**Discussion points regarding coding of the elicitation survey**

Did we have blocks? Are they marked in the data? Yes, possible. **DONE**

First trial of each block at the individual trial level not as outlier? **DONE**

**This could be done.**

Use pre-registered criteria, with an example that is not included in the pre-reg. Include the example. **DONE**

What does trim mean: remove or replace? Remove. **DONE**

Apply a hoaglin and iglewicz (1987) outlier removal procedure. Check what this means. **DONE**

[***https://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478551***](https://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478551)

***I don’t have access but it appears to be a sliding rule like Van Selst***

**TH: As I understand it, it is a variant of Tukey’s Q1-1.5IQR Q3+1.5IQR outlier definition, where instead of 1.5 they use a constant (k) depending on sample size and the proportion of data you want to remove, assuming normality. In the paper, they have a table that specifies k for different sample sizes and outlier proportions (5 or 10%). Given that the response didn’t specify the latter, I’m inclined to categorize it as not operationalizable.**

**EB: makes sense to me as well, so let’s say it’s not possible.**

“Removing participants who are outside of 2.5 sds from the mean” not trials? **DONE**

Remove participants who alternate or keep it constant. Necessary? **DONE**

MAD? How does this work (for accuracy in particular). **DONE**

[***https://www.mathworks.com/help/matlab/ref/isoutlier.html#bvolfgk***](https://www.mathworks.com/help/matlab/ref/isoutlier.html#bvolfgk)

[***https://dipot.ulb.ac.be/dspace/bitstream/2013/139499/1/Leys\_MAD\_final-libre.pdf***](https://dipot.ulb.ac.be/dspace/bitstream/2013/139499/1/Leys_MAD_final-libre.pdf)

***See bottom.***

**TH: makes sense now. One more question though: for RT, shouldn’t it also be about the mean RT per participant? This was the input from the survey “ Based on both the mean (tranformed) RT (for all trials) and the accuracy of each participant, exclude participants that are outliers for either measure based on a on-parametric criterion (i.e., more than three scaled MAD above and below, respectively for RT and accuracy, from the median.” I tentatively coded this as *Across trials: Calculate each participant's accuracy and remove those whose accuracy is more than three scaled MAD above and below the median accuracy, with scaled MAD defined as c\*median(abs(accuracy-median(accuracy))), where c=-1/(sqrt(2)\*erfcinv(3/2))* and *Across trials: Calculate each participant's mean RT and remove those whose mean RT is more than three scaled MAD above and below the median, with scaled MAD defined as c\*median(abs(mean RTs-median(mean RTs))), where c=-1/(sqrt(2)\*erfcinv(3/2)).***

**EB: yes agree, by participant, except accuracy which is overall by the nature of it.**

Remove participants who did not provide the correct response greater than 30% of the time. **DONE**

check for multimodality and exclude participants with multimodal RT distributions, see Fisher et al. 1994 Testing for multimodality and in R one could use the mclust package and or the silvermantest package. Need to check. **DONE**

use the R package {bestNormalize} to automatically test all available transformation method and pick the best one. Need to check. **DONE**

[***https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html***](https://cran.r-project.org/web/packages/bestNormalize/vignettes/bestNormalize.html)

***Looks doable***

**TH: question is whether we want to include transformations (see later)**

**EB: we decided against this**

What to share? **OPEN**

**I vote we share the coded data - unsure how identifiable the raw data would be?**

R\_3EtY5xnBXHE6k4q

“Remove data points with an RT below 200 msec or above the upper quartile plus three times the IQR.” Not operationalizable or Across trials: RTs beyond third quartile + 3\*interquartile range are removed

***I would assume across trials.* DONE**

R\_p65Iho57xz7gbsZ

“I would exclude RTs above/below 2.5SD to make sure that I have as homogenous and coherent data as possible. To this end, I would do it separately for each variable; namely RTcar-dog and RTcat-dog” coded as “Per condition (related vs. unrelated): RTs +- 2.5 SD from the condition-specific mean are excluded” was wondering whether it’s “Per condition (related vs. unrelated) and item combination: RTs +- 2.5 SD from the condition-by-item specific mean are excluded”

***Good call.* DONE**

R\_p65Iho57xz7gbsZ

“The priming is measured as follows: zRTcar-dog - zRTcat-dog, but I would still divide it by zRTcat-dog which would take into account the individual differences (similar to interference effect measured by Stroop-like tests).” seems wrong as it would result in negative effects. Keep as is?

