Data collection protocols: Reflection note from shared learning workshop

Background

This note draws on learning from a capacity-building project in which NatCen worked with several organisations that run programmes aimed at helping young people to enter the workplace.

This reflection note was created following a workshop, which was part of a suite of learning outputs to help support these organisations.

This note is intended to provide a starting point for thinking about experimental and quasi-experimental methods and their use in evaluations. It includes some considerations, hints and tips built up from our experience of using these methods within evaluation work. We have provided references at the end for looking at this topic in more depth.

What are data collection protocols?

A document for describing and standardising the procedures for collecting and storing data. They cover everything from the purpose of data collection, to who collects what data, how, and when, to how impartiality, safety, and privacy are ensured. They outline systematic procedures to ensure that high quality data is collected in all cases. While commonly used for surveys, protocols can be used for various data collection methods (administrative data, monitoring data, interviews, focus groups, observations etc.).

Why are data collection protocols important?

They help ensure that data collection practices are transparent, consistent, and clearly documented. This is especially important when multiple members of staff are involved in data collection, or when there are changes in staffing within an organisation. They can also be useful for establishing who needs to do what, and when.

A data collection protocol can help organisations address a number of problems they might face. Common issues around data collection include:

- missing data
- incomplete/improperly completed measures
- inconsistent data collection practices/schedules/formats
- duplicates
- accessibility issues
- survey fatigue (becoming bored, uninterested or unwilling to take part in surveys or other research studies)
- failure to collect appropriate outcomes

- irrelevant/redundant data collection

How can data collection protocols be used?

They have a number of purposes:

- to guide and standardise data collection processes
- to establish staff's responsibilities in different areas
- to inform staff of data collection processes
- to provide transparency, allowing relevant parties to understand data collection processes in full
- to ensure accountability, and verify whether data is being collected as stated

When should data collection protocols be updated?

Protocols should be updated any time a substantial change in data collection practice is made, for example changes to data collection instruments, schedules, and procedures, changes in named staff members, changes in staff responsibilities, and changes in privacy/information security practice.

Data collection protocol template

An example template is provided at the end of this note (pages 8-12).

It is intended for a single data collection instrument (instrument meaning a mode of data collection, such as a survey). If you are using multiple instruments, then a data protection protocol should be completed for each one.

It covers the following areas of data collection:

1. **Data collection instrument**

   Instrument details (name, content, format, who completes it, storage location)

2. **Who are the data about?**

   Your participants, for example, the young people completing a given programme

3. **Data collection process**

   The context of how and where the data is collected

4. **Team composition**

   Staff involved in data collection and their responsibilities

5. **Data collection schedule**

   When data is collected, including differences between groups

6. **Impartiality, privacy and safety**

   Consent, how data can or cannot be used, incentives, who has access to the data and how that access is controlled, data sharing

7. **Data storage and protection**

   File storage location, removing personal identifiers from the data (anonymisation/pseudonymisation), personal data (information that relates to an identified or identifiable individual), password protection, data storage, data transfers, data deletion

8. **Data processing and validation**

   Data digitalisation, inputting, checking the accuracy and quality of data (validation), missing data

9. **Notes**

   Any other details

Common questions about data collection processes

In this section we outline some elements of data collection that need to be considered when putting together a data protocol. These issues were often raised in our discussions with the organisations we were supporting.

Consent

Informed consent needs to be collected at the start of data collection. If there is more than one data collection timepoint, this information should be provided with the participant, and consent should be obtained for all planned future data collection activities and purposes. It should be made clear that participants can opt out any time and they should be told how to opt out at each timepoint. The materials should also be reviewed to ensure they are age-appropriate, inclusive and accessible (see Accessibility below).

Accessibility

The materials used in data collection (particularly consent materials and survey instruments) should be appropriate for the group(s) or individual young people that are involved in the research. It may be necessary to review these materials and adapt them so that they are inclusive, age-appropriate, and accessible for young people with different needs.

