Complete free-text responses in the expert-consensus survey

(corrected for spelling errors).

Data exclusions due to events beyond the researcher's control, such as: *Research assistants collecting data from participants who do not match participation criteria (e.g., age, gender, visual acuity); Equipment malfunctions, technical errors. Do you have any feedback to provide on this processing step?*

1. Exclusion due to present criteria I always report and this is common. If trials fail due to some unforeseen circumstances we usually exclude the entire participant. So the number of trials excluded is not reported as the remedy is the exclusion of the participant.
2. Have had experienced this for the need for stimuli exclusion (hence, responded 'never'), but agree that same criteria should be applied.
3. There isn't an NA option for if I never exclude using these criteria
4. I always report fully the exclusion criteria and the outcome of exclusions.
5. Most important criteria for me are to define exact measurement points, I. E. When does RT start/end. This is often left in publications and requires interpretation.
6. Exclusions for these reasons especially equipment malfunction or noise not data. In terms of participants, we usually test everyone who agrees to participate but may exclude the occasional participant for example a child who has some kind of characteristic like a lower queue, but we didn’t know that ahead of time
7. Conduct robustness check with and without excluded data
8. The last point (stimuli) is listed as 'never' but should really be 'NA' (I've never encountered a situation where this would be necessary)
9. This should occur not that often anyway
10. The criterion mentioned here is mainly for excluding participant instead of trials in my case.
11. Some things are not relevant to me, for example, exclusion of stimuli, because my stimuli repeat and it makes no sense to exclude them. Also, if a participant doesn't match criteria, I'd exclude everything, not just some trials. So that's also partially not relevant to me.
12. It’s necessary but I doubt it’s making big impact on my results though.
13. It almost never happens, but obviously I report it if it does.
14. "reporting" depends on if it is in the paper or the SM. If you share all your code/data (as I do now), all steps and values are, in some sense, "reported" but they may not be made salient in the paper proper
15. If the experiment isn't working properly, it needs repeating; you can't just remove 'bad' data.
16. Sometimes if students collect data, it is unknown whether they excluded participants
17. This question is too vague to answer.
18. Not relevant for my personal research

19. If sample size is large enough, noise from such factors will not affect the results too much. My philosophy is usually to limit arbitrary exclusion criteria. If my effect is real, it should survive such small external factors that are quite rare. Of course, if I am aware of a mistake, I would exclude those participants and report it.
20. These criteria should always be checked, but I don't typically encounter trial or stimuli level errors of this sort and participant level errors are only found occasionally.
21. Technical errors (power cuts etc.) should be reported to illustrate not everything went smoothly, I get skeptical if there are no missing data due to this kind of events
22. On the previous page I mentioned never excluding data unless it is unrealistic… this was assuming it indeed comes from the population I am interested in and the experiment/stimuli are appropriate. If I find that they are not (for any reason) I always report.
23. This step always forms part of the preprocessing steps based on my own training in human-subjects research
24. There should be an "NA" option above because we would never have to remove stimuli during a study (after the pilot phase I guess). The "Nevers" above should be NA -- but if we were to do such a thing, we would report it.
25. I think that most outlier exclusion techniques bias the results. Therefore, it is very difficult for me to give a clear recommendation. I recommend using fixed cutoffs for outlier exclusion. I also must say that this part of your questionnaire is not entirely clear to me. Are you concerned about replication?
26. I had not such experience and in all studies all stimuli and all participants were included in the initial set.

Data exclusions based on fixed criteria, such as: *A minimum percent correct answers; A minimum duration needed to visually perceive the stimulus. Do you have any feedback to provide on this processing step?*

1. Excluding trials or stimuli is often not possible, since we often have binary (0 / 1) answers on trials, which are often only administered once per task
1. I choose not to exclude stimuli because they are part of the design choices for the study, and typically I use within-subject designs where performance happens in a task environment where all stimuli were presented. I do perform analyses by items (Anova) or include stimuli as a random factor in linear mixed models (unless there where the model for an effect does not converge).
2. The importance depends strongly on the number of trials, i.e., missing some when a participant was measured many times is less critical than if a participant's RT was only measured once.
3. If the response time is zero or less than 150ms, then this sort of fix criterion works. However, it’s not all of it always that obvious what fixed criteria you should use. So, we don’t necessarily decide this ahead of time it might depend on what the task looks for that sample. But we would always report any exclusions.
4. As before, I've never needed to exclude stimuli, so 'never' should be read as 'NA' in this case.

