We Don't Know What You Did Last Summer. On the Importance of Transparent Reporting of Reaction Time Data Pre-processing

Hannah Dorothea Loenneker[1], Erin M. Buchanan[2,a], Ana Martinovici[3,a], Maximilian A. Primbs[4,a], Mahmoud Medhat Elsherif[5], Bradley J. Baker[6], Leonie A. Dudda[4], Dušica Filipović Đurđević[7], Ksenija Mišić[7], Hannah K. Peetz[4], Jan Philipp Röer[8], Lars Schulze[9], Lisa Wagner[10], Julia Katharina Wolska[11], Corinna Kührt[12,b], & Ekaterina Pronizius[13,b,c]

[1] Diagnostics and Cognitive Neuropsychology, Tübingen University, Tübingen, Germany

[2] Analytics, Harrisburg University of Science and Technology, Harrisburg, USA

[3] Department of Marketing Management, Rotterdam School of Management, Erasmus University, Rotterdam, Netherlands

[4] Radboud University, Nijmegen, Netherlands

[5] Department of Psychology, University of Birmingham, Birmingham, UK

[6] Temple University, Philadelphia, USA

[7] University of Belgrade, Belgrade, Serbia

[8] Witten/Herdecke University, Witten, Germany

[9] Freie Universität Berlin, Berlin, Germany

[10] Jacobs Center for Productive Youth Development, University of Zurich, Zurich, Switzerland

[11] Manchester Metropolitan University, Manchester, UK

[12] Faculty of Psychology, Technische Universität Dresden, Dresden, Germany

[13] Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria

Author note

[a] These authors contributed equally and are listed in alphabetical order.

[b] Shared last authors.

[c] Correspondence concerning this article should be addressed to E. Pronizius, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria. E-mail: ekaterina.pronizius@univie.ac.at

Funding

CK was partially funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, https://www.dfg.de; grant number SFB 940/2). Open access publication was funded by Open Access Publishing Fund of the University of Vienna. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors information

- Hannah Dorothea Loenneker hannah-dorothea.loenneker@uni-tuebingen.de, https://orcid.org/0000-0003-0327-6507.
- Erin M. Buchanan ebuchanan@harrisburgu.edu, https://orcid.org/0000-0002-9689-4189.
- Ana Martinovici martinovici@rsm.nl, https://orcid.org/0000-0003-1940-0293.
- Maximilian A. Primbs max.primbs@ru.nl, https://orcid.org/0000-0002-3398-5569.
- Mahmoud Medhat Elsherif mahmoud.medhat.elsherif@gmail.com, https://orcid.org/0000-0002-0540-3998.
- Bradley J. Baker bradley.baker@temple.edu, https://orcid.org/0000-0002-1697-4198.
- Leonie A. Dudda Leonie.dudda2@ru.nl.
- Dušica Filipović Đurđević dusica.djurdjevic@f.bg.ac.rs, https://orcid.org/0000-0001-5044-5428.
- Ksenija Mišić ksenija.misic@f.bg.ac.rs.
- Hannah K. Peetz hannah.peetz@ru.nl, https://orcid.org/0000-0002-1486-5342.
- Jan Philipp Röer jan.roeer@uni-wh.de, https://orcid.org/0000-0001-7774-3433.
- Lars Schulze lars.schulze@fu-berlin.de, https://orcid.org/0000-0003-4720-2411.
- Lisa Wagner l.wagner@psychologie.uzh.ch, https://orcid.org/0000-0002-1925-2676.
- Julia Katharina Wolska j.wolska@mmu.ac.uk, https://orcid.org/0000-0001-8675-4388.
- Corinna Kührt corinna.kuehrt@tu-dresden.de, https://orcid.org/0000-0002-6418-6479.
- Ekaterina Pronizius ekaterina.pronizius@univie.ac.at, https://orcid.org/0000-0003-1446-196X.

Abstract

In behavioral, cognitive, and social sciences, reaction time measures are an important source of information. However, analyses on reaction time data are affected by researchers' analytical choices and the order in which these choices are applied. The results of a *systematic literature review*, presented in this paper, revealed that the justification for and order in which analytical choices are conducted are rarely reported, leading to difficulty in reproducing results and interpreting mixed findings. To address this methodological shortcoming, we created a *checklist* on reporting reaction time pre-processing to make these decisions more explicit, improve transparency, and thus, promote best practices within the field. The importance of the pre-processing checklist was additionally supported by an *expert consensus survey* and *a multiverse analysis*. Consequently, we appeal for maximal transparency on all methods applied and offer a checklist to improve replicability and reproducibility of studies that use reaction time measures.