Collecting *informed consent* means that participants enter the research freely and with full information about what it means for them to take part, so the materials explaining the data collection and how data will be used must be written in a way that is easy for your participants to understand.

Data collection instruments (such as surveys) should also be designed to be accessible and appropriate for the intended participants. Some examples of things to consider are given below, but this is not exhaustive:

- The mode of the data collection (e.g. online, telephone, face-to-face, paper and pen): will all young people be able to participate in the chosen mode or should an alternative be offered?
- Question wording: are the questions written in plain English and will the concepts covered be understood by the intended participants?
- Presentation of information: are the questions presented clearly? Think about font, size of text, colours used.
- Language: should the data collection instruments be translated into different languages?

If you are able to do so, testing the materials with the intended audience can help you to identify and address any accessibility issues.

Ensuring data is collected consistently

It is important to communicate the purposes of data collection to the staff members who will be collecting it (e.g. mentors). We suggest carrying out inductions or briefings with staff before data collection starts, to explain how and why data is being collected, and to allow them opportunities to ask questions. Involving previous programme participants in these inductions, to get insight into how they experienced data collection, may be valuable.

Making survey questions mandatory (and avoiding missing data if not)

It is good practice to require a response to each question in a questionnaire (so that it can't be left blank), but do always ensure you have options for "don't know" and "prefer not to say" in case participants do not want to answer your questions. Indeed, patterns of these responses can be informative.

We strongly advise that questions are not made mandatory in surveys without a "don't know" and "prefer not to say" option. This is because (a) it can be unethical to force participants to answer questions they are uncomfortable with, and (b) doing so can be counterproductive, causing participants to give inaccurate or unconsidered responses, or even not complete any of the questionnaire.

Attrition at follow-up

It is common to lose participants at follow-up/endline data collection points. After participants have taken part in the programme, it becomes less likely that they will take part in research associated with it. This is known as attrition. There are a number of strategies to deal with this. These include using incentives, such as vouchers, and planning data collection time points to occur either before young people's involvement ends or at later 'touch points' when you know you will come into contact with them again.

Baselining measures and initial over-scoring

In some circumstances (e.g. when measuring self-report employment skills in a job training setting), participants may report inaccurately high scores at baseline, for example due to social desirability. As participants engage with training their perceptions of their skill levels may change. It is possible that their responses will become more realistic (i.e. lower) as they begin to understand the limits to their current knowledge and skills base. This can lead to a misleading decrease in outcome scores between baseline and endline, when no true decrease occurred.

During our discussions with organisations, multiple solutions have been suggested to account for this issue, some of them are listed here for reference:

- Collect baseline data a few weeks into the programme and/or allow staff discretion in when data is collected. *(We do not recommend this, as it risks introducing bias and inconsistency)*.
- Collect baseline data after a first meeting to build rapport, but before the programme starts. *(This is better, but the staff's involvement may still have started to have an effect by this point)*.
- Collect baseline data at multiple points in the weeks leading up to the programme and take the average. *(This again can help, but still risks bias and puts additional pressure on staff and participants)*.
- Compare the pre-post change for the participants with a control group of comparable non-participants, as in a randomised controlled trial or quasi-experimental design like difference-in-differences. Baseline inaccuracy should then not be an issue if it is equally present in each group. *(This is the best solution but is likely impractical for organisations' routine monitoring and evaluation activities)*.

Regardless of the decision taken, data collection practices should be recorded and reported in a transparent manner. For example, report when data was collected and why.

Appropriate length of data retention

There is no universal standard for how long you should hold on to data after it has been collected and analysed. Generally, data should be retained for as short a time as necessary and no longer than dictated in consent forms and other such documentation. Procedures for data retention and deletion should be transparently established.

Conclusion

A data collection protocol enables you to formalise and standardise data collection processes. It can be very helpful for ensuring that staff practices and responsibilities are well laid out. The process of drafting a data collection protocol can itself help organisations to think about what data they are collecting, how, when, and why.