5. This is common practice in psycholinguistics, if the item or participant is below 80%, we are told to remove them. In memory, it is more common if it is below 50%.
6. Combining the two reasons makes most of my responses invalid because I take different actions at each of them. If a participant doesn't pass minimum accuracy, they're out and fully reported. If some trials are too fast, they are excluded but I don't report the specific percentage: I just report the overall percentage of trials excluded due to being faster than 200ms and +/- 2.5SDs. So it's different actually. Excluding stimuli is not relevant to me
7. Outliers can make big impact on my results
8. If subject can't perform the task (not sig better than chance), the experiment is not working.
9. This varies by whether I am doing things like assessment of tests versus surveys/personality stuff.
10. in my research, this step is mostly a technical necessity (e.g., models cannot deal with missing values), not a conscious decision
11. Ubiquitous in the processing of indirect measurement tasks such as the IAT.
12. There is always at least one participant that either misunderstood the task or rushes through it. We pre-reg these exclusion criteria.
13. In my previous answer I forgot to mention that I also typically exclude responses shorter that 200ms from analyses.
14. In our studies, this typically doesn't apply. We always assume that our participants can see the stimuli and because we test pre-verbal infants, there is no way to ask them if they can see the stim (but we do design the stimuli w/ their visual acuity in mind).
15. I do not exclude stimuli or participants, but rather specific RT.

Data exclusions based on data-dependent criteria, such as: *Participants with an incorrect response rate that is too far from the average incorrect response rate; Trials with a response time smaller or larger than, respectively, the average response time minus or plus twice the standard deviation; Stimuli with an incorrect response rate larger than a minimum absolute deviation of the median. Do you have any feedback to provide on this processing step?*

1. I believe this is similar to the previous step, and in a way, an alternative. I have always approached analyses as: either you use a predefined threshold (e.g., impossible RT as a minimum threshold), OR you use a threshold (say, deviation from mean) based on the data. I tend to opt for the first option because it's then more consistent across experiments, and the second option is not always a good idea when individual differences in RTs are of interest.
2. Next to exclusion, one could also think about trimming RT data to 3 SD above the mean (for example). For my most recent research, I did both, and reported the results of trimming instead of removing in the appendix to show what deviations from the results this leads to
3. I write a registered report or at least a preregistration for each of my studies, and it is sometimes hard to tell in advance which data exclusion procedures are reasonable, but this way I inform myself before data collection and this makes it easy to stick to my

own set rules and to not forget any data exclusion step because the procedure is made public online a long time before even starting the analysis.

4. I don't explicitly report number of trials excluded, but I do report % of total that were excluded, which could be used to easily find the number of trials
5. There may be situations where the kind of exclusion makes sense. But normally we wouldn't use these on the first step we would look at the overall pattern of data. In my experience it makes very little difference to start slicing and dicing like this. If you need to do that to get an effect then probably there's something else going on and you should go back to the drawing board with your task when and where your stimuli
6. As before, I've never needed to exclude stimuli, so 'never' should be read as 'NA' in this case.
7. Our lab usually used M +/- 3 SD but I recently I switched to Median +/- 3 MAD
8. It is extremely important that this is *never* done. 'Outlier' exclusion is p-hacking. End of story.
9. Not relevant for my research
10. Used these types of criteria in for other tasks, but not common in IAT data.
11. We design it so that a few outliers should not affect the main result, hence we extremely rarely exclude participants by SD or other criteria; rarely exclude trials, can't remember a single case to have excluded stimuli based on that
12. Apply my own methods to those provided

Data transformation, aggregation, and scoring. *Transformation, such as log-transforms for normalization of skewed response times; Aggregation, such as averaging all trials for participant means; Scoring, such as creating mean differences or D-scores. Do you have any feedback to provide on data transformation, aggregation, and scoring?*

1. Data transformation and aggregations before analyses can skew the analysis results, so should be used cautiously / avoided where possible.
2. If I transform data, it is following published guidance that I understand the benefit of. If in doubt I seek advice.
3. We never use transformations because I think with response time data, you're assuming that those response times are flex and cognitive processing. We do use aggregation although more recently we've been using multi-level models to reduce the amount of aggregation. Much of my research has evolved individual differences in Kong in the processing and so aggregating across trials is equally as bad as aggregating across individuals because you may be of scary what's going on. When I first started doing research search in this area however there were no good ways to deal with those kinds of data so instead, we try to have as uniform of conditions as possible.
4. I generally use diffusion model to estimate underlying decision and memory process and I use raw RT data without transforming them considering that the diffusion model can handle it and right-skewed data is an inherent characteristic of RT data.
5. No blanket rules; it depends

6. Norms and usage depend on the task. For many tasks such as the IAT, best practices have been established based on psychometric analyses of alternatives scoring methods.
7. I don't use these methods as am yet to analyse my data
8. we aggregate across trials we rarely transform (done previously) unless reviewers force us to do so for certain experiments (sc-IAT) beregner D-means
9. Transforming the distribution to resemble normality is what I do and recommend. However, I do not use aggregation and scoring, I do not recommend these.
10. Lately I've transformed data less and use different likelihood families.
11. Transformation of RTs is a topic which I have struggled to understand, as there are arguments for and against log-transformations, and in my own research this decision has substantially change the results from significant effect to no significant effect