Keywords. Checklist, Open Scholarship, pre-processing, Reaction Time, Transparency

Word count: 10,108

1. Introduction

Chronometric methods, such as reaction time (RT) measurements, are used as a proxy to understand cognitive processes underlying behavior (e.g., Baayen & Milin, 2010). In the behavioral, cognitive, and social sciences, RTs are used to depict the consecutive processes from stimulus input (e.g., visual or auditory) to any output response (e.g., written, speech, or motor, Baayen & Milin, 2010). Given factors such as participant inattention or technical malfunction, it is common among behavioral scientists to pre-process RT data because those affected RTs do not depict the underlying process of interest. Once the influence of these random and extraneous processes has been addressed, behavioral scientists believe that only relevant factors remain to drive the RT pattern, such as the effect of experimental manipulation or group differences. A wide range of methods has been suggested to reduce the data collected to observations that are representative of the underlying cognitive process (see the following for methodological examples: André, 2022; Baayen & Milin, 2010; Ratcliff, 1993; Tukey, 1962; Van Selst & Jolicoeur, 1994), but so far there is no consensus on how to report the actions taken to prepare for statistical analysis. These pre-processing methods lead to different samples of items, participants, and observations, and correspondingly, *different* skewness of data distributions, measures of central tendency, and linear relations between RT and the independent variable [e.g., groups or conditions; Ulrich and Miller (1994)]. Therefore, pre-processing actions can alter conclusions from hypothesis testing, as shown in different proportions of significant tests in empirical studies (e.g., André, 2022; Morís Fernández & Vadillo, 2020) or in increased false positive rates in simulated data (e.g., André, 2022; Berger & Kiefer, 2021; Morís Fernández & Vadillo, 2020). Pre-processing not only changes effect sizes, but also the reliability with which constructs are measured (Parsons, 2022). Therefore, this article provides a checklist on reporting RT pre-processing actions, informed by discrepancies between a literature review on an exemplary

cognitive phenomenon and an expert consensus survey, a multiverse analysis, as well as personal scientific experiences.

The characteristics of RT data which make pre-processing necessary can be categorized according to different procedural levels: 1) artefacts outside of the researcher's control (e.g., technical malfunction, attrition, missing data, Morís Fernández & Vadillo, 2020; Woods et al., 2021; Woods et al., 2023), 2) data points exceeding pre-defined criteria (e.g., based on psychophysiological considerations, very fast RTs/anticipations and very slow RTs/omissions cannot depict the process of interest, Luce, 1986; Pain & Hibbs, 2007), and 3) observations deviating from the empirical individual or group average reaction pattern (e.g., RTs exceeding 2 median absolute deviations from the sample median). Transparency regarding which pre-processing has been used, in which order, and based on which rationale is utterly important, because the researchers' degrees of freedom caused by the many available methods can lead to different conclusions (Leys et al., 2013), as they result in different samples of included participants and observations (see below for results from the multiverse analysis conducted in this project).

Consequently, the effect of pre-processing decisions remaining unnoticed weakens inferences from the body of scientific evidence. Analytical choices made by one or a few researchers can be seen as different but equally rational, leading to conflicting empirical results and therefore, false positives and negatives in the literature (e.g., André, 2022; Berger & Kiefer, 2021). Therefore, as a first step, awareness needs to be raised for the effect of choosing certain pre-processing method(s) on the results (see Aguinis et al., 2013; Berger & Kiefer, 2021). Second, disclosing pre-processing decisions is necessary for reproducing and replicating published results (Morís Fernández & Vadillo, 2020). Reproducing numerical results by using the

same data and analysis of the original study is considered a minimum standard for evaluating the reliability of scientific findings (Nosek et al., 2022; Peng, 2011). Access to the original dataset used in a study is necessary, yet not sufficient, for assessing reproducibility. Chances of reproducing all the numerical results reported in a study increase as authors share information about pre-processing actions, code, information about the software they used, and adopt good computing practices (Wilson et al., 2014, 2017).

Given the increase in popularity of pre-registrations and registered reports (Christensen et al., 2020; Hardwicke et al., 2022), scientists have improved their ability to thoroughly relate their research questions to pre-defined analyses. However, as RT pre-processing takes place prior to data analysis and hypothesis testing, these decisions need to be reported with the same rigor and openness. Although there are guidelines on how to deal with outliers and missing data (Aguinis et al., 2013; Ratcliff, 1993; Woods et al., 2021; Woods et al., 2023), the field of psychological science lacks consensus on how to report the pre-processing of RT data. As a result, guidelines providing best practice examples of the latter are needed to further increase transparency. Therefore, we quantified the frequency of reporting pre-processing actions and assessed their completeness and transparency in the literature by conducting a systematic literature review on an exemplary cognitive phenomenon (the Simon effect), which is summarized in the following section.

