially important if you developed the instrument or it is going to be used with a population different from the one(s) that it was developed for and has been used with in the past.

Pilot participants should be asked about the clarity of the items and whether they think any items should be added or deleted. Then, use the feedback to make modifications in the instrument before beginning data collection. **Content validity** can also be checked by asking experts to judge whether your items cover all aspects of the domain you intended to measure and whether they are in appropriate proportions relative to that domain.

Collect the data. The next step in the research process is to collect the data. There are several ways to collect questionnaire or survey data (such as telephone, mail, or e-mail). We will not discuss them here because that is not the purpose of this book. The Salant and Dillman (1994) book, *How to Conduct Your Own Survey*, provides considerable detail on the various methods for collecting survey data.

You should check your raw data after you collect it even before it is entered into the computer. Make sure that the participants marked their score sheets or questionnaires appropriately; check to see if there are double answers to a question (when only one is expected) or answers that are marked between two rating points. If this happens, you need to have a rule (e.g., "use the

average") that you can apply consistently. Thus, you should "clean up" your data, making sure they are clear, consistent, and readable, before entering them into a data file.

Let's assume that the completed questionnaires shown in Fig. 2.1 and 2.2 were given to a small class of 12 students and that they filled them out and turned them in at the end of the class. The researcher numbered the forms from 1 to 12 as shown opposite ID.

