ATELIER GR NUMÉRIQUE, ARDAA

A shared research space: Using R to share data analysis procedures in Second Language Acquisition and Didactics

Marie-Pierre Jouannaud, Laboratoire Transcrit, Université Paris 8

Introduction

Scientific research is a collaborative enterprise: science rests on the sharing of observations, ideas and methods, and on debate among competing hypotheses and interpretation of results. Traditionally, this played out in scientific journals and society meetings and conferences, or in private correspondence between researchers. More recently, the advent of the internet has opened up new spaces for collaboration, and a number of academic journals have gone "open access". Awareness has grown, however, that sharing ideas and results may not be enough. In the context of the replicability crisis in social sciences (Open science collaboration, 2015), it is also important to share data (e.g. CHILDES) and data collection materials such as questionnaires, tests, etc (Marsden et al., 2015).

In this paper, I describe the advantages of using the statistical software R (R Core Team, 2022) to go one step further and share data analysis procedures together with collected data. R is a free, open source environment for statistics and graphics. Because analyses are run using lines of code that can easily be shared with other researchers and run again using the same or similar data, it is easy to collaborate with colleagues who can reproduce analyses and graphics, point out flaws, and perhaps tweak them to improve them. Although R requires initial training (as any statistical software), there are books, MOOCs and videos dedicated to its use, as well as a vibrant support community that can be relied upon when encountering difficulties.

1. Background: The Open Science movement

At the end of the 1990s, scientists started to rebel against the power of the major scientific publishers, whose publishing costs decreased in the Internet era and yet charged steadily rising fees which did not seem justified (Larivière et al., 2015). Researchers did the actual work of writing and reviewing the articles, but publishing companies forced them to sign

away their rights, and charged university libraries the world over for access to the scientific output of their own employees. Most articles were behind paywalls and thus difficult to access for the general public (who had, albeit indirectly, paid for the research), or for researchers with little or no resources or institutional support. This state of affairs inspired reactions at different levels:

- Some researchers went 'rogue' and called for rebellion, by illegal means if necessary (Swartz, 2008): the "Sci-Hub" repository is a child of this movement (Siew, 2017);
- Some journals decided to go "open access": their articles are available without cost on the web (see Al-Hoorie, 2021 for a list of open access journals in linguistics);
- Governments have also adopted resolutions (at the Europen level) or laws (at the national level) pushing for government-funded research to be available to the public (Enserick, 2016). In France the HAL platform is conceived as an open repository of each public researcher's published output, after a short embargo or immediately upon publication if the CC-BY licence is used (MESR, 2022).

However, Marsden (2020) argues that methodological transparency (under the banner of open science) goes much further than open access and the free and permanent availability of research output to the general public. It is also a philosophy that has consequences at all stages of the research process:

- First, when designing research protocols and before data collection, permission must be sought to make (anonymized) data publicly available; provision must be made for public storage of all instruments designed and used, such as questionnaires, pre- and post-tests and experimental protocols. The Iris database (Marsden et al., 2015) is a natural repository for such documents, as it was designed specifically to facilitate sharing and collaboration between SLA/ AL researchers and is free and open to the public.
- Second, knowing that our instruments will be publicly available and open to criticism by the community has a positive washback effect on instrument design, Marsden (2020) argues. If we want other teams to re-use our instruments or replicate our study, care must be taken to make sure their usage is self-explanatory (otherwise, they need to be accompanied by a tutorial, to describe the scoring scheme, for example). The washback effect also works on instrument quality (better proofreading, better quality items through prior piloting), as has been shown to happen in other related domains (Wicherts et al., 2011).

- Finally, after data collection, all the steps taken to 'clean' the data before analysis must be described and justified (for example, stating whether outliers, or incomplete data from some participants, were removed, and on what grounds). When reporting the results, visual means of presentation are important (Larson-Hall, 2017), but detailed descriptive statistics (means, standard deviation, skew, reliability of instruments, etc) must also be provided to enable reanalysis, future comparison with other data sets or meta-analyses.

