Data management plan

Existing Data. The current project will extend the existing z-proso study (http://www.cru.ethz.ch/en/projects/z-proso.html) which, by the end of the current project, will include ten waves of survey data. It will also (depending on the success of other funding applications) potentially include brain imaging, gene expression and other biological data (such as substance use levels derived from hair samples) on a subset of 200 participants. Whilst z-proso has rich data over longer time lags, it does not contain any data at the 'day-to-day' level to inform on momentary processes. Combining the newly collected experience sampling data described in the current proposal with the ten waves of existing longitudinal data will be critical to answering key outstanding questions relating to how criminal and aggressive behaviour develops over momentary and developmental timescales. To the best of our knowledge, there are no existing datasets that combine longitudinal data with experience sampling data for the study of crime and aggression. The principal investigators of z-proso have agreed that for our research we can combine the 10 waves of z-proso data with our experience sampling data (see *Data access, sharing and re-use agreement*). Investigators on the current project will have priority on and responsibility for research projects using the experience sampling data. The newly collected data can be easily integrated into existing z-proso datasets and its collection, processing and storage implemented in accordance with existing z-proso protocols. These data management protocols have ensured the security and fidelity of the existing data across z-proso's lifetime. The use of this existing data source and its integration with the new data from the current project will present no additional difficulties.

New Data. The current project will generate ~3360 new variables yielding a dataset with ~1000 rows by 3361 columns (including participant unique identifiers). The majority of items will use a 5-point Likert scale response format, resulting in ordered-categorical data. The data will be in the form of a comma-delimited ('.csv') file. As the data dimensions are relatively low, no new specialist processing or storage will be required; the data can be accommodated within existing z-proso IT systems at the Swiss Federal Institute of Technology Zurich (*ETH*), Switzerland with the existing and other to-be-collected data. Data will be collected via participants' smartphones using an application provided by *LifeData LLC*. Data is then transmitted to and securely stored on a *LifeData LLC* server during data collection period. As detailed in their *Privacy Policy*: https://www.lifedatacorp.com/lifedata-privacy-policy/, *LifeData LLC* will treat this data as confidential and safeguard it with security protections and precautions.

We will download data from these servers in comma-delimited ('.csv') files. Question numbers will form the column headings and columns are populated with participant responses. At the point of download, personal data will be deleted in order to anonymise the dataset. Data will be downloaded at regular intervals during the data collection period to create back-up copies in the event of software failures. These back-ups will be given names such as 'z-proso_ESM_data_BU_DATE.csv' to differentiate them from the final datasets where 'DATE' indicates the date of data download.

The full experience sampling datasets for each burst will be downloaded at the end of the two data collection periods. In these, column headings will be re-labelled manually by the research assistant and checked by the PI. Column headings will distinguish the same variable measured in the first and second experience sampling bursts using the suffixes '_ES1' and '_ES2'. Responses will be stored in character form e.g. 'strongly agree' with numerical recoding performed at the point of data analysis. Once processed, the new data (the '*ESM data*') will be stored as datafiles independent of the larger z-proso dataset containing the pre-existing data under filenames such as 'z-proso_ESM_data_master_VX.csv' where 'master' is used to distinguish from other versions of the datasets that will be kept for quality control and back-up purposes. The *ESM data* or specific variables from it will be merged with other variables from the main dataset via unique participant identifiers contained in all z-proso datasets at the point of data analysis and sharing.

Meta-data will be created and maintained in accordance with the *UK Data Service* guidance. Basic variable information will be stored in the *ESM data dictionary* in the form of a comma-delimited ('.csv') file that includes: variable labels (corresponding to the labels in the dataset); items as presented to participants (in German); English translations of items; coding information (e.g. '999'= missing); variable classes (e.g. 'nominal'); and brief explanations of variables. Explanations will include information such as the scale to which an item belongs (plus citation) or, in the case of summed scores or other derived variables, explanations of how they were derived. In addition, details on data collection, cleaning, coding, quality and version control procedures will be provided in a *ESM data protocol* document. This document will also record the version history of the *ESM data*. Finally, .doc and .pdf

copies of the questionnaires as they were administered to participants (*ESM questionnaire ES1 and ESM questionnaire ES2*) will be stored with the data.

Datasets will be named in accordance with the following convention, with filenames taking the form: z-proso_ESM_data_type_VX.csv'. 'Z-proso' identifies the project to which the data belong and 'ESM' identifies the sub-project. This is desirable because z-proso includes a number of sub-projects which are- due to the collective volume of data involved- stored separately from one another and linked only as required. The 'data' term identifies the type of file (to distinguish it from other types such as draft manuscripts, meta-data, policy documents etc.). The 'type' term will vary across datasets in order to distinguish temporary back-up copies created during data collection ('BU'); back-up copies stored long-term for quality control and back-up purposes ('QC'); data variants created for the purposes of sharing for specific projects ('SUB_XXX'), where 'XXX' identifies a short project code recorded in the *ESM Data Protocol*; and the master dataset ('master'). The 'VX' term identifies the file version. Major versions will be given names such as 'V1, V2, V3' etc.; minor versions names such as 'V1.1', 'V1.2' etc.