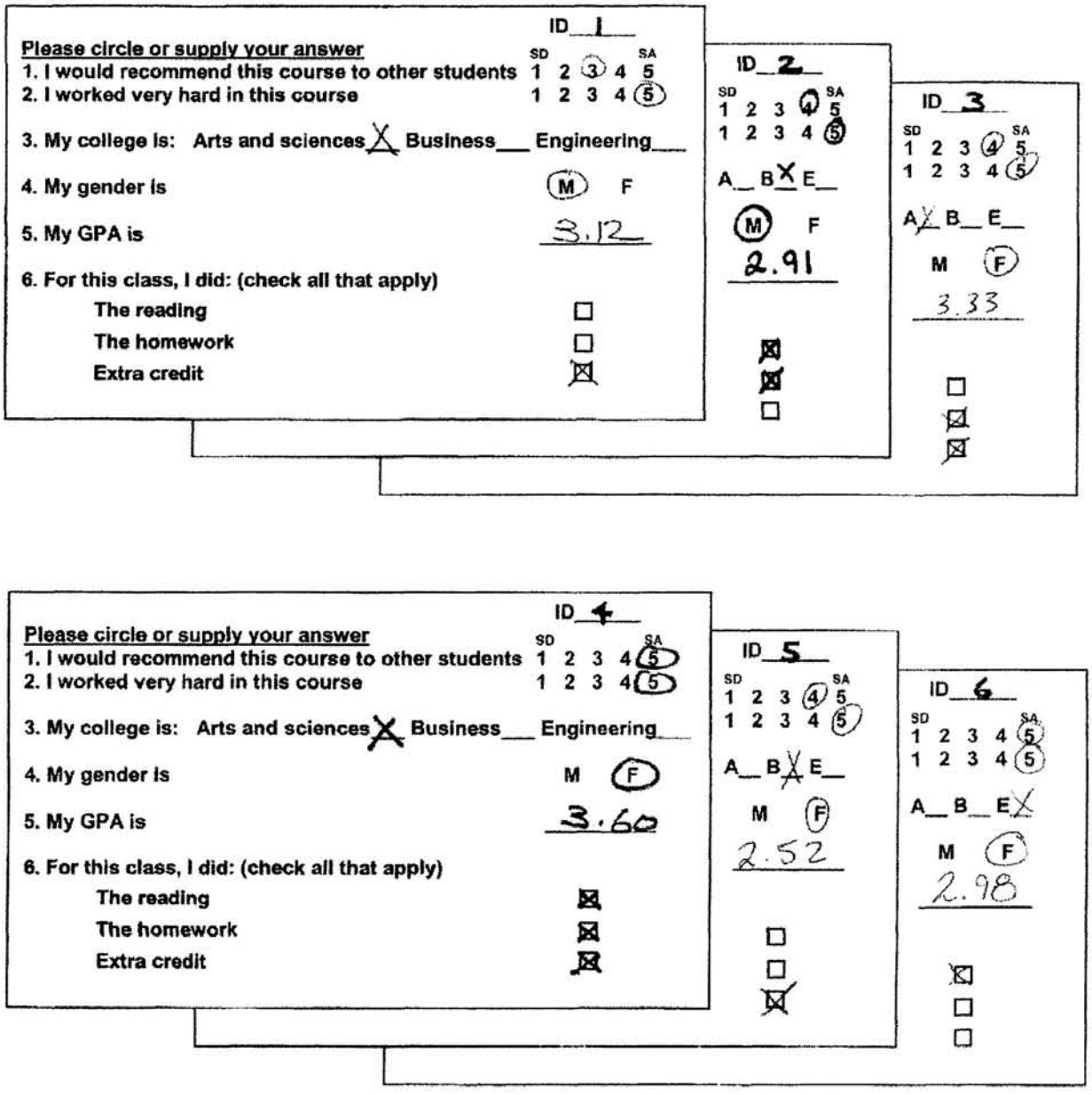


Fig. 2.1. Completed questionnaires for participant 1 through 6.

