

Dear Professor Di Cook,

In *A Study in Reproducibility: The Congruent Matching Cells Algorithm and cmcR package*, we describe the process of implementing an open-source version of an algorithm used for cartridge case comparisons called the Congruent Matching Cells (CMC) method. Our paper discusses the usage of the `cmcR` implementation, but this is ancillary to the paper's narrative: instead, we focus on the challenges of implementing an algorithm which was released without any code, intermediate data, or even pseudocode.

We are highlighting the shortcomings of a purely verbal description of an algorithm - a discussion that does not include pseudocode, function references, or any discussion of parameter settings. Our discussion of this process centers on the required components for true computational reproducibility: that both source code and intermediate data are necessary to assure that the new implementation is qualitatively similar to the implementation described in the paper.

Because we did not have these luxuries, we conducted a limited grid search through the different parameter settings and algorithm options, and developed a method to quantify the differences between the results of each combination of settings and procedures. This required modularizing the algorithm (which is of course good programming practice in any case) and examining many of the different ways any particular description could be interpreted. As a result, we think this paper extends well beyond the `cmcR` package vignette, and in fact contains useful observations for anyone interested in computational reproducibility or anyone who is operating in fields where open-source software is not an expectation.

Thank you for your consideration,

Joseph Zemmels, Susan Vanderplas, and Heike Hofmann

Below is an explanation of non-standard files/folders included with the submission.

- data folder: contains data to demonstrate the `cmcR` package and reproduce various results discussed in the paper. A description of each file in this folder follows.
 - fadul1-1.x3p, fadul1-2.x3p, fadu2-1.x3p: Two same-source and one different-source cartridge case scans. They are used to demonstrate usage of the `cmcR` package.
 - `cmcCountData.Rdata` file: contains CMC count data from the Fadul et al. (2011) set of 40 cartridge case scans. These data are used to generate Figures 12 and 13.
 - `cmcCountData_script.R` file: a script to reproduce the `cmcCountData.Rdata` file
- images folder: contains manually-created images and photos used in the paper
- figures folder: contains automatically-generated images (while running `zemmels-vanderplas-hofmann.R`) used in the paper