Step 4: Clean and process data

To prepare the data for analysis, some checking and cleaning of data will need to be done to ensure that the data is of good quality. Data cleaning can be done on Microsoft Excel after exporting the raw data from the survey platform.

Below are some common issues and possible ways to resolve them:

> Other common software that can be used for data cleaning, processing and analysis (for Step 5):
> - SPSS
> - STATA
> - Jamovi
> - R Studio
> - Python

Duplicate data

Example

Having two rows of data which are completely identical.

This issue could have happened because the service user clicked on the "Submit" button twice for online surveys.

Resolution

Once it has been verified that the two rows are 100% identical and that both responses belong to the same service user, you may proceed to delete either one of the responses.

Service User	Item 1 (Y/N)	Item 2 (Y/N)
Sarah Tan	Y	N
Sarah Tan	Y	N

Missing data

Example

Certain questions do not have data even though the questions are meant to be compulsory.

This could be due to the online survey link having an erroneous skip logic, or if a question was accidentally marked as "optional" instead of "compulsory".

Service User	Item 1 (Y/N)	Item 2 (Y/N)
Sarah Tan	Y	

Resolution

For online survey links, it is important to conduct at least 2 rounds of checks to ensure that skip logics have been applied correctly. Once data collection has begun, it is important to check the data to ensure that there are no missing data. Should there be an error in the skip logic or question settings, the survey link will need to be updated immediately to avoid further errors.

Depending on the nature of the missing data (i.e., missing 'sex' vs missing one of the items in the outcome metric), SSAs may be able to input objective data such as sex or age, but would likely need to discard the response should there be a missing rating for one of the items in the outcome metric.

Illogical data

Example

There are two post-intervention entries for the same service user, but the responses to both are different.

This issue could be due to human error where the service user mistakenly selected "post-intervention" when he/she should have selected "pre-intervention" for the first entry.

Resolution

The system-captured data and time for online surveys are typically accurate. Hence, there might be a need to recode the entry that was completed earlier (by the same service user) to be "pre-intervention".

Service User	Item 1 (Y/N)	Item 2 (Y/N)
Sarah Tan	Y	N
Sarah Tan	N	Y

Following the checking of the data to ensure that the dataset is free of data quality issues, SSAs may proceed to process and organise the data into the desired format (i.e., arranging the columns in the dataset, renaming variables, etc.) to prepare for analysis.

For SEF onboarded programmes, the NCSS team will provide assistance to SSAs in data cleaning and processing. At the end of this step, SSAs should have a dataset that is clean (i.e., free from data issues), paired (i.e., each service user's pre-intervention data has been matched with the respective post-intervention data), and processed (i.e., the data is in an analysis-ready format).

Checkpoints:

☐ **Are you noticing a lot of incomplete data?**

- If you notice a lot of incomplete data (i.e., service users end the survey prematurely), you may need to consider the following:
  - ○ Service users might not understand the survey items and would require further assistance
  - ○ Survey is too lengthy, resulting in service users losing interest
  - ○ Service users might prefer other modes of data collection (e.g., physical forms vs online forms)

☐ **Have you observed any patterns within the service user's response?**

- Some examples include:
  - ○ Having the same rating for consecutive items in the survey (e.g., "Strongly agree" for 10 consecutive items)
  - ○ Ratings which follow sequential or repeated patterns (e.g., Alternating responses between "Strongly agree" and "Agree")
- It would be useful to flag these responses out for further investigation.
- You could approach the service user's caseworker to check on the service user's condition throughout the programme. This may help you make an informed decision on whether the responses are valid or if they should be removed from the dataset.