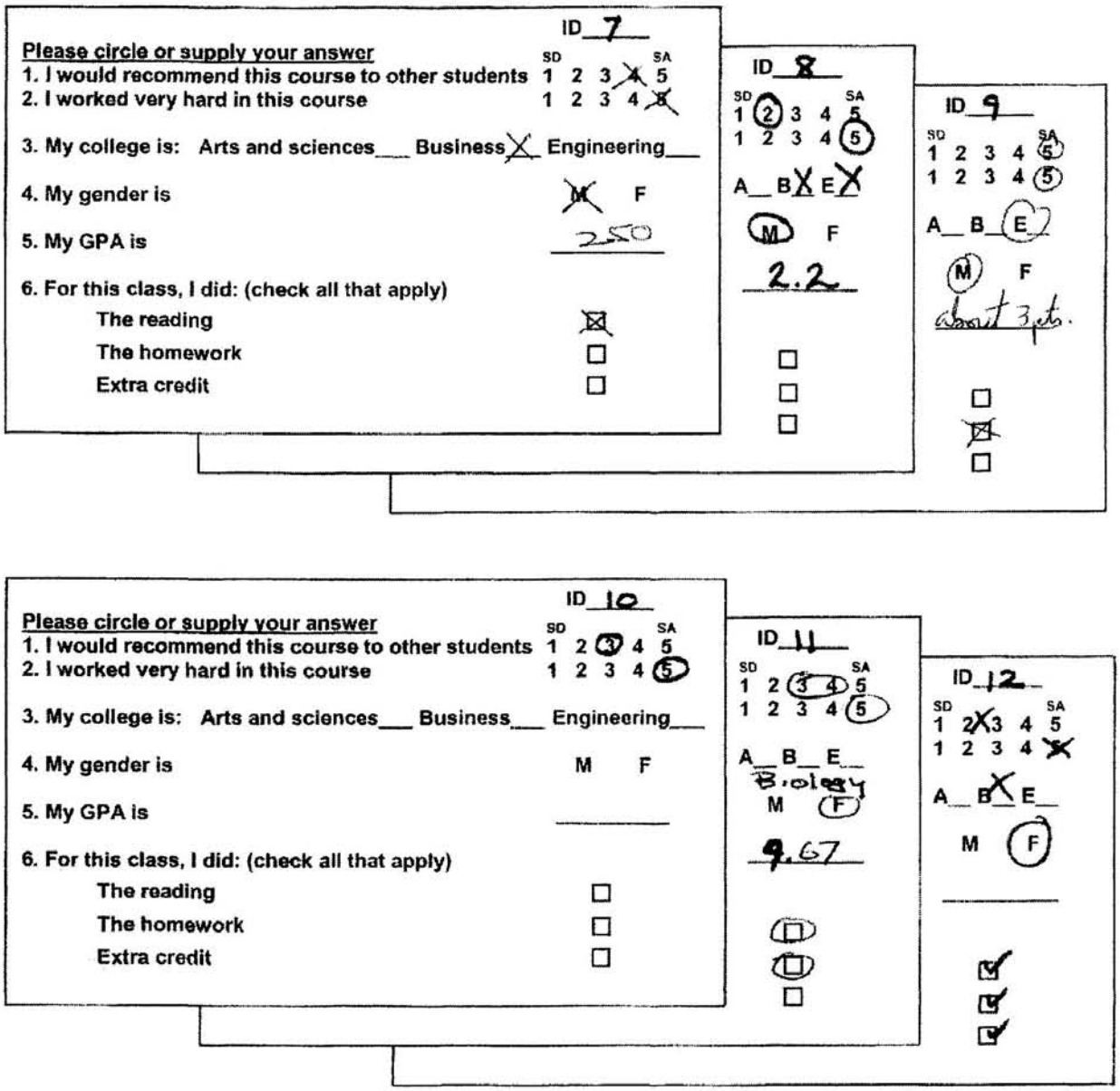


Fig. 2.2. Completed questionnaires for participants 7 through 12.

After the questionnaires were turned in and numbered, the researcher was ready to begin the coding process, which we will describe in the next section.

Code Data for Data Entry

Rules for Data Coding

Coding is the process of assigning numbers to the values or levels of each variable. Before starting the coding process we want to present some broad suggestions or rules to keep in mind as you proceed. These suggestions are adapted from rules proposed in Newton and Rudestam's (1999) useful book entitled *Your Statistical Consultant*. We believe that our suggestions are appropriate, but some researchers might propose alternatives, especially for "rules" 1, 2, 4, 5, and 7.

1. ***All data should be numeric.*** Even though it is possible to use letters or words (string variables) as data, it is not desirable to do so with SPSS. For example, we could code gender as M for male and F for female, but in order to do most statistics with SPSS you would have to convert the letters or words to numbers. It is easier to do this conversion before entering the data into the computer. As you will see in Fig. 2.3, we decided to code females as 1 and males as 0. This is called **dummy coding**. In essence, the 0 means "not female." We could, of course, code males as 1 and females as 0, or we could code one gender as 1 and the other as 2. However, it is crucial that you be consistent in your coding (e.g., for this study all males are coded 0 and females 1) and have a way to remind yourself and others of how you did the coding. Later in this chapter we will show how you can provide such a record called a **codebook**.

2. ***Each variable for each case or participant must occupy the same column in the SPSS Data Editor.*** With SPSS it is important that data from each participant occupies only one line (row), and each column must contain data on the same variable for all the participants. The SPSS data editor, into which you will enter data, facilitates this by putting the short variable names that you choose at the top of each column, as you saw in Chapter 1, Fig. 1.3. If a variable is measured more than once (e.g., pretest and posttest), it will be entered in two columns with somewhat different names like *mathpre* and *mathpost*.

3. ***All values (codes) for a variable must be mutually exclusive.*** That is, only one value or number can be recorded for each variable. Some items, like our item 6 in Fig. 2.3, allow for participants to check more than one response. In that case the item should be divided into a separate variable for each possible response choice, with one value of each variable corresponding to yes (checked) and the other to no (not checked). For example, item 6 becomes variables 6, 7, and 8 (see Fig. 2.3). Usually, items should be phrased so that persons would logically choose only one of the provided options, and all possible options are provided. A final category labeled "other" may be provided in cases where all possible options cannot be listed but these "other" responses are usually quite diverse and, thus, usually not very useful for statistical purposes.

4. ***Each variable should be coded to obtain maximum information.*** Do not collapse categories or values when you set up the codes for them. If needed, let the computer do it later. In general, it is desirable to code and enter data in as detailed a form as available. Thus, enter actual test scores, ages, GPAs, etc. if you know them. It is good to practice to ask participants to provide information that is quite specific. However, you should be careful not to ask questions that are so specific that the respondent may not know the answer or may not feel comfortable providing it. For example, you will obtain more information by asking participants to state their GPA to two decimals (as in Fig. 2.1 and 2.2), than if you asked them to select from a few broad categories (e.g., less than 2.0, 2.0-2.49, 2.50-2.99, etc). However, if students don't know their GPA or don't want to reveal it precisely, they may leave the question blank or write in a difficult to interpret answer.

