

SPSS data file specifications

The following specifications are designed to ensure that the resulting data file can be quickly, easily and accurately analyzed. It is usually cost-effective to provide these specifications to data collectors before they commence collecting the data.

When these specifications are not followed it may mean that considerable work is required to appropriately analyze the data, and, sometimes, that it is impossible for the data file to be appropriately analyzed.

SPSS data files have the file extension `.sav`. There are a variety of other SPSS data files with different extensions – SPSS can convert these to `.sav` files (they cannot be converted merely by changing the file extensions).

Dates Date variables, such as interview time, should be stored as data variables in the SPSS data file. If the intent is to change dates when reporting (e.g., move the last couple of days of interviewing in a month into the next wave's reporting), a second date variable should be created which contains this recoded date data.

Non-Response Respondents who were not asked a particular question (i.e., were intentionally or unintentionally skipped), should have a `SPSS SYSTEM-MISSING VALUE`. It is never appropriate to record all missing values in a data file as having a value of 0 (this is very important, as for many binary variables the `NO` response is often coded as a 0, making it impossible to determine which respondents said `NO` and which were not asked the question).

Sometimes it is appropriate to treat missing values for some of the questions as being equivalent to a "No" response (e.g., giving them a value of 0). For example, if people are asked which brands they have consumed, but are only shown brands that they are aware of. In this instance, the question should be included in the data file twice, once with the `SPSS SYSTEM-MISSING VALUE` values and once with the "No" responses instead.

Don't Know The `Don't Know` code needs to be different to the non-response code.

Single Response Questions Single response questions need to be represented in SPSS as one variable. A data file that uses a different variable for each unique response code of a single response question is not useful.

Multiple Response Questions Where there are multiple response variables, a binary variable should be created for each possible response. For example, a question in a

questionnaire may have been:

Q1	Which of the following products do you own?	
	MULTIPLE RESPONSE	
	Savings account.....	1
	Checking Account.....	2
	Credit Card.....	3
	Home loan.....	4

but the data file should be structured as if you had asked the following four questions.

Q1a	Do you have a savings account?	
	No.....	0
	Yes.....	1
Q1b	Do you have a checking account?	
	No.....	0
	Yes.....	1
Q1c	Do you have a credit card?	
	No.....	0
	Yes.....	1
Q1d	Do you have a home loan?	
	No.....	0
	Yes.....	1

Ideally, multiple response questions should be marked as *Multiple Response Sets* in the SPSS data file. If this is not done, Q will automatically guess whether or not variables need to be combined into multiple response questions – this guessing may be inaccurate and require additional work by the user of the data.

Looped questions and image grids Care needs to be taken with the creation of labels for looped questions and image grid questions. Consider a study containing the following three questions:

Q1a	When you think of soft drinks that are <i>sexy</i> , which ones come to mind?	MULTIPLE RESPONSE
	Coke.....	1
	Pepsi.....	2
	Fanta.....	3
	Other.....	4
Q1b	When you think of soft drinks that are <i>masculine</i> , which ones come to mind?	MULTIPLE RESPONSE
	Coke.....	1
	Pepsi.....	2
	Fanta.....	3
	Other.....	4
Q1b	When you think of soft drinks that are <i>powerful</i> , which ones come to mind?	MULTIPLE RESPONSE
	Coke.....	1
	Pepsi.....	2
	Fanta.....	3
	Other.....	4

If the variable labels set up for such questions follow identical structures, this will make the use of the file considerably more straightforward.

<i>Variable Name</i>	<i>Variable Label</i>
Q1a1	Sexy brands: Coke
Q1a2	Sexy brands: Pepsi
Q1a3	Sexy brands: Fanta
Q1a4	Sexy brands: Other
Q1b1	Masculine brands: Coke
Q1b2	Masculine brands: Pepsi
Q1b3	Masculine brands: Fanta
Q1b4	Masculine brands: Other
Q1c1	Powerful brands: Coke
Q1c2	Powerful brands: Pepsi
Q1c3	Powerful brands: Fanta
Q1c4	Powerful brands: Other

By contrast, the following labels would likely cause many hours of frustration for the user:

Q1a1	Sexy brands: Coke
Q1a2	Sexy brands: Pepsi
Q1a3	Sexy brands: Fanta
Q1a4	Sexy brand: Other
Q1b1	Masculine brands: Coke
Q1b2	Masculine brands: Pepsi
Q1b3	Masculine brands: Fanta
Q1b4	Masculine brands: Other
Q1c1	Powerful brands: Coke
Q1c2	Powerful brands: Pepsi
Q1c3	Powerful brands: Fanta
Q1c4	Powerful brands: Others

The problem with this second set of labels is that they are inconsistent. An additional space precedes Pepsi for **Q1b2**, there is no *s* with brands in **Q1a4** and an *s* has been added to *Others* in **Q1c4**. While these may seem like minor issues, they prevent Q from identifying the looped structure in the data – when Q cannot identify this structure, the users cannot conduct many appropriate analyses.

