August 20th, 2021

Dr. Brendan Johns

Associate Editor

*Behavior Research Methods*

Dear Dr. Johns,

Our revision of BR-Org-21-206 “The lrd Package: An R Package and Shiny Application for Processing Lexical Data” has been resubmitted for your consideration. We appreciate the thorough examination and constructive feedback provided by each of our reviewers and are encouraged that *lrd* was viewed as “an excellent tool” and “useful to researchers.” Below, we list each reviewer’s comments and provide our responses. We cite page numbers for sections in which specific revisions have been made and note that modifications to the manuscript have been made in blue-colored font to facilitate review. We look forward to hearing from you regarding our revision and hope that it is now suitable for publication in *Behavior Research Methods*.

Sincerely,

Nicholas P. Maxwell

School of Psychology

The University of Southern Mississippi

Email: mark.huff@usm.edu

Phone: (601) 266-5411

Fax: (601) 266-5580

Cc:

Mark J. Huff, PhD

The University of Southern Mississippi

Erin M. Buchanan, PhD

Harrisburg University of Science and Technology

**Reviewer 1**

**Comment #1:** My main suggestion is that the paper needs to make contact with the other work that uses fuzzy-matching to score written responses in research (e.g., Autoscore and the Token Sort Ratio method). Lrd goes beyond those, but it would be helpful to have references to the previous work and an explanation of the differences to make it more clear what lrd is offering that the others are not.

***Response:*** Thank you for bringing these tools to our attention. Based on your review and the suggestions from other reviewers, we have added a brief General Discussion beginning on page 42 that discusses each of these tools and compares them to *lrd* (pgs. 43-45).  
  
**Comment #2:** Another problem with manual scoring you might mention in the introduction is that if scoring is not done blind to condition, it provides an opportunity for bias to sneak in.

***Response:*** Great point. We have added a sentence to the introduction on page 5 listing this an additional pitfall of manually scoring data.  
  
**Comment #3:** When introducing the different datasets in the intro, it would be nice to include a short statement of why it’s interesting to look at multiple different types of tasks. I think it is, but it would be helpful to have it spelled out for us (e.g., would we expect that free-recall should show different responses than a sentence-recall task?)

***Response:*** The introduction now contains additional discussion on page 6 regarding the benefits of using *lrd* to score data from multiple test types.  
  
**Comment #4:** I didn’t quite follow why you simulated participant data rather than just using the datasets you use for validation? It seems like it might be more straightforward to just select a few participants from the datasets you end up using in the validation section so you don’t have to explain how you did the simulation (see my comments on length below)

***Response:*** Simulated data were used to provide examples of a variety of different types of errors. Since most participants provide very few errors in their recall, using a participant dataset would not be able to show the breadth of participant errors that *lrd* can adjust for.   
  
**Comment #5**: Really smart to make a shiny app in addition to the R package.

***Response:*** Thank you. We felt it was important to make *lrd* accessible to researchers who may not be familiar with using *R*.  
  
**Comment #6:** I wonder if it might be useful as part of the “arrange data” tab in the shiny app to give the option to download a CSV that has the headers formatted appropriately? Just to help along folks who are struggling with figuring out the form the data should take? Might be overkill, just trying to think through all the problems that people who aren’t familiar with this might run into.

***Response:*** Yes, it is important to make *lrd* as user friendly as possible. We have added a link in the “arrange data” tab’s description that allows users to download a .csv file from our OSF page with headers in the correct format.

**Comment #7** I’d encourage you to look for ways to make the manuscript shorter. I’ve flagged a few spots that seem like candidates for cutting. Of course you are to be commended for being thorough, just trying to increase readability.

***Response:*** Thank you for the suggestion, we have removed some of the extra details suggested in comment #8.  
  
**Comment #8:** I wonder if you might cut some of the detail about the datasets you’re using for the validation study? Looking at details like “stimuli pairs used in this study were originally selected based on their levels of forward associative strength…etc.” Seems like what people will care about primarily here is the nature of the tasks themselves, perhaps a bit about sample size, but the details of the words doesn’t see super important? Sorry if I’m missing concerns others may have.