***How would this be negative in a bad way … not following.***

***Let’s say it’s 50ms - 45ms (5) / 45 ms = .11***

***Versus 40ms - 50ms (10) / 50ms = -.20***

***So, negatives mean the wrong direction (slower processing). I just didn’t think it was useful because we are already z-scoring on participant. However if you are z-scoring overall maybe this is useful? Why divide twice to control for individual differences (not that I think this is individual differences more like stimuli level differences)?* TH: yes, I read this wrong. DONE**

R\_1Pc6XqO5Rh3SJV9

“not include trials beyond certain thresholds (e.g., 5th and 95th percentile) per individual” = > not included in the coding?

***I put it in the other box - it’s a complicated answer.***

**There are many ways in which response time data can be preprocessed. I would start with more general and conceptual steps (the way how I would think) and then proceed to more concrete steps (the way how I would proceed).**

**To prepare an optimal strategy suited for the project, I would check if there exist any recent standards/guidelines that I can follow. If not I would prepare it by myself (see information provided below) based on the literature and forums. Also, (if my programming skills would allow) I would try to simulate data or reserve some portion of pilot data or at least try the whole procedure on myself and simulate various potential problems (e.g., wrong answers, keyboard layouts, distractions, to fast or to slow responses and so on) to contextualize it. Based on these ‚Äútrain‚Äù data, I would conduct a pilot data-processing procedure. After the pilot testing procedure, I would pre-register the procedure before the final data processing to limit my degrees of freedom when working with real datasets. Optimally, besides the primary analysis, I would also conduct and report the sensitivity analysis without cleaning procedure (as not treating outliers is an option similar to treating them in some way) or even include a few key decision points that can substantially alter results (multiverse analysis). I would be as transparent as possible.**

**To provide more concrete steps suitable for the present task, I would follow these steps:**

**First, I would check invalid traits and check if the stimulus presentation or response recording was not faulty (for example, I would check stimulus onset times). I would not include participants that have faulty presentations/recordings.**

**Second, I would check errors where participants gave the wrong answer or failed to respond. In particular, If the errors are too abundant, I would not include participants beyond a certain threshold (e.g., 75% accuracy) as these participants could have technical or other problems. Also, I would not include wrong trials as these can bias results.**

**Third I would consider dealing with outliers. By outliers, I mean response data that are extremely fast (e.g., a response time shorter than 100-200 ms. is probably not valid) or slow (response beyond 2s could be biased due to attention lapses or distractions). The outliers can be handled in various ways (see e.g., Berger and Kiefer 2021 for simulation and comparison of different response time outlier exclusion methods). I would consider these recommendations and my technical abilities and available resources. If possible, I would try 2-3 approaches and provide a comparison of results.**

**For example, (3a) in the software that I tried for data collection of reaction times, it was very easy to implement a percentile method (not include trials beyond certain thresholds (e.g., 5th and 95th percentile) per individual. Based on this experience, I would include only trials that pass this criterion and then follow the suggested data analytics strategy (i.e., ‚ÄúResponse times to the critical targets will be z-transformed for each participant separately...).**

**Alternatively, to provide results with a more comprehensive alternative approach, according to Berger and Kiefer (2021), the methods excluding RTs as outliers based on z-scores (‚Äú2sd,‚Äù ‚Äú3sd,‚Äù and ‚Äútransform‚Äù) showed considerably small (absolute) biases, few Type-I errors and excluded only small proportions of reaction times and, as z-scores will be used in the following analysis. Therefore, I would try to implement one of these methods; i.e., (3b) z-transformed values exceeding a particular z-score (e.g., 3) will be excluded before subtracting z-transformed related and unrelated trials response times; or, optimally, (3c) I would use the formula mentioned in Berger and Kiefer (2021): ‚ÄúFor each transformed value, the square root of the untransformed value minus the minimum value of the sample divided through the sample range is calculated. The fraction bounds all values between 0 and 1, while the square root enlarges small values‚Äù. Afterwards, these values are z-transformed and values exceeding a particular z-score (e.g., 3) are excluded; however, to be honest, at this phase, I¬¥m not sure if this is fully compatible with the suggested analytic strategy.**