These issues might lead you to provide a number of categories, each with a relatively narrow range of values, for variables such as age, weight, and income. Never collapse such categories before you enter the data into SPSS. For example, if you had age categories for university undergraduates 16-18, 18-20, 21-23, etc. and you realize that there are only a few students in the below 18 group, keep the codes as is for now. Later you can make a new category of 20 or under by using an SPSS function, **Transform => Recode**. If you collapse categories before you enter the data, the information is gone.

5. ***For each participant, there must be a code or value for each variable.*** These codes should be numbers, except for variables for which the data are missing. We recommend using blanks when data are missing or unusable, because SPSS is designed to handle blanks as missing values. However, sometimes you may have more than one type of missing data, such as items left blank *and* those that had an answer that was not appropriate or usable. In this case you may assign numeric codes such as 98 and 99 to them, but you must tell SPSS that these codes are for missing values, or SPSS will treat them as actual data.

6. ***Apply any coding rules consistently for all participants.*** This means that if you decide to treat a certain type of response as, say, missing for one person, you must do the same for all other participants.

7. ***Use high numbers (value or codes) for the "agree," "good," or "positive" end of a variable that is ordered.*** Sometimes you will see questionnaires that use 1 for "strongly agree," and 5 for "strongly disagree." This is not wrong as long as you are clear and consistent. However, you are less likely to get confused when interpreting your results if high values have positive meaning.

Make a Coding Form

Now you need to make some decisions about how to code the data provided in Fig. 2.1 and 2.2, especially data that are not already in numerical form. When the responses provided by participants are numbers, the variable is said to be "self coding". You can just enter the number that was circled or checked. On the other hand, variables such as *gender* or *college* have no intrinsic value associated with them. See Fig. 2.3 for the decisions we made about how to number the variables, code the values, and name the eight variables. Don't forget to number each of the questionnaires so that you can later check the entered data against the questionnaires.

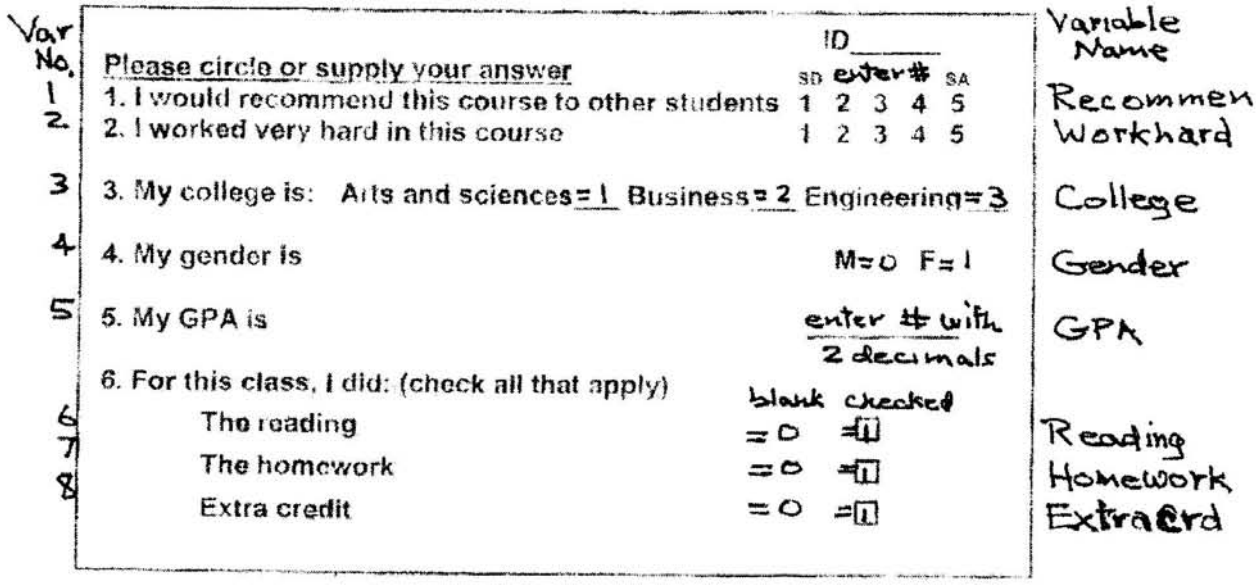


Fig. 2.3. A blank survey showing how to code the data.

