Interactive Intelligence
Checklist for Review of Dataset
(Version 1)

We recommend that students or employees wishing to publish on their data and results for a given research project in the form of a dataset asks a fellow student or colleague to review this dataset with regard to the points in this checklist. The purpose of the checklist is to ensure that all data that can be made available is made available, that all analyses were conducted conscientiously by the researchers, that all results are reported accurately, and that all methods are transparent and sufficiently clear to be reproducible.

If you choose to have your code reviewed according to this checklist, we advise you to upload this document together with your dataset to the research data repository of your choice (e.g. 4TU Research Data) upon publication of your work.

I. Basic Data

Paper title:	**Using Reinforcement Learning to Personalize Daily Step Goals for a Collaborative Dialogue with a Virtual Coach**
Name(s) of researcher(s):	**Martin Dierikx**
Name of the reviewer:	**Andrei Stefan**
Data repository platform (e.g. 4TU Centre for Research Data):	**4TU Centre for Research Data**

II. Checklist

Statement	**Yes**	**No**
1. The dataset contains a README file that fulfils the requirements of the data repository platform that the researcher wishes to use. If no such requirements can be found, the dataset nonetheless contains a README file that clearly explains the contents of the dataset?	X	
2. Either within the README file or within an extra, easily findable file, the researchers have explained their data. This means that, for example, for every column of a tabular dataset, all column names and possible cell values are explained.	X	
3. All data is in readily readable file formats. If this should not be the case, the README (or similar) clearly explains the file format and which software can be used to access the contents.	X	
4. All data has been anonymized in accordance to promises made in the Data Management Plan.	X	
5. The analysis file or files contain a header with meta-data (name of author, date of writing, required input files and generated output files).	X	
6. All required input files for the analysis are available in the dataset.	X	

Statement	Yes	No
7. There is an output file that is generated by the analysis script that neatly combines code and commentary (e.g. markdown output file). This output file is in a readily readable file format (e.g. pdf).	X	
8. The analysis script is clean and comprehensible in the sense that: • There is sufficient, useful, and clearly written commentary • Irrelevant code (such as old analyses) has been removed • The details of analyses that are not reported in the paper (e.g. assumption checks) are proportional to those that are reported in the paper. This means that unreported analyses should not clutter up the script, making it long and unreadable.	X	
9. The analysis script can be run successfully.	X	
10. All preprocessing steps are clearly described and traceable, especially when preprocessing code cannot be executed because raw data is not available.	X	
11. The analyses and results reported in the manuscript can be found back in the analysis script with labels according to where they appear in the manuscript.	X	
12. All results reported in the manuscript accurately correspond to the output produced by the analysis script.	X	

III. Additional comments by reviewer

Please state any additional things you noticed in reviewing the dataset or possible points of improvement for the reviewer.

Comments Round 1:

data README

preprocessed_pre_screening_data.csv
- missing explanation for Steps_on_average (so item 2 in the checklist is a "No")
- might be good to add a reference rather than just saying "the questions are given in the paper from Godin"

main README
- Database data preprocessing.ipynb | Code **fo r** reproducing the preprocessing of the database data file. - typo
- multiple places in the file - Python, not python
- https://data.4tu.nl/s/documents/Guidelines_for_creating_a_README_file.pdf - missing title of the dataset, contact info

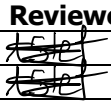
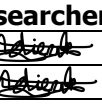
Analysis files:
- could specify when there is no output file in the header, just for clarity
- Database data preprocessing - preprocessed_database_data_reproduced is not a csv, so comparing with the existing csv file is tedious
- Demographic results
  - Table 4.1 – missing print statements for the other means of recording steps
- G algorithm results
  - Also, the **plitting** of the data could create a warning running the anova - typo
  - Also after separating the data over multiple **bucket** to compare them, it could be that one of the buckets has too few samples in it which creates a warning running the anova. This warning will not impact the results, however. - typo
  - Not clear which feature corresponds to which number
  - Very difficult to translate the G-algorithm runs into the table
  - Unclear which entry corresponds to which iteration of the G-algoritm

- Table hides some of the results by only specifying that a feature is significant in combintaion with other features (no way of knowing which the other features are without running the code)
    - "Feature 2" is both the context state and the rest (which is part of context state)
    - To see why feature 2 (rest) and feature 3 (available time) have a Yes** in the table, modifying the code is necessary, so it's not immediately clear why the reported results are the way they are
- General analysis results
    - plt.ylabel("Number of **occurances**") – typo (also in many other places)
    - 18% - it is ambiguous since it doesn't say 18% of what
    - state_action_pair_**encounteres** – typo (also in other files)
    - Distribution of the rewards of samples, Figure 4.6 and Appendix E: needs more comments, not immediately clear what the for loops do (so item 8 in the checklist is a "No")

Comments Round 2:

Most of the previous comments are implemented. The comments which made me answer "No" to some of the checklist items have been resolved. The remaining ones are minor things which could be better (e.g. the G-algorithm outputs) but the code is explained well enough to reproduce the results in the manuscript, so I don't think it's worth taking another round for changing those.

IV. Review log

Round	Date	Paper Status	Checklist Items	Signature Reviewer	Signature Researcher
1	25/09/2023	Pre-submission	1-12		
2	26/09/2023	Pre-submission	1-12		