**Data screening instructions**

Before you do anything else with your data, you need to screen it. Data screening is the process of identifying, fixing, or removing errors from the dataset.

Keep a record of the steps that you take and any changes that you make to the dataset, as you’ll need to summarize these in a short paragraph at the beginning of your results section. This needs to be detailed enough so that someone could replicate your process, but not include any details that don’t make sense to people unfamiliar with your data (e.g. case numbers).

* Skim down the dataset, looking for any completely empty cases (rows).
  + If there is no participant data in these, then this is a situation where the data enterer has accidentally added an extra row – you can delete this, and not worry about recording this step, as no real participant data has been edited or removed.
* Check for out-of-range values:
  + Look at either the survey or data entry guidelines to see what the range is for each item (e.g. 1 to 5), and keep this next to you as you look over the data.
  + Obtain descriptive statistics to see what the minimum and maximum values are for all variables of interest.
  + If you identify any variables with out-of-range values (i.e. for an item with response options from 1 to 5, anything below 1 or above 5 is out of range), apply a filter to your dataset to select only these out of range values.
* For each error, you need to decide whether to accept, replace, or remove it, and *have a good reason for thi*s:
  + Can you presume what the value is? As an example, if 77 was entered in the best/worst marks column, the data enterer likely meant to type 7.7. If there is only one logical option for what this value should have been, you can replace incorrect values.
  + If it is not clear what the value should be, then we should treat this as a cell of missing data – click in the cell and delete it.
  + If that whole case has a lot of erroneous data, then you may choose to remove the entire case (i.e., delete the whole row).
* Next, check for unrealistic responses, for example:
  + Are there any participants who report their age is under 17 years old? Check using min/max score.
  + Check that all reported best/worst hours have realistic values. The easiest way to do this is using filters in *jamovi* (select cases that have a value on these variables that is above the highest reasonable value – you decide and defend the value chosen).
  + Check whether any cases report a higher worstmark than bestmark, which is an unrealistic pair of responses. Again, a filter is the easiest way to do this.
* Once you have finished screening your data, remember to **save the screened data**.
* Write up the summary of this process for your lab report right away. You can then move on to the recoding step.

**Transformations**

After you have removed the errors from your dataset in the data screening stage above, there are several transformations to complete on your data.

The table below explain how your data are currently coded, and any recommended transformations.

| **Variables** | **Current coding/interpretation** | **Recommended transformation** |
| --- | --- | --- |
| Age | Age in years (higher value = older) | - |
| Gender | 1, 2, 3 (1 = male; 2 = female; 3 = other) | Apply labels to categories. |
| Personality1-17 | The interpretation of these variables will depend on results from your EFA. For now, go through the items and identify which most likely measure each of the Big 5 personality traits (openness, conscientiousness, extroversion, agreeableness, emotional stability). | Reverse score items as needed so that higher scores = greater openness, conscientiousness, extroversion, agreeableness, or emotional stability. |
| Procrast1-6 | Values from 1-5 (not at all-very much)  0 – not applicable | Recode 0 scores into missing data. |
| Bestmark, worstmark | Score between 0 and 10, interpreted as expected final mark is between 0 and 100% | Multiply all scores by 10 to convert to a percentage. |
| Besthours, worsthours | Number of hours (higher value = more hours) | - |
| Perfect1-18 | The interpretation of these variables will depend on results from your EFA. For now, go through the items and identify which items are worded so that agreement indicates higher perfectionism, and which are worded in the reverse (1 = strongly disagree; 2 = disagree; 3 = neutral, 4 = agree; 5 = strongly agree) | Reverse code items as needed so that for all items, higher scores = greater perfectionism. |

**How to apply labels to response options**

Select the Gender variable, and in the Levels menu, double click into each level (1, 2 and 3) and type over this with the correct category label so that it is clear what the numbers represent.

Note – *jamovi* does not currently support labels on continuous variables, so it is best to keep in mind what higher values mean for your continuous-type data. Reverse coding them will help with this.

**How to reverse code variables**

Step 1. Check what the response options were – e.g. 1-5 or 1-7

Step 2. Think about what value you would need to subtract your scores from in order to reverse code them, e.g. subtracting scores from 6 would reverse code a 1-5 Likert scale (try it if you want to check!)

Step 3. Create your transformations – select the first variable to recode, and under the data menu, click transform. Enter details of the transformation (name – e.g. recode 1-5, description – e.g. reverse code item scored from 1-5, suffix – e.g. rv5, instructions for the recode - e.g. 6-$source).

Step 4. Apply the transformation to other variables that need to be reverse coded in the same way – do this by selecting the variable, clicking transform, then selecting the transformation.

**How to recode to missing values**

Create a transformation following similar steps above, but the instructions for the recode will be conditional – if participants scored between 1 and 5, they will keep their original value, else they will get missing data.

**How to convert best/worstmark values into a percentage**

The best and worst mark estimates were given on a 1-10 sliding scale. To improve the interpretability of these items, you can multiply original scores by 10 in a Transformation (e.g. $source\*10), and they are then interpreted as the expected mark as a percentage.