Problem 2.1: Check the Completed Questionnaires

Now examine Fig. 2.1 and 2.2 for incomplete, unclear, or double answers. **Stop** and do this now, before proceeding. What issues did you see? The researcher needs to make rules about how to handle these problems and note them on the questionnaires or on a master "coding instructions" sheet so that the same rules are used for all cases.

We have identified at least 11 responses on 6 of the 12 questionnaires that need to be clarified. Can you find them all? How would you resolve them? Write on Fig. 2.1 and 2.2 how you would handle each issue that you see.

Make Rules About How to Handle These Problems

For each type of incomplete, blank, unclear, or double answer, you need to make a rule for what to do. As much as possible, you should make these rules before data collection, but there may well be some unanticipated issues. It is important that you apply the rules consistently for all similar problems so as not to bias your results.

Interpretation of Problem 2.1 and Fig. 2.4.

Now, we will discuss each of the issues and how we decided to handle them. Of course, some reasonable choices could have been different from ours. We think that the data for participants 1 – 6 are quite clear and ready to enter into SPSS with the help of Fig. 2.3. However, the questionnaires for participants 7 – 12 pose a number of minor and more serious problems for the person entering the data. We have written our decision in numbered callout boxes on Fig. 2.4, which are the surveys and responses for subjects 7 – 12.

1. For participant 7, the *GPA* appears to be written as 250. It seems reasonable to assume that he meant to include a decimal after the 2, and so we would enter 2.50. We could instead have said that this was an invalid response and coded it as missing. However, missing data create problems in later data analysis, especially for complex statistics. Thus, we want to use as much of the data provided as is reasonable. The important thing here is that you *must* treat all other similar problems the same way.

2. For subject 8, two colleges were checked. We could have developed a new legitimate response value (4 = other). Because this fictitious university requires that students be identified with one and only one of its three colleges, we have developed two missing value codes (as we did for ethnic group and religion in the HSB data set). Thus, for this variable only, we have used 98, for multiple checked colleges or other written-in responses that do not fit clearly into one of the colleges (e.g., business engineering or history and business). We treat such responses as missing because they seem to be invalid and /or because we would not have enough of any given response to form a reasonable size group for analysis. We used 99 as the code for cases where nothing was checked or written on the form. Having two codes enables us to distinguish between these two types of missing data, if we ever wanted to later. Other researchers (e.g., Newton and Rudestam, 1999) recommend using 8 and 9 in this case, but we think that it is best to use a code that is very different form the "valid" codes so that they stand out if you forget to tell SPSS that they are missing values.

3. Also, subject 8 wrote 2.2. for his *GPA*. It seems reasonable to enter 2.20 as the *GPA*. Actually, in this case if we enter 2.2, SPSS will treat it as 2.20 because we will tell SPSS to use two decimal places.

4. We decided to enter 3.00 for participant 9's *GPA*. Of course, the actual *GPA* could be higher or, more likely, lower, but 3.00 seems to be the best choice given the information provided by the student.

5. Participant number 10 only answered the first two questions, so there are lots of missing data. It appears that he or she decided not to complete the questionnaire. We made a rule that if 3 out of the first 5 items were blank or invalid; we would throw out that whole questionnaire as invalid. In your research report, you should state how many questionnaires were thrown out and for what reason(s). Usually you would not enter any data from that questionnaire, so you would only have 11 subjects or cases to enter. To show you how you would code someone's *college* if they left it blank, we have not deleted this subject.

6. For subject 11, there are several problems. First, she circled both 3 and 4 for the first item; a reasonable decision is to enter the average or midpoint, 3.50.