  As we will now see, the R software is ideally suited for this.

2. The R software

Applied linguistics and SLA researchers often need to use quantitative analyses to summarize their observations, describe their results and answer their research questions. However, most of us come from the humanities and are not entirely comfortable with the use of statistics. Although statistics are ubiquitous in our daily environment (the news especially), and are required to understand a majority of Second Language Acquisition research articles, training in statistical reasoning is lacking. Several studies have pointed out that statistical literacy is not high, and that many researchers feel underprepared for the comprehension and use of statistics (Loewen et al., 2014). Nevertheless, the same studies show that most researchers know how to use basic descriptive statistics (mean, median, and standard deviation), are familiar with a few language testing concepts (validity and reliability) and can interpret some inferential statistics (p-value, t-tests, ANOVA, effect sizes, etc.).

In order to conduct quantitative analyses, a majority of applied linguists use the SPSS software, and a small minority (15 to 20%) use R (Loewen et al., 2014; 2020). R is a statistical environment for statistical computing and graphics, and it has several advantages. It is free, so that there are no financial barriers to entry. It is also open source: statisticians and computer scientists can (and do) add extensions to the core functions in R. Because of this, the range of possible analyses and graphics is virtually unlimited and very flexible.

Reproducibility of data analysis, however, is "perhaps the most compelling advantage provided by R" (Mizumoto & Plonsky, 2016). R makes it easy to reproduce analyses because each action it performs corresponds to a line of code (one or several functions) that is controlled by the user. Each line of code can be commented so that users can explain to someone else using the code (or to themselves at a later date) what it does and why. For example, there is a function to import files in table format (.csv), functions to add or delete rows or columns, or to perform basic operations (means, medians, etc). Other functions allow users to produce

graphics. In addition, libraries of functions called “packages” automate additional operations or offer more elaborate graphics. Analyses are performed using successive function calls stored in an R “script” (a sequence of commented functions) that can then be run again as is, by the same or a different person, without the need to go through each step anew.

By contrast, in less open software such as SPSS or even Excel, the sequence of actions necessary to perform an analysis with a given data file is not memorized automatically, and the steps taken to clean the files before analysis are not transparent. Previous research has shown that this transparency is essential, because in social sciences, dealing with human subjects vastly increases the complexity of analyses needed to answer the questions we ask. We need to admit that we are not certain what the best methods are, and that using a different method might yield different results (Silberzahn et al., 2018). Even using the same method but with different decisions at each step of the process (selection of participants, etc.) can greatly alter the conclusions reached (Simmons et al., 2011). Being open about these decisions is the easiest way to mitigate this unavoidable problem, and can hopefully inspire others to try to replicate results to arrive at a clearer picture in the end.

3. Luciole, a research project using R

Luciole is one of the applications developed in the Fluence project (Mandin et al., 2021). It is a serious game in which children play the role of a young secret agent whose mission is to find and free kidnapped animals. Its aim is to develop listening comprehension skills in English as a foreign language during the first years of elementary school in France. Two other applications were developed in Fluence: EVAsion and ELARGIR train the cognitive mechanisms (visual and visuo-attentional processing) and holistic processing (orthographic units, prosody, breath groups) inherent to reading, with the goal of improving students' reading fluency in L1 French.

The project is a longitudinal study following a cohort of more than 500 students from first to third grade (2018-2020). The three applications are used symmetrically. Luciole (English listening comprehension) serves as a control for EVAsion and ELARGIR (L1 reading), while they in turn act as control groups for Luciole. The effect of each application is measured with pre- and post-tests administered at the beginning of the first year and then at the end of each school year. Because of the scale of the project and thanks to the funding obtained (e-FRAN, PIA 2), a post-doc researcher was hired to help with statistics. For the pre-test analyses, this post-doc shared the R files with us and walked us through them. In this way, we were able to replicate the initial analysis on our own time and understand the effect of each line