**DONE**

R\_1mPHRSyxhucS9rL

Coded as: Exclude negative RTs (Data exclusions; Trial-level, so not coded as outlier treatment, because it’s more indicative of an experiment issue), Across trials: Participants with more than 10% negative RTs removed (Data exclusions; Participant-level), and Across trials: Participants with an error rate above 40% removed (Data exclusions; Participant-level). The latter because “I would calculate the accuracy scores for each participant and ensure that participants' accuracy was at least 10% above chance-level performance” translates to 60% correct or more in SPAML, right?

***Yes, would agree with that. Is that 60% of what’s left after the experiment issues are handled?***

**TH: presumably, because it was the next step in the pathway proposed by the respondent. But that’s not to be mentioned in the validation survey, I’d argue, because it would include info on the order of certain steps. DONE**

R\_3j98VqmfoEPS1gO

“I would remove trials in which reaction times deviate more than 2.5 sd from the mean, calculated separately for related/unrelated trials.” was coded as Per condition (related vs. unrelated) and participant combination: RTs +- 2 SD from the condition-by-participant-specific mean are removed. However, the response doesn’t mention participants, right? I was thinking coding it as Per condition (related vs. unrelated): RTs +- 2.5 SD from the condition-specific mean are removed

***Agree on the reread - don’t think this is a good choice but more closely approximates what they said.* DONE**

R\_2OO9NKGNhdP0pUB

“First, I'd exclude participants with low accuracy (<70%) in either word- or non-word targets” was coded as “across trials Ps must have at least 70% accuracy”. However, it seems to make a distinction between different conditions (words vs non-words). So, I was thinking coding it as Per lexical status (words vs. nonwords): Participants with an error rate above 30% are removed

***Good call, I agree - I suppose you could have 99 percent accuracy in words and less in non-words and so they would get excluded?* TH: yes, well less than 70% in this case. DONE**

R\_3PcMc0dwaXNvaHY

“I usually winsorize in my reaction time studies at 3 SD.” code as Across trials: RTs +- 3 SD from the mean are replaced by the mean +- 3 SD (i.e., the end of the distribution) (whichever applies)

***Would they be replaced by the mean or by the end of the distribution with traditional winsorizing? I may have selected the wrong thing here. It’s vague on the level of the winsorizing but I assume across trials.* DONE**

R\_2dTsyyot1SdYW7N

“We considered all the observations with a studentized (t - student) deleted residual greater than 4 in relation to incorrect answers as outliers” = > you can only do this after performing some analysis right? So, it is considered part of the analysis choices and hence, not included here.

***Probably - you could calculate these Mahalanobis style on randomized data, but not sure that makes sense on studentized residuals.* DONE**

library(dplyr)

# generate random data

DF <- data.frame(

person = rep(1:100, each = 100),

accuracy = sample(c(0,1), replace = TRUE, size = 100\*100, prob = c(.2, .8)),

rt = rnorm(100\*100, m = 1700, sd = 500)

)

View(DF)

# calculate accuracy

accDF <- DF %>%

group\_by(person) %>%

summarize(acc = sum(accuracy)/n())

# suggested by paper

cutoffRT <- median(DF$rt) + 3\*mad(DF$rt)

cutoffacc <- median(accDF$acc) - 3\*mad(accDF$acc)

DF$outlier <- (DF$rt > cutoffRT) # also do low

accDF$outlier <- (accDF$acc < cutoffacc) # high doesn't make sense

table(DF$outlier)

table(accDF$outlier)

