Confidence Database contributor instructions

What is the Confidence Database?

The Confidence Database is a collection of datasets from a variety of studies employing subjective ratings. The goal is to enable future model development and discoveries by making a large amount of data available in a standardized format.

Why contribute?

There are at least 2 reasons for contributing your dataset:

1. Your data could support and become the basis for new discoveries and model development that require the availability of large amounts of data
2. If your dataset is published, you'd be increasing the visibility of your paper and may receive more citations than otherwise

What studies qualify for inclusion in the database?

Any study that collected confidence (or visibility/wagering/difficulty/etc) ratings is eligible. Any dataset with confidence is welcome as long as it can be formatted in the structure described below.

Format of the data

All data should be submitted in .csv format such that each column presents a different feature of the data. All experiments should have the following columns:

- Subj_idx (the subject index, preferably in 1:N format; do not use participant initials)
- Stimulus (this could be numeric or string; for 2AFC designs, it may be necessary to include two fields [e.g., Stim1 and Stim2] corresponding to the two stimuli presented)
- Response (this should have the same format - numeric or string - as the Stimulus field)
- Confidence
- RT_dec (RT of decision in seconds; if decision and confidence are given at the same time, this should instead be named RT_decConf)
- RT_conf (RT of the confidence in seconds; this should be the time taken after the decision is made in experiments where decision and confidence were given separately; if RT for confidence was not recorded, do not include this column but make a note about it in the readme file)

In addition, if applicable, submissions should have the following extra columns:

- Contrast/coherence/noise level/difficulty (an indication of the difficulty in the given trial; this could be on an abstract scale [e.g. 1-3 if there are 3 contrast levels] or the raw contrast/coherence/etc can be reported)
- Condition (if more than one condition is present [e.g., TMS vs. no-TMS], then the condition can be indicated here)
- Accuracy (this can usually be inferred from the Stimulus and Response columns but not always)
- Training (if training data are included, then this field can be used to indicate whether each trial is from the training or the real experiment; if this field is not included, then it is assumed that none of the trials reported are training trials)
- Stim alternatives (if different stimulus alternatives are present on different trials)

If your experiment requires additional columns to fully describe the dataset, feel free to add custom columns.

Importantly, all raw data should be included exactly in the order in which it was obtained. For example, even if a subject didn't respond on time on a given trial, this trial should still be included in the dataset (this could be important, for example, in analyses on serial dependence). Missing values should be coded with “NaN”.

Datasets should be as complete as possible. Even if you excluded subjects for bad performance or other reasons, all subjects who completed the experiment should be included in the dataset (exceptions could be entertained but these should be rare; such exceptions should be explicitly described in the accompanying readme.txt file described below).

If subjects completed several different tasks, include all data in a single file and add a column “Task” that specifies which task was performed in a given trial.

Accompanying readme.txt file
Each dataset should be accompanied by a readme.txt file. This file should give some basic details regarding the experiment. Contributors are welcome to make this file as detailed as they see fit. However, at a minimum, the file should contain the following information:

- Contributor(s): Full name and email of contributor(s).
- Citation: A full citation of the published paper and indication as to which experiment in the paper the dataset comes from (for cases of multi-experiment papers). If there is no associated paper (either a published paper or a preprint), this should still be indicated here.
- Stimulus: A short description of the stimulus categories. If numerical coding is used in the .csv file, then the corresponding categories for stimulus and response should be indicated here.
- Confidence scale: An indication of the confidence scale used, as well as the type of confidence collected (confidence rating, wager, visibility, etc.) Cases where confidence goes from “certainly wrong” to “certainly correct” (rather than the more usually “guess” to “certain”) should be specifically highlighted. If the different options were named (e.g., “not all confident”, “somewhat confident”, “very confident”), this should be indicated here too.
- Manipulations: Full description of any manipulations used (or indication that there were no manipulations).
- Block size: The block size and number of blocks completed by each subject. If blocks are of variable size, then the dataset should include an extra column “Trial_in_block” which can be used to unambiguously identify when new blocks start and how long each block is.
- Feedback: was trial-by-trial feedback included or not.