Quality Assurance. ***Pilot studies.*** Quality assurance will begin with pilot studies conducted before the start date of the proposed research. These will test the items, data collection method and the data management protocol, allowing us to optimise our methodology for the main study. Data will be collected on the acceptability, reliability and validity of items, response rates and times, and participant experiences. At time of writing, one small pilot study with n=20 participants has been completed and a larger n=200 study is underway. These data will be treated as independent of the data created in the current project and managed according to a separate data management plan.

Manual data entry and coding. The only manual data entry will be re-labelling dataset columns to ensure that labels are intuitive, brief and include no characters that would create difficulties when the data is imported into data analysis programmes. Re-labelling will be completed by the research assistant and checked by the PI.

Data checking. Ten per cent of responses will be selected and manually checked against the data held on the application server by the research assistant to ensure no errors have been introduced in data conversion, download, and column re-labelling. Data will be screened for respondents with large numbers of missing responses, out of range values, and random responding or responding according to response sets. This will use simple functions written and implemented by the PI in *R Statistical Software* and supplemented by manual checking by the research assistant. *a priori* protocols for dealing with suspect values will be written into the *ESM data protocol*. As the newly produced data will be integrated with the existing z-proso data, the same conventions as have been used in the existing datasets will be used to code for missing data, anomalous responses etc. No missing data will be imputed; all missing data treatment will occur at the point of data analysis.

Data maintenance. One master copy of the *ESM data* will be kept on the z-proso servers at ETH. Editing rights will be restricted to the research assistant, the PI and the z-proso main study PIs with all changes to be authorised by the PI. Overall responsibility for this dataset will be with the PI. Changes will be documented in the *ESM data protocol* and corresponding updates made to the data dictionary. The consistency of these files will be checked at six-month intervals. Previous versions will be archived on the z-proso servers for reference and as a back-up in case errors are introduced into later datasets. Old versions of *ESM data* will be discarded in accordance with existing z-proso protocols and documented in the *ESM data protocol*.

Security and Backup. *ESM data* will be stored securely in accordance with existing z-proso protocols. Data will be stored on password protected computers located at the Swiss Federal Institute for Technology (ETH) along with the other z-proso scientific datasets and the database containing participant personal details. These PCs are in locked offices in secure University buildings at ETH. The data files will be password protected. The separate files containing participant personal information and the *ESM data* will be linkable to the participants' data via unique identifiers. The *ESM data* files will be anonymised, with only unique identifiers given. Only the PI, research assistant and z-proso PIs will have access to the personal information files. Access to the anonymised scientific data will follow the procedures already in place in z-proso; namely the completion of a confidentiality agreement and project proposal to be approved by a project PI. Data will be backed up by saving to an external drive (encrypted) systematically every 2 weeks, after any alterations are made to any files, and after any new data is downloaded. Data on the ETH servers are also backed-up automatically at regular intervals.

Data Sharing – Issues and Solutions. Prior to data deposit, data access and sharing will follow existing z-proso protocols. Files will be shared securely through direct download from the secure z-proso

server accessed via a username and password or shared as password protected files on a USB or by email. At the end of the grant period or on publication (whichever is earlier), the *ESM Data* and associated meta-data and documentation will be prepared for deposit in the UK Data Service data repository to provide long term open access to the data. This will be completed by the research assistant under the supervision of the PI. GPS data will be deposited only in summary form because individual GPS data could risk identifying participants.

Consent and Anonymity. Expressed written consent will be collected for all participants. The request for consent will make clear that their data may be used and shared with other researchers in anonymised form but that their individual data will not be identifiable in any outputs. All participants will have unique identifiers assigned to them. Only this identifier will appear in the datasets. A separate file with personal details will be kept secure. These files will not be shared with other researchers.

Copyright and Intellectual Property Ownership. Copyright and intellectual property will be held by the PI of the current project.

Responsibilities. The PI of the current project will have overall responsibility for ensuring the integrity and security of data. The research assistant will have responsibility for the day to day management and upkeep of the datasets, meta-data and supporting documentation production. This will be supervised by the PI under the guidance of the mentor and z-proso project PIs (also see *Staff Duties*).

Appendix: Data access, sharing, and reuse agreement

The project D2M will generate two bursts of experience sampling data '*D2M data'* that can be combined with existing and to-be-collected z-proso data '*main data'* to form '*combined data'.* In all cases, the data excludes any identifying participant information. The arrangements for data reuse, access and sharing of these three datasets will be as follows:

- The D2M and z-proso study principal investigators agree to combining the *D2M* and *main data* to form the *combined data*.
- The *combined data* can be used and accessed by the principal investigators of the D2M and z-proso main study without restriction.
- The *D2M* and *combined data* can be shared with other members of the z-proso team and external researchers in accordance with the existing z-proso protocols.
- The *D2M data* will automatically be subject to existing z-proso data management, access and sharing protocols unless otherwise stated in the D2M *Data Management Plan*
- The *D2M data* will be made publicly available on publication or on the end of the D2M grant funding period by deposit in the *UK Data Service* repository.
- An exception is the GPS data from D2M which can only be shared with external researchers in summary form due to risk of de-anonymization.
- The *main data* may become publicly available at a later date, in which case the full *combined data* will be publicly available.

9th July 2017

Signed **Date**

8 July 2017

Signed **Date**