of code (to which we added our own comments). Thanks to this initial training, we largely managed to conduct the next iteration of analyses independently. Below is an example of the code produced for the cleaning of the data file before running any statistical analysis. The lines beginning with a hashtag correspond to comments, and the other are lines of code, with 'foo' being the new name of the data file uploaded to R, 'read.csv' the function used to upload it, and 'filter' the function used to keep participant data (rows in the data file). In this way, the fact that we did not use data from a bilingual class, or from the students whose parents did not sign the consent form or who were absent for a test is stated explicitly.

```
#import data file
foo <- read.csv ("PrePo_CP_Luc_Scores_data.csv", sep = ";", na.strings=c("","NA"))

#enlever 22 eleves classe bilingue (classe num 32)
foo <- filter (foo, Classe!=32)

#enlever 43 non consentement
foo <- filter (foo, consent==1)

#enlever 30 absents prétest ou posttest
foo <- filter(foo, abs_pretest==0)
foo <- filter(foo, abs_posttest==0)

#il reste 520 enfants
```

Each decision to remove data or participants is completely transparent, and any user running these lines again will see the effect of these decisions immediately. It is also easy to tweak the code and observe the potential effects of making different decisions on the end result. This transparency makes it difficult to "doctor" the data or the analysis in such a way that they correspond to the hoped-for results.

After the initial data cleaning comes the analysis and the visualization of results. R is known for the quality of its graphics, and the range of data visualization it allows. Figure 1 shows different ways of presenting the same results depending on the level of detail required, comparing the effect of Luciole training on listening comprehension scores after 10 hours playing on the application, and showing a significant effect.

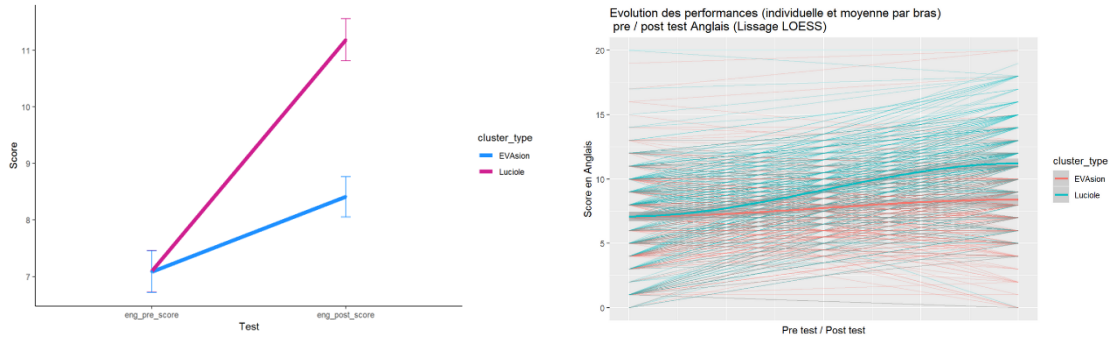

Figure 1 – Two visualizations of the evolution of English listening score at the beginning and end of CP (first grade) for Luciole and EVAsion groups

Both graphs in Figure 1 represent the evolution from pre- to post-test, but the line graph on the right (with one line per participant) also allows the reader to visualize the amount of individual variation among both groups, in terms of both initial level and learning trajectory. It is a much more 'data accountable' graph in Larson-Hall (2017)'s terms, because it attempts to convey all the information available (in a hopefully still understandable way).

4. Perspectives

The aim of the Open Science movement is to produce transparent and trustworthy research whose results are freely available to everyone, and to encourage collaboration among researchers. In this paper, I have argued that the *R* software is a useful tool to partially fulfil these aims and open up new spaces for collaboration, because it helps with the sharing of data analysis methods, including pre-treatment of data files and production of nice-looking data-accountable graphics.

It must be said, however, that while most of us share the lofty goals of open science presented here, the reality of conducting research is sometimes far removed from them. *R* does have a steep learning curve (mainly because we usually need to learn both the statistical methods and the code for them at the same time), but learning how to use it can also be done collaboratively, with colleagues working on the same project, with the help of user websites, or with one of several available MOOCs (e.g., Falissard & Lalanne, 2017).

