Notebooks - Exploratory study data analysis

This folder contains the notebooks used to perform analyses of data collected in the exploratory survey—see ***section 5.4*** and ***5.5***, and ***Appendices F*** to ***N*** in the paper.

Files

Notebooks in this folder are numbered following their running order:

- `1-results data structuring.ipynb`: survey data preprocessing before running analyses (excluding data from participants based on attention checks, separating data from individual stimuli), and descriptive reports (demographics).
  - **INPUT file**:
    * `Results/results_cleaned.csv`
  - **OUTPUT files** in `Results/Data_Analysis/`:
    * (root)
    * `data/` (candidate items ratings alone)
    * `others/` This folder was not included in our OSF repository as we did not include this data directly in the paper, but the files it contains serve as input to produce the `all_comments_by_stimulus.csv` file with the next notebook.
- `2-data analysis.ipynb`: comments counts and PREVis ratings frequencies heatmap plots (this file is not mandatory to run: the outputs were not necessary to validate PREVis and we did not include them in the paper).
  - **INPUT files** from `Results/Data_Analysis/`:
    * `valid_answers-[...].csv` for [A, B, C, D, E, F]
    * `others/intial-results-[...]_others_comments.csv` for [A, B, C, D, E, F]
  - **OUTPUT files** in `Results/Data_Analysis/`:
    * (root) : comments and comments counts for each item x stimulus
    * `generatedPlots/`: heatmaps
- `3-EFA-+-all.Rmd` and `4-EFA factor loading analysis.ipynb`: Exploratory Factor Analyses (see ***section 5.5*** and ***Appendices F*** to ***J*** in the paper). We used `knitr` to produce an html output from the R notebook: `3-EFA-+-all.html`, which embeds important outputs, and notably those related to ***Appendices F*** and ***H*** in the paper.
  - **INPUT files** from `Results/Data_Analysis/`:

    * `data/ratings-[...].csv` [1 to 7]:
        · files numbered `1---6` correspond to A—F individual stimuli ratings
        · file numbered `7` corresponds to all collected ratings across stimuli (related to output files labeled `Agg`)
  - **OUTPUT files** in `Results/Data_Analysis/`:
    * `generatedPlots-EFA/`: contains files related to ***Appendix I***.
    * `generatedData-EFA/`: in particular, the `Agg/` subfolder contains the factor loadings tables we used to conduct analyses described in ***Appendix J***

- `5a-reliability.Rmd`: reliability analyses described in ***Appendix L***.
  - **INPUT file** from `Results/Data_Analysis/`:
    * `data/ratings-stimulus.csv`
    * with `full_Readability_factors` variable to filter the dataset, extracted from the 4-factors structure described in ***Appendix K*** (see also ***Fig. 28*** in the paper, and corresponding OSF file: `Phase 2 - Scale development/Figures supp/EFA_loadings-4_factors.pdf`)
  - **OUTPUT files** in `Results/Data_Analysis/`:
    * `generatedData-CFA/reliability/`

- `5b-MG-CFA-[...].Rmd` [Model1, Model_2, Model_3, Model_final]: Multi-Group Confirmatory Factor analysis for each model explored as described in ***Appendix M***.
  - **INPUT file** from `Results/Data_Analysis/`:
    * `data/ratings-stimulus.csv`
    * each notebook uses a different `this_model` variable to filter the dataset, as described in ***Appendix M***
  - **OUTPUT files** in `Results/Data_Analysis/`:
    * `generatedData-CFA/`
    * `generatedPlots-CFA/correlations/`

- `6-reliability-MG-CFA-analysis.ipynb`data pre-processing from `5-a` and `5-b` (above) notebooks outputs, to facilitate analysis described in ***Appendix M.2***.
  - **INPUT files** from `Results/Data_analysis/generatedData-CFA`:
    * `fit_indices-[...].csv` [Model1, Model_2, Model_3]
    * `/reliability/to aggregate/[...][...]-reliability.csv` [understand, layout, dataFeat, dataRead]x[Model1, Model_2, Model_3, Model_final]. *Note: we built this notebook with 3 models and edited it after selecting the final model to obtain our final instrument's reliability metrics as reported in* ***Appendix M.3***.

  - **OUTPUT files** in `Results/Data_Analysis/`:
    * `CFA/`: fit indices ranking and metrics summary outputs to compare the 3 models' goodness-of-fit performances.
    * `generatedData-CFA/reliability`: concatenated reliability metrics tables for each model (as well as summary .csv and .tex files for each).

- `7-IRT-graded.Rmd`: as stated in ***Appendix M***, we did not use IRT as part of our analysis—although we had considered it originally—as our data was not appropriate due to a low number of items in each individual factor. We ran it as a supplementary confirmation of our final choice. This notebook creates a graded response model IRT analysis and outputs Item Characteristics Curves for a 1-factor solution (which was the most likely to produce reliable results as it contained enough information with 29 items) and subscale-by-subscale in our 4-factor solution from ***Appendix J***.
- `8-rating plots.ipynb`: subscales average scores point plots with 95% CI, as PDF files.
  - **INPUT file** from `Results/Data_Analysis/`:
    * `data/ratings-stimulus.csv`
  - **OUTPUT files** in `Results/Data_Analysis/`:
    * `generatedPlots/ratings-factors` contains ***Fig. 48*** to ***Fig. 65*** in the paper: plots generated for each item in the 4-factors solution from ***Appendix J***, with the item's and the factor's average ratings and 95% CI, across stimuli.
    * `generatedPlots/ratings-PREVis`: contains ***Fig. 30*** to ***Fig. 47*** in the paper: data generated from invidiual PREVis subscales, for each stimulus, with individual items' average ratings and 95% CI.
  - **OUTPUT files** in `Results/Data_Analysis/generatedPlots/ratings`

Running notebooks

Requirements

Python code

Pyhton Jupyter notebooks require Pyhton 3+ with the following libraries (and their dependencies):

- `pandas` *version 2.0+*
- `seaborn` *version 0.13+*

R code

R Markdown files can be run with the R Studio software with the following libraries (and their dependencies):

- `dplyr`
- `knitr`
- `psych`
- `lavaan`
- `corrplot`
- `RColorBrewer`
- (`misty`) *loaded as part of original functions, but not used as there is no missing data in this survey*
- (`mice`) *loaded as part of original functions, but not used as there is no missing data in this survey*
- (`mifa`) *loaded as part of original functions, but not used as there is no missing data in this survey*

How to run notebooks

1- Prepare folders

To run these notebooks, you will need to download the `results_cleaned.csv` file from the `Results/` folder on OSF, and arrange a local folder structure as shown below:

```
/root_folder

    /Notebooks

        *[all notebooks]*

    /Results

        - results_cleaned.csv
```

2- Update R markdown files

Update the `setup` cell (first cell) in all `.Rmd` files so as to set the working directory to the absolute path of your `/root_folder`.

3- Run individual notebooks

Run notebooks in order (1 to 5): the `Results/Data_Analysis/` folder and nested structure will be generated from the code.