Rankings Ranking questions need to be recorded with a single variable for each item being ranked. Ideally, the most preferred item will have the highest value and the least preferred the lowest, except where the questionnaire expressly indicates an alternative coding.

Open-ended questions Open-ended questions should be coded as per traditional single and multiple response variables. An additional string variable should store the raw responses. Alternatively, Q can be used to recode data. Neither Q nor modern versions

of SPSS have any limit on the length of text that can be contained in open-ended questions (some programs that write SPSS data files have a limit of 256 characters).

Variable labels Variable labels should communicate the information contained in the variable. Variable labels such as `How important is this on a scale of 1 to 10`, provided for each of a set of variables, are of no use as it is impossible to determine what is being rated without referring to the questionnaire. A better variable label would be `Importance: Price`. Where practical, the variable labels should correspond to the actual wording used in the questions. Most programs that write SPSS data files automatically truncate variable labels to 120 characters, which can cause automatically generated labels from looped questions to be uninformative (e.g., the first 120 characters may not include all of the information about the loop).

Value labels Value labels should be taken directly from the questionnaire, provided that their length is 60 characters or less (this is because most programs that write SPSS data files automatically truncate to 60 characters).

Variable Names Variable names should relate to the question numbers. It is often useful if separate question numbering is used for screeners, general questions and classification variables (i.e., `S1, S, ..., Q1, Q2, ..., C1, C2,...`). Where a question is represented by multiple variables, please use a common prefix (e.g., `Q4a, Q4b, Q4c`), rather than out-filing each variable with a different question number (e.g., `Q231, Q232, Q233`). Where a question is a loop of a multiple response question, this is generally best represented via a common prefix and two separate looping suffixes (e.g., `Q4a1, Q4a2, Q4b1, Q4b2`). While these are only guidelines, the core principle is to employ a convention that is easily understandable, whereby the variable names are informative as to the structure of the data.

Variables To Be Excluded Variables that have no possible meaning to the user of the data file should be excluded from the data file. Some data collection programs automatically export useless variables that only relate to the way in which the questionnaire was set up. Examples of variables with no possible meaning that may be exported include:

- ⇒ Looped variables, where one variable will have a value of 1 for every respondent, another will have a value of 2 for all respondents, and so on.
- ⇒ Variables representing un-used codes in multiple response questions.

Variable Formatting Numeric variables (e.g., estimates of number of flights taken in the last 12 months) should be formatted as *Numeric Variables* in SPSS, not as alphanumeric variables. This will prevent the out-filing of illegal numeric values, such as the

always popular -. If possible, please indicate in the data file whether the measure is *Scale*, *Ordinal* or *Nominal*.

ID variable The ID variable in the file should be a *Scale* variable type. That is, it needs to have been out-filed as a numeric rather than text or categorical field.

Weighting Any weighting variables constructed by the data collection company. The variable label should, ideally, include an explanation of the weighting procedure used (e.g., age-by-gender-by-country). If the weight variable is given the variable name of `weight` in the data file, it will automatically be available as a weight within Q.

Max-Diff Max-Diff experiments are questions that present respondents with a series of options and ask them which is best and which is worst. For example, a very simple max-diff experiment may consist of the following two tasks:

Q5a. Which of these companies do you like the most and which do you like the least? *RANDOMIZE ORDER*

Like most		Like least
<input type="radio"/>	Apple	<input type="radio"/>
<input type="radio"/>	Microsoft	<input type="radio"/>
<input type="radio"/>	IBM	<input type="radio"/>
<input type="radio"/>	Google	<input type="radio"/>

Q5b. Which of these companies do you like the most and which do you like the least? *RANDOMIZE ORDER*

Like most		Like least
<input type="radio"/>	IBM	<input type="radio"/>
<input type="radio"/>	Toshiba	<input type="radio"/>
<input type="radio"/>	Dell	<input type="radio"/>
<input type="radio"/>	Apple	<input type="radio"/>

Max-diff questions are setup as follows:

- ⇒ The number of variables required is AT where A is the total number of alternatives (e.g., the total number of brands, or, the total number of attributes being evaluated) and T is the number of tasks. In the example shown above, $T = 2$ and $A = 6$ (i.e. even though only four brands appear in each task, the total number of brands appearing in the two tasks is six), so 12 variables are required in total.
- ⇒ The first A variables relate to the first task, variables $A+1$ to $2A$ relate to the second task, etc.