Bibliography

Al-Hoorie, A. (2022). *Applied Linguistics Open Access Journals*. Ali H. Al-Hoorie. https://www.ali-alhoorie.com/applied-linguistics-open-access-journals

Enserink, M. (2016). In dramatic statement, European leaders call for 'immediate' open access to all scientific papers by 2020. *Science*. https://doi.org/10.1126/science.aag0577

Falissard, B., & Lalanne, C. (2017). *Introduction à la statistique avec R* [Cours en ligne]. FUN MOOC. http://www.fun-mooc.fr/fr/cours/introduction-a-la-statistique-avec-r/

Larivière, V., Haustein, S., & Mongeon, P. (2015). The Oligopoly of Academic Publishers in the Digital Era. *PLOS ONE*, *10*(6), e0127502. https://doi.org/10.1371/journal.pone.0127502

Larson–Hall, J. (2017). Moving Beyond the Bar Plot and the Line Graph to Create Informative and Attractive Graphics. *The Modern Language Journal*, *101*(1), 244-270. https://doi.org/10.1111/modl.12386

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical Literacy Among Applied Linguists and Second Language Acquisition Researchers. *TESOL Quarterly*, *48*(2), 360-388. https://doi.org/10.1002/tesq.128

Loewen, S., Gönülal, T., Isbell, D. R., Ballard, L., Crowther, D., Lim, J., Maloney, J., & Tigchelaar, M. (2020). How Knowledgeable are Applied Linguistics and SLA Researchers about Basic Statistics? : Data from North America and Europe. *Studies in Second Language Acquisition*, *42*(4), 871-890. https://doi.org/10.1017/S0272263119000548

Mandin, S., Zaher, A., Meyer, S., Loiseau, M., Bailly, G., Payre-Ficout, C., Diard, J., Valdois, S., Blavot, A., Bosse, M.-L., Briswalter, Y., Chalon, N., Godde, E., Ingremeau, S., Jouannaud, M.-P., Lequette, C., Magnat, E., Masperi, M., Piat-Marchand, A.-L., … Zanoni, M. (2021). Expérimentation à grande échelle d'applications pour tablettes pour favoriser l'apprentissage de la lecture et de l'anglais oral. *EIAH 2021 - 10e Conférence sur les Environnements Informatiques pour l'Apprentissage Humain*, 118-129. https://hal.archives-ouvertes.fr/hal-03292798

Marsden, E. (2020). Methodological transparency and its consequences for the quality and scope of research. In J. McKinley & K. Rose, *The Routledge Handbook of Research Methods in Applied Linguistics* (p. 15-28). Routledge. https://doi.org/10.4324/9780367824471-2

Marsden, E., Mackey, A., & Plonsky, L. (2015). The IRIS Repository : Advancing Research Practice and Methodology. In A. Mackey & E. Marsden, *Advancing Methodology and Practice* (p. 1-21). Routledge.

Ministère de l'Enseignement supérieur et de la Recherche. (2022). *Mettre en œuvre la stratégie de non-cession des droits sur les publications scientifiques* (Ouvrir la Science). https://www.ouvrirlascience.fr/mettre-en-oeuvre-la-strategie-de-non-cession-des-droits-sur-les-publications-scientifiques

Mizumoto, A., & Plonsky, L. (2016). R as a Lingua Franca : Advantages of Using R for Quantitative Research in Applied Linguistics. *Applied Linguistics*, *37*(2), 284-291. https://doi.org/10.1093/applin/amv025

Morgan-Short, K., Marsden, E., Heil, J., Issa II, B. I., Leow, R. P., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018). Multisite Replication in Second Language Acquisition Research : Attention to Form During Listening and Reading Comprehension. *Language Learning*, *68*(2), 392-437. https://doi.org/10.1111/lang.12292

OPEN SCIENCE COLLABORATION. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Siew, K. (2017). The open science movement. *The Physiological Society*. https://www.physoc.org/magazine-articles/the-open-science-movement-revolution-is-underway/

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many Analysts, One Data Set : Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337-356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology : Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632