**Discussion points regarding the pathway validation survey**

Anonymity survey? Not anonymous. Monitor quality. **DONE**

Attention check (deliberately bad choices), reliability? We could potentially use RT as a cutoff to identify “suspicious” participants (how do you get RT per question?). Hopefully, it will not be necessary given that the survey won’t be anonymous now. Also, I added a few attention/comprehension questions. We can exclude participants from the analysis, but we probably won’t exclude anyone from co-authorship, unless there are compelling reasons to do so. OPEN (see to do list)

**Decided to grab two criteria:**

* **Answer questions at the beginning that indicate knowledge of the instructions**
* **Time reading the instruction pages**

**If these two criteria are lower, we would exclude them.**

Do we present options in random order? Looks weird in some cases. Downside: the first items might have a higher preference, because of anchoring. **DONE**

**Decided against because of potential issues on survey delivery.**

Do wo provide a link to English and German data so participants could try out some things. Rationale: some people say/think: only do X if Y is the case (e.g., only remove data above a certain cutoff if the data aren’t normally distributed or if it affects a certain percentage of the data and so on). Downside: I don’t know whether the data is ready and whether we want to share the data already. Another potential downside: participants could in principle try out different strategies to get a desirable end-result. **OPEN**

**Decided to make a small representative dataset.**

About the relative outlier criteria: do we group them together or split them somehow? **DONE**

**Decided to leave it.**

What about transformations? Now we mention the z-transformation as a step with no option to change it, except the timing of this step. **DONE**

**Leave with z but give the option to change it in the open ended question.**

What about the order of steps. They are in a particular order, but not per se corresponding with how you’d do it. Do we suggest an order (explicitly or implicitly)? **DONE**

**Leave it as is.**

The experiment doesn’t give an error when the order is not consistent (e.g., remove filler trials when they are needed to determine whether participants made an error on the previous trial). **DONE**

**On our end we will filter options that are possible.**

Are all word filler trials, prime trials? **DONE**

**No - trial types:**

* **Word word related**
* **Word word unrelated**
* **Nonword word**
* **Word nonword**
* **Nonword nonword**

Age, are there any below 18 or are they automatically excluded? They take part and are excluded afterwards. **DONE**

I’ve added a “Back” button, but I don’t know whether that gives any issues. Have to test this. I initially did it so that participants could go back to the instructions, but I also added comprehension checks in between and then it would have been odd to go back and change their response. So now, you can only go back to certain points. For example, participants can go back and change their response to one of the data-processing decisions. But they can’t go all the way back to the instructions, which would be annoying anyway presumably. Instead, I’m considering adding a link to a google doc (or something else) with all the instructions underneath one another. I tested this for the first question with an empty google doc, and it seems to work (wanted to do this for a doc in the multiverse folder, but it didn’t allow me to share anything, I believe). Let me know what you think. **OPEN**

**Gave you all the power in the SPAML (include it a google doc).**

**Other discussion points**

Pre-registration? I’m leaning towards not pre-registering, given the pilot character of the study and the time it would take to do a proper pre-registration before sending the survey to participants. **DONE**

**Nah**

What do we eventually share about the coding process (not urgent)? **OPEN**

**I think we could summarize the coding process and include the coded information?**

<https://www.psychologicalscience.org/ampps/ampps-invited-papers-call-for-proposals> **DONE**

**We stick with Psych Methods as the first option.**

**To do**

* Have a look at the only remaining open issue about coding elicitation survey: Katja/Erin **DONE**
* Add timing to instruction “questions” in Qualtrics: Erin/Tom/Katja **DONE**
* Create a small representative dataset: Erin
* Link to dataset and instructions in all questions (currently only implemented for the first “real” question (so not the instructions or the attention checks), but without a link to the data). Also, the link to the data is mentioned in the instructions, so we need to insert it there too, both in Qualtrics and the google doc with instructions: Erin
* Add the final instructions to the google doc and Qualtrics: Tom **DONE (except for the link to the dataset)**
* Somehow, I can’t figure out how to change the sharing settings for the google doc with instructions to view only. Maybe I don’t have the rights for that, or, more likely, I’m a noob, so could someone help? Alternatively, we could just link to a pdf: Erin/Katja
* Add informed consent and ask approval from ethical committee: Erin