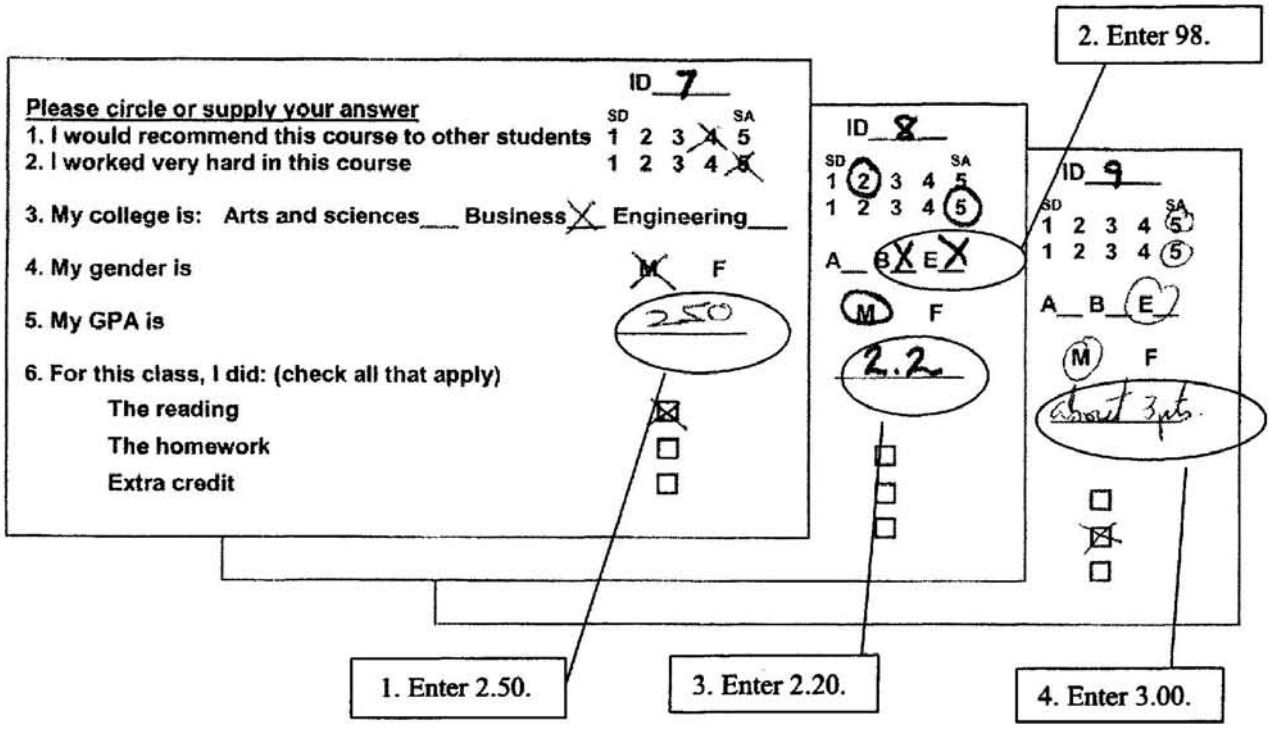
7. Participant 11 has written in "biology" for *college*. Although there is no biology college at this university; it seems reasonable to enter 1 = arts and sciences in this case and in other cases (e.g., history = 1, marketing = 2, civil = 3) where the actual college is clear. See the discussion of issue 2, above, for how to handle unclear examples.

8. Participant 11 also entered 9.67 for the *GPA*, which is an invalid response because this university has a 4-point grading system (4.00 is the maximum possible *GPA*). To show you one method of checking the entered data for errors, we will go ahead and enter 9.67. If you examine the completed questionnaires carefully, you should be able to spot errors like this in the data and not enter them.

9. Enter 1 (checked) for reading and homework for participant 11. Also enter 0 for extra credit (not checked) as you would for all the boxes left unchecked by other participants (except number 10). Even though this person circled the boxes rather than putting X's or checks in them, her intent is clear.