***Response:*** We have made a point of reworking the introductions to the cued- and free-recall validation studies so that the nature of the task is focal. For the cued-recall validation study introduction, we have removed the paragraph which described details of the stimuli used in each cued-recall task (pgs. 30-31) such as semantic relatedness measures. This section now only reports the sample size and details of the cued-recall tasks. For the free-recall section, we have reworked the description of the dataset to be more concise (pgs. 35-36).  
  
**Comment # 9:** Sensitivity/specificity analysis is cool! And cool that researchers can change the Levenshtein distance if they have a reason to. Any concerns that that’s a researcher degree of freedom that can be gamed? Maybe helpful to make the recommendation more strongly that unless there is a clear reason to change it, stick with the default?

***Response:*** Thank you. We made the Levenshtein distance variable so that researchers could tailor the scoring algorithm to their dataset. Of course, it could allow researchers to unfairly manipulate the scoring, but the same could be said of standard recall scoring by a biased experimenter. We mention this point in the General Discussion (pgs.42-43) where we discuss optimal Levenshtein cutoffs for each task type and make recommendations for best scoring practices.

**Reviewer 2**

**Comment #1:**  As I was trying to use the package in R, and via Shiny, I realized that there is one particular type of study design that the package cannot straightforwardly handle - one where each participant or a group of participants are shown different sets of study lists. Initially, I thought the study design from Huff et al (2018) on page 34 was similar to this, but if I understood correctly, all participants studied all six lists in that study. I think the package would benefit from a function that can adequately handle different study lists for different participants/groups -- stimuli randomization is becoming increasingly common in the field, and this type of function/processing capability would add to the strength of the package and truly make it flexible for use across simple and complex lexical/memory experiments. I would suggest that the authors look into this and if possible, provide a function and/or guidelines about how to go about analyzing such types of data in an automated manner via lrd.

***Response:*** We appreciate your constructive feedback. As originally written, *lrd*’s free recall scoring function, prop\_correct\_free() along with its corresponding tab in the shiny application, could only score free-recall responses if all items came from the same study list. This meant that if a dataset contained responses coming from more than one study list (such as the Huff et al. 2018 data in which participants studied words taken from ad-hoc, categorical, and unrelated lists), the dataset would first have to be subset by list, and then scored separately. We originally scored the Huff et al. data by splitting the dataset into six parts (one corresponding to each list), scoring each subset separately, and then combining the lists by type. This process resulted in three scored datasets (one for each of the three list types).

While this process of subsetting can be used, we realized that this would become complicated in research designs with more groups/conditions and that complications would increase further if researchers employed multiple counterbalances. Therefore, we have developed an additional function for scoring free-recall responses when multiple lists are used (prop\_correct\_multiple(), which is introduced on page 18). This function requires the inputs as prop\_correct\_free() and contains additional arguments for denoting list identifiers (i.e., which columns in the dataset and answer key contain information about which study list a key item or response belongs to). We have updated the shiny application with a “Multiple Free-Recall” tab, which is based on the prop\_correct\_multiple() function. This new tab is described on page 28.

**Comment #2:** A related question I had was also about how the data was processed for the above study -- the description on page 34 states that the scoring was done separately for each list, it might be helpful to have a vignette or small example of how one may go about scoring when there are multiple lists involved.

***Response:*** This section (now page 36) has been updated to reflect processing using the new prop\_correct\_multiple() function. Additionally, include a vignette (Multi\_Recall) which details how one would process and score free-recall data when multiple study lists are involved.

**Comment #3:** The Shiny app is a bit glitchy -- I was able to use it once, but after that I kept getting a “disconnected from server” error. I tried reloading it several times, the page would let me upload the data but not score it.

***Response:*** This may reflect a connectivity issue. We have since tested the application at multiple times across different browsers and have not encountered this issue.

**Comment #4:** I think the paper would also benefit from a demonstration of how to plot the serial position curve using the package and base R functions, as well as via Shiny.