2. Systematic Literature Search on Pre-processing Reports

To get an overview of how pre-processing is generally reported, we conducted a systematic literature review on the Simon effect (i.e., a cognitive phenomenon stemming from the difference between congruent and incongruent trials, Craft & Simon, 1970). For details on

methods of the literature search and a meta-analysis see **Supplementary Materials A**. The Simon effect has been chosen as an example because it has been repeatedly replicated since its introduction in 1970 (Cespón et al., 2020; Craft & Simon, 1970) and is representative of the classical behavioral experiment including a within-subject factor with two levels (comparing congruent and incongruent trials). The authors are not aware of a study investigating the influence of different pre-processing pipelines on effect size estimates in the Simon task, but there is at least one study (Morís Fernández & Vadillo, 2020) showing it for the Stroop task. Morís Fernández and Vadillo (2020) showed that as the set of pre-processing pipelines grew, the proportion of false positives also increased successively. It is important to note that only when a single pre-processing pipeline was applied, the proportion of false positives equaled to the α-level of the t-test, 0.05. There were no specific pre-processing pipelines that had the strongest influence, and it was shown that it is not even necessary to conduct all possible pre-processing pipelines every time: Simply considering them is already sufficient to increase the false-positive rate. The Simon task shares conceptual similarities with the Stroop task, and the Simon effect is classified as a spatial Stroop effect (Morís Fernández & Vadillo, 2020). As shown in *Figure 1*, we found a large variety in the number of RT pre-processing actions that were reported. Only five of 55 articles explicitly stated the order in which the pre-processing actions were applied, 32 articles indicated the rationale for the basis for choosing the respective method(s) applied, 16 reported how many or if any participants were excluded, and 21 documented how many data points were excluded through these actions.

Looking at the six articles not reporting any RT pre-processing, it remains unclear whether the authors did not mention data pre-processing they performed or whether they did not conduct any pre-processing. It is self-evident that this literature search does not allow us to make claims on discrepancies between pre-processing actions being conducted and pre-processing actions

being reported, as we cannot back-trace what has been done with the data in these studies retrospectively. However, observing that we cannot infer if any or which pre-processing actions have been applied to the data is an alarming fact for the endeavor of reproducible and robust research, which is why this article proceeds with the development of a checklist supporting complete reporting of pre-processing actions.

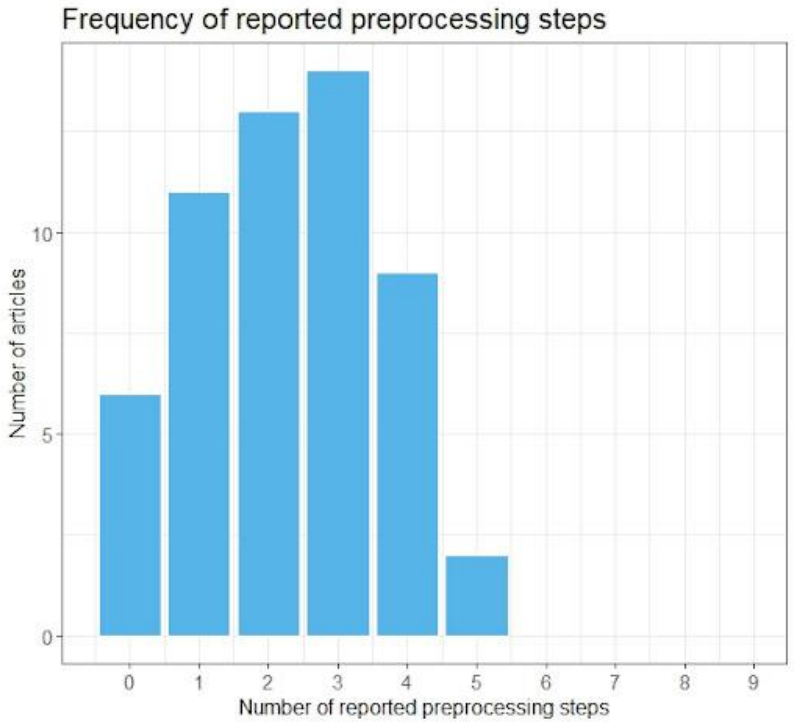

Figure 1. Number of reported pre-processing actions in articles included in the systematic literature search on the Simon effect. *Note*: Pre-processing actions include exclusion of error trials, exclusion of participants, handling of missing data, exclusion of fillers, exclusion of trials, outlier trimming with a fixed minimum, outlier trimming with a fixed maximum, relative outlier trimming based on measures of central tendency and variability, and data transformation. The number of reported actions per article can range from 0 to 9.

3. Development of a Checklist

Based on our literature review on the exemplary Simon effect, we observed that researchers rarely report the complete set of pre-processing actions for RTs. We assume that

these results cannot be interpreted as an exception, but rather as a case study representative of the field of behavioral sciences. To address this shortcoming, the present paper proposes a checklist to facilitate the accurate and complete reporting of pre-processing choices for the analysis of RTs. This overview of common pre-processing actions (*Table 1*) is based on previous surveys of the literature (Kerr et