10. As in point 6 above, we decided to enter 2.5 for participant 12's X between 2 and 3.

11. Participant 12 also left *GPA* blank so, using the SPSS general (system) missing value code, we left it blank.



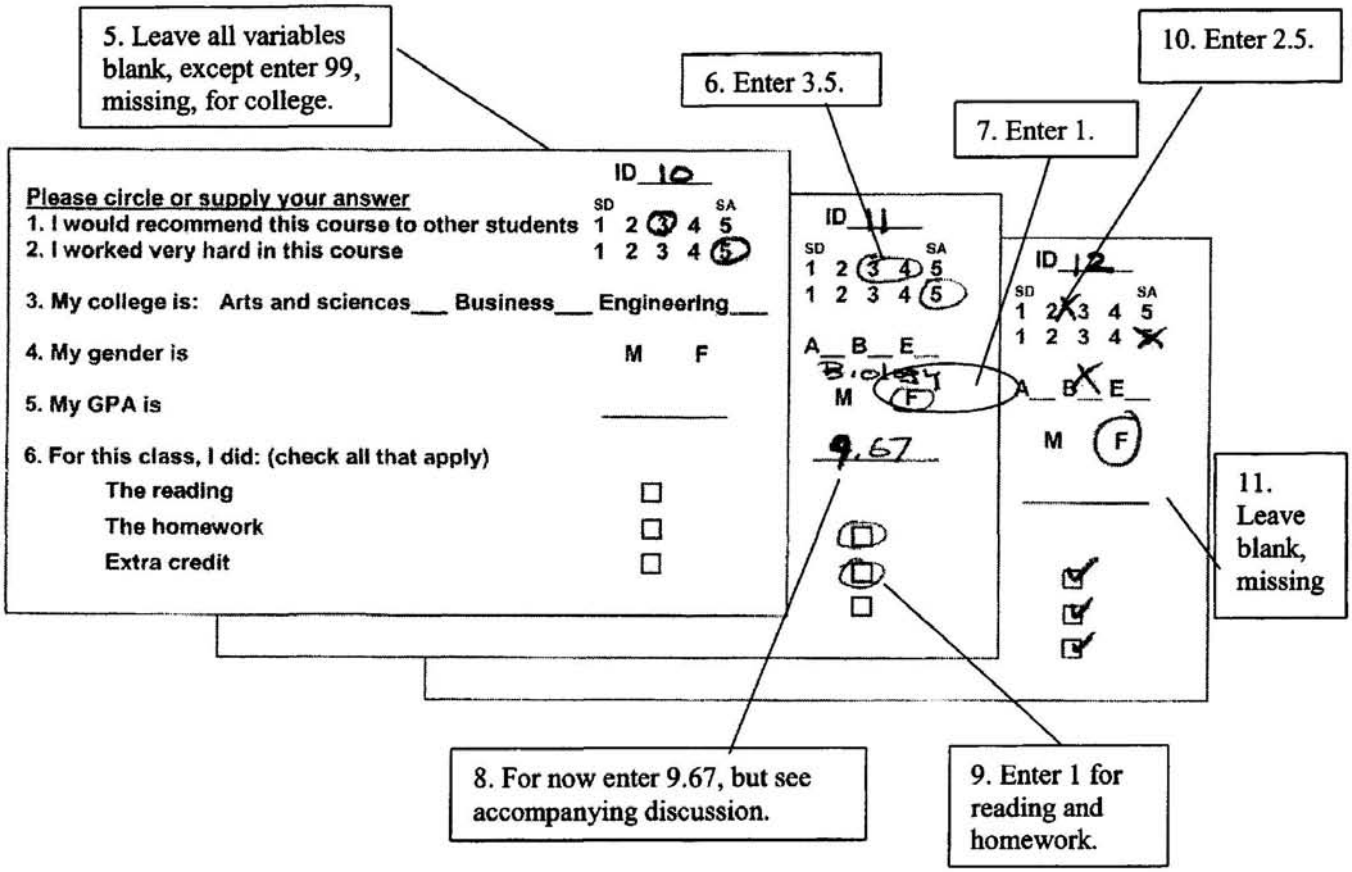


Fig. 2.4. Completed survey with callout boxes showing we handled problem responses.