Processing Order. *Indication of processing steps order. Do you have any feedback to provide on this item?*

1. Importance depends on the content of the steps involved.
2. The importance of order really depends on what the steps actually are and whether/how they might interact with one another to alter the results. I often report using the processing methods outlined by previous work rather than detailing them in my own methods.
3. we provide scripts
4. So important as the steps need to be chronological for reproducibility

Checklist feedback. *Here are the proposed checklist items in a table format. Given this summary, do you have any final thoughts about items or concerns we might have missed?*

1. The most important issue that has not been raised is the reason for exclusion. If you exclude a participant, stimulus, or trial based on some criterion that criterion should be accounted for in your theory. If your theory cannot justify why anything above a 2 SD mark does not count your theory will need to explain why and this is almost never the case. People typically state that anything above some arbitrary mark does not count which cannot be justified by the theory they aim to test, making the theory thoroughly underdetermined. As such, transparency with regard to the reasoning for an exclusion and why that makes *theoretical sense* is highly important in my opinion.
2. I think "order" should just be implicit in that the reporting order should reflect the order in which things were done to the data
3. encourage reasons for decisions, e.g., power calculations.
4. Measurement timings
5. Fine
6. I think that would be a very useful table to use especially when you need to go and write up the results. I think would also be super helpful for students. I wish I could save it!
7. Assumptions fulfilled and simulations

8. Statistical models that are used? RM ANOVA or linear regression or models like DDM.
9. Proportion of trials excluded *per participant, per condition*. Or at least per condition. One needs to make sure they don't end up with very little in only one of their conditions.
10. Really nice idea to include transparency! Checklist is great. Only for the item "Total number of participants (per condition)" I am not sure whether it should really be "per condition" because in a within-subject design you would typically exclude a participant for all conditions - maybe you mean "by task" or "by group" or whatever? Another thing to add would be "case-wise" vs. "list-wise" exclusion.
11. nothing missing comes to mind, but what is reported where depends somewhat on the structure of the paper
12. The table looks great, please write a Ten Simple Rules article about that (or joined with me :-). Most of our papers do report all that, if not it was often due to silly word limits
13. Criterions for each of the decision on exclusions
14. Great checklist, thanks for creating it. I would like to use it in the future.
15. It looks great to me!!
16. As before - data sharing format and guidelines.
17. I like this as a template for standardizing research reports. I think there would be pushback from some on reporting exclusion information before the analysis/results.
18. I think you are concerned with replication. Have you ever thought about the base rate of true effects?

Final thoughts. *Thank you for your thoughts on our proposed processing checklist. Now that you've reviewed these items, do you have any final thoughts about items or concerns we might have missed?*

1. Sharing data processing and analyses scripts (e.g., on GitHub) is also important!
2. If I have never transformed my data so far I would respond with 'I never report this'. That is then due to not applying it and not due to not wanting to report it
3. It would be great to have a public checklist, maybe even approved by the APA or another big research institution of the field, to which researchers can stick when reporting analyses, so that no important details remain unclear.
4. Thanks. Interesting survey.
5. I like this study!
6. We need to include whether you check the statistical assumptions of reaction time or report any reliability checklist, because this is rarely discussed or if you simulated data.
7. I guess it is important to note when do research set the criteria: before data analysis or during data analysis.
8. split the different suggestions of procedures
9. Reaction time is a big thing in the emerging drift diffusion model

10. I'm not allowed to go back my previous answer which I wanted to edit. I accidentally went forward before finishing one of my answers regarding the steps I take to adjust my data before analyzing.
11. Unfortunately, I couldn't go back in the questionnaire. I just understood the categorization of criteria a bit later but couldn't correct previous responses. Also I was not sure about the if the scale for "did you report" relates back to "did you use" or was independent.
12. Please do not shy away from recommending that data are never excluded apart from for technical errors.
13. I publish the preprocessing & analysis code, so reproducibility does not solely depend on steps reported in the manuscript
14. should these steps be pre-registered? Could one "skip" over describing it if one uploads the analysis script?
15. Do you make your code available
16. I was a little confused at first as to whether “how often you apply these methods” was referring to only cases when necessary (at least for data exclusions), or if it was referring to how often I would do it if the occasion arises (e.g. I’ll always be transparent about the exclusions, but this only sometimes occurs in my research)
17. Asking about data sharing and its level (trial / aggregated / pre-post exclusion) - this is a mess when it comes to standards and would be good to know.