Useful resources

Better Evaluation (website with evaluation resources):
https://www.betterevaluation.org/

Causal Inference: The Mixtape (more advanced free online textbook on methods for establishing causality in the social sciences, with Stata and R examples):
https://mixtape.scunning.com/

The Magenta Book (HM Treasury guidance on impact evaluation):
https://www.gov.uk/government/publications/the-magenta-book

Top 15 most common data quality issues (and how to fix them).
https://towardsdatascience.com/top-15-most-common-data-quality-issues-and-how-to-fix-them-c1ef0854dca6

UNICEF Overview of Impact Evaluation (website and videos on evaluation):
https://www.unicef-irc.org/KM/IE/impact_1.php

What Works Network (website providing information on the What Works centres):
https://www.whatworksnetwork.org.uk/what-works-centres/

Example Data Collection Protocol

1) Data collection instrument

a) Name of the instrument

b) Mode/method of data collection

i. Paper questionnaire
ii. Online questionnaire
iii. In person interview
iv. Telephone/virtual interview
v. Focus groups
vi. Other (please specify)

c) If it is a questionnaire, who completes the instrument?

i. Young person
ii. Staff/Mentor
iii. Other (please specify)

d) Instrument storage location

i. File cabinet X
ii. Website address
iii. Spreadsheet name
iv. Other (please specify)

e) Content of the instrument

i. List of topic areas
ii. Summarise what you are measuring and how

2) Who are the data about?

a) Young people

i. Young people
ii. Young people in a particular programme (please specify)

b) Staff/Mentor
c) Other (please specify)

3) Data collection process

a) Where are the data collected?

i. On site
ii. At home
iii. At school
iv. Other (please specify)

b) Are the data collected in private?

i. Yes, the young person is alone
ii. Yes, but the young person and staff/mentor are together
iii. No, the data is collected in group
iv. Not known
v. Other (please specify)

4) Team Composition

a) Briefly describe the team and team's responsibilities in the data collection process

b) Report the full names of the staff collecting the data

c) How will you ensure compliance with this protocol?

5) Data collection schedule

a) First point of data collection

b) Future points of data collection

c) How frequently are the data going to be collected?

d) Are there different data collection schedules for different groups?

i. Please specify how many groups and data collection time for each group.
ii. Is there a plan in place in case a data collection point is missed?

6) Impartiality, privacy and safety

a) How will consent be obtained?

b) State location of any consent forms or any similar materials

c) What can this data be used for?

d) Has consent been obtained to use this data for the stated research purposes?

e) What can this data NOT be used for?

f) Describe incentives for participating to the data collection (if applicable)

g) Is completing this instrument mandatory for young people to access your services/programme?

h) Can young people opt out of their data being used for the stated research purposes?

i) Who has access to the data?

j) How is access controlled?

k) Is the data shared with third parties?

7) Data storage and protection

a) Specify location and file name of data files

b) Is the data anonymised, pseudonymised, or neither?[1]

c) If the data is anonymised/pseudonymised, do you keep the raw data or not?

[1] Anonymised data has all identifying information removed such that individuals cannot be re-identified. Pseudonymised data has identifying information removed and replaced with a unique identifier (e.g., a string of numbers), enabling individuals to be re-identified.

d) What personal data are included?

e) Is the file password protected? If yes,

i. Who has the password?
ii. Where can the password be obtained from?

f) How long can the data be stored?

g) Who is responsible for deleting the data?

h) What is the procedure for sharing this data with people outside of your organisation?

8) Data processing and validation

a) How is the data digitalised if it is collected using a paper questionnaire?

b) Was a procedure established for data inputting?

c) Were data entry staff carefully trained? If yes,

i. What training did they receive?

d) Is data validation[2] implemented? If yes,

i. Is it implemented manually?
ii. Is it implemented through computer-based controls?
iii. Can it be part of the data collection process?

e) What procedures are in place for identifying and dealing with missing data?

9) Notes

[2] Data validation is the process of checking the accuracy and quality of source data before using, importing or processing data.