- ⇒ The *Variable Name* for each variable in the data file should be unique. It is useful to create informative variable names, such as: Task1Alt1, Task1Alt2, Task1Alt3, Task1Alt4, Task1Alt5, Task1Alt6, Task2Alt1, Task2Alt2, Task2Alt3, Task2Alt4, Task2Alt5 and Task2Alt6.
- ⇒ The *Variable Label* for each variable should be the description of the option, even if the option was not presented. For example, for the two tasks, labels would be: Apple, Microsoft, IBM, Google, Dell, Toshiba, Apple, Microsoft, IBM, Google, Dell and Toshiba. Note that:
- The variable labels are repeated the same number of times as there are tasks.
 - The variable labels are always in the same order.
 - No additional information is contained in the variable labels (e.g., no question name, nothing indicating the task).

For example, the variable names and corresponding labels for the questions shown previously could be laid out like this (a different ordering of the brands could be used):

Task1Alt1	Task1Alt2	Task1Alt3	Task1Alt4	Task1Alt5	Task1Alt6	Task2Alt1	Task2Alt2	Task2Alt3	Task2Alt4	Task2Alt5	Task2Alt6
Apple	Microsoft	IBM	Google	Dell	Toshiba	Apple	Microsoft	IBM	Google	Dell	Toshiba

- ⇒ Each variable should contain one of four possible values for each respondent:
- A NaN if the option was not shown.
 - A 1 if the option was chosen as best.
 - A -1 if the option was chosen as worst.
 - A 0 if the option was shown but not selected.

With the example, if the options are laid out in the following order: Apple, Microsoft, IBM, Google, Dell and Toshiba, then the data should look something like this:

Task1Alt1	Task1Alt2	Task1Alt3	Task1Alt4	Task1Alt5	Task1Alt6	Task2Alt1	Task2Alt2	Task2Alt3	Task2Alt4	Task2Alt5	Task2Alt6
0	1	-1	0	NaN	NaN	1	NaN	-1	NaN	0	0
0	1	-1	0	NaN	NaN	0	NaN	-1	NaN	1	0
1	0	-1	0	NaN	NaN	1	NaN	-1	NaN	0	0
0	0	1	-1	NaN	NaN	1	NaN	0	NaN	0	1

In this example, in the first task the respondent chose Microsoft as Best and IBM as worst, while in the second task they chose Apple as Best and IBM as worst again.

Tracking studies Tracking studies should use cumulative data files. That is, a single data file should contain all waves of the study. And, this data file should be prepared from by the data collector (although Q has facilities for merging files, it is usually undesirable to use these for trackers, as it is generally much, much more efficient for the data collector to resolve inconsistencies in the questionnaire and data file that often occur from wave-to-wave).

Where a data file provider is unable to provide a cumulative data file containing all the waves of a study then it will usually not be possible to use Q to analyse the waves of the data that are not in the cumulative data file.

Questionnaires with multiple versions It is common for tracking studies contain slightly different versions of questionnaires. For example, an option in a question previously labelled "New York" may be relabelled as "New York – New York", either because the questionnaire itself changed, or, because of instructions regarding how data from the study should be prepared. Or, a new brand may be added. The following principles will save a lot of time in the analysis of a study with different versions of the questionnaire:

- a) Where at all possible, any changes should be made retrospectively to the raw data file. That is, even if respondents were shown "New York", the system should export the data as if they had been shown "New York – New York". Q has tools to make changes of labels quite straightforward. However, if the data file is created in such a way that it contains different response options for different waves of respondents it will cause a massive increase in the workload required to analyse the data.

- b) Variable names must not change. That is, if variable name Q2a means "Awareness – Coca Cola" in the initial data file, then this name should be retained forever. The variable name is used by Q to work out which data means what. Any changes to the variable name will cause all analyses of the data to "break" and fixing the Q project file to address such changes will generally be extraordinarily difficult.
- c) Where a response option is removed from a questionnaire, then if it is a multiple response question, the variable should remain in the data file but be assigned missing value codes in waves where it did not appear. If it is a single response question the value and label should be left in the data file.
- d) Where a response option is added to a question (e.g., a new brand), this should involve adding new codes (values) if they are single response and new variables if they are multiple response. A variable from a previously deleted response option should not be re-used. Where a new variable is created, respondents from earlier waves of the study need to be assigned missing values.