***Response:*** Given concerns with our manuscript’s length expressed by our other reviewers, we have elected to forego including an in-depth demonstration of plotting the position curves in *R*. The free-recall vignette, however, contains an example of plotting each curve in *R* using a combination of *lrd*, *ggplot2*, and base *R*. We now refer interested readers to this vignette on page 19. Finally, these plots are automatically generated with the Shiny application, a point which is also emphasized on page 19.

**Comment #5:** Also -- when the scored data preview is presented within Shiny -- is it possible to explicitly display the “scored” column -- otherwise the user has to download the data to check if the scoring even worked -- might just improve the general usability of the app if users can see the new variables that have been created in the preview itself.

***Response:*** The scored column was most likely not displaying due to the size the output tables (i.e., if a dataset contained many variables and/or if it was being displayed on a small screen, the scored column would be “pushed” off the edge of the screen). We have now added horizontal scroll bars to all tables that allow users to view all columns without having to download the scored dataset.

We appreciate your helpful “beta-tester” suggestions as they have improved the quality of the application!  
  
**Several minor wording edits**

***Response*:** We appreciate your attention to detail. Each of these minor spelling and grammar mistakes have been corrected.

**Reviewer 3**

**Comment #1:** The content in pages 9 to 20 provides a tutorial on using specific functions. I noticed the authors also have an online tutorial document, as well as vignettes in the R package (often used for documentation and examples). For purposes of streamlining the paper, the authors could considering moving this tutorial style content into the online documentation/vignettes.

***Response:*** Thank you for taking the time to thoroughly review *lrd.* While we agree with the notion of streamlining the manuscript, we feel that it is important to include our tutorial information collected in one place to provide a complete walk-through of the package. Therefore, we have elected to keep the tutorial information on pages 9 – 22.

However, for the sake of concision, we have made cuts in other places. Specifically, we have reworked the introductions to the cued- and free-recall validation studies (pgs. 30 and 35, respectively) so that they now focus less on the details of the stimuli and more on the nature of the recall tasks. In the cued-recall validation introduction, we removed the paragraph describing details of the stimuli used for each cued-recall task (pgs. 30-31) and instead only report the sample size and details of the cued-recall tasks. For the free-recall validation, we have edited the dataset description for concision (pgs. 35-36).

**Comment #2:** I could see some value in adding a general discussion section. As it stands, the paper moves rather quickly from the validity tests to the conclusion. Some discussion points could include an assessment of the limitations of the lrd package, how it compares to other automated methods (if there are any), how the authors plan to support the package over the long term (or ideas for future additions to package), how others might contribute to the package given that it is being hosted on Github. Additionally, the authors mention that they developed lrd so that users with minimal programming experience could use the package. This is a laudable goal. However, it was unclear to me how much programming experience may or may not be necessary for researchers to use lrd. A more complete assessment of the kinds of pre-requisite knowledge that a user might be expected to possess before using lrd could be useful. To be clear, I don't think that it is necessary for  lrd to be novice friendly, just that some additional discussion on those issues could be useful.

***Response:*** We agree that adding a General Discussion section would be beneficial to our manuscript. We now include a General Discussion (pgs. 42-45) which summarizes scoring results for each recall type and provides recommendations regarding cutoff values based on different factors that may influence the selection process. Finally, we use this section to compare *lrd* with two other scoring tools (Autoscore and the Token Sort Ratio) and discuss the advantages of using l*rd* to score output from recall studies compared to these tools. Of course, we are cognizant of your recommendation above regarding keeping the manuscript short and therefore kept our discussion brief.  
  
**Comment #3:** Based on the different validation studies, the authors give slightly different recommendations for Levenshtein cutoff throughout the manuscript. It could be helpful to create a section the general discussion that focuses on these recommendations for future researchers, delineates each recommendation in one place in the manuscript, and discusses potential alternatives. For example, should other researchers just adopt a cutoff of 1? What other options could researchers avail themselves of to determine the best cutoff for their study? What kinds of study parameters might influence these choices? I think some additional discussion of these issues would benefit the manuscript.

***Response:*** We agree that a discussion of these issues would make for a stronger manuscript. These topics are now addressed in the general discussion on page 43.

**Several minor wording edits.**

***Response:*** Thank you. We have corrected each of these minor grammatical mistakes.