In addition, here are some ideas about additional (optional) information that could be indicated (it would be useful if most submissions indicate most of these):

- NaN fields: Explanation as to where NaN fields come from (e.g., there was a deadline on the response and so no answer was recorded)
- Subject population: give an age range and perhaps mean and SD. If a dataset contains different populations (e.g., young/old, disease/control), a list of the subjects in each group can be provided here (this can also be indicated in a separate column in the csv file). In addition, sex/age of each participant could be provided separately.
- Response device: most studies will have “keyboard” but others possibilities include mouse, joystick, touchscreen, button box, eye tracker, etc.
- Experiment setting: e.g., in lab/online, individual booth vs. social setting, etc.
- Training: amount of training given prior to the experiment
- Experiment goal: original goal of experiment
- Main result: main finding
- Special instructions: special instructions given to the subjects (e.g., was there any instruction about the speed of responses, were subjects encouraged to use the whole confidence scale, etc)

- Link to material/codes: If the material/codes for experiment are available anywhere, a link can be provided here
- Experiment dates: Dates when the data were collected
- Location of data collection: lab, university, city, and country where the data were collected (could also include language in which the experiment was conducted)
- Any other important information about the dataset

Basic dataset information

To keep track of the different datasets, basic information for each dataset needs to be provided. The information will be added to the master spreadsheet titled “Database_Information.xlsx”. To provide the needed information, use the file “Spreadsheet_template.xlsx”. Make sure that everything you include in this file matches the csv files you’re sending. In the field “Num_tasks_x_conditions” indicate the number of different tasks, conditions, or task X conditions (if you have both) in your experiment. In general, different difficulty levels and different subject populations are not considered different tasks/conditions. If you’re unsure how to fill any of the fields, check out what was done for some of the other datasets in the Confidence Database.

Naming the files

The files should be named “data_FirstAuthorLastName_year.csv” and “readme_FirstAuthorLastName_year.txt”, where the year [in YYYY format] indicates the year of the corresponding publication or preprint. For unpublished datasets, use “unpub” instead of the year.

Before you send your contribution

Over half of the contributions that we receive do not meet all requirements listed above. This places a large burden on the moderators of the database and often results in long back-and-forth email chains. Please, double- and triple-check every detail of your submission before sending it. Open all of your files again, re-read the readme ones and scroll through the csv files to make sure that everything looks good. In addition, check for these common issues:

- The data in the csv file should not appear in a single column. This happens sometimes due to incompatibility between different software packages or operating systems (and is often country-dependent). Make sure that your csv files always open with data in separate columns.
- The names of the required columns should be *exactly* as listed (e.g., don’t use “Subject” instead of “Subj_idx” or “stim” instead of “Stimulus”).
- The order of the columns should be as listed (the first few columns should be “Subj_idx”, “Stimulus”, “Response”, “Confidence”, “RT_dec”, and “RT_conf” *in this order*)
- RT should be in seconds, not ms.
- All trials and subjects should be included. Regardless of whether you excluded any subjects or trials in your analyses, include everything in the dataset that you contribute. If anything must be omitted, make a note about it in the readme file.
- Missing data should be replaced by “NaN” (e.g., do not use “NA” instead of “NaN”).
- There should be no empty columns or rows, or individual cells anywhere.
- All data should be numbers and letters. Make sure that your submission doesn’t contain nonsensical characters (which sometimes happen during conversions between different file types).
- If you composed the readme file in Word and saved it in .txt, it will likely contain a number of nonsensical characters. Open the readme file in a text editor and edit such nonsensical characters there.
- The files should be named exactly according to the rules above. If Smith & Jones are submitting an unpublished dataset, its name should be “Smith_unpub”. Names like “SmithJones_unpub”,

“smith_unpub”, “Smith_2019”, Smith_et_al_unpub”, “Expt1_Smith_2019”, etc. are all incorrect. If, and only if, a name would overlap with another dataset already in the database (or another dataset that you’re sending), then you should add custom additional information *at the end* of the name (e.g., “Smith_unpub_Expt1” or “Smith_unpub_memory”).

Data quality checks

We have written simple scripts in Matlab, Python, and R Markdown to help with initial data quality checks. These files are named quality_check.m, quality_check.py, and quality_check.Rmd, respectively and are available on the OSF page. Run one of these files before sending your contributions (this is a REQUIREMENT for submission). To run the files:

- Edit the “Spreadsheet_template.csv” document to include the information about your dataset(s).
- Place one of the quality_check files, all files you're sending, and the Spreadsheet_template.csv document (don’t rename this file) in a new folder.
- Run the quality_check file. If there are any errors or warnings, update the files until you have addressed all issues.

Where to send files

To contribute, send an email to confidence.database@gmail.com with:

- All dataset files (in .csv format)
- All readme files (in .txt format)
- A single csv file with information about the dataset(s) you’re contributing. This information will be added to the master spreadsheet “Database_Information.csv”. Please use the template provided called “Spreadsheet_template.csv”.
- A confirmation (within the email) that you have ran one of the data quality checks and they return no errors.

Your files will undergo additional basic quality check and then be uploaded to the database (https://osf.io/s46pr/). You can check the website for examples of datasets and readme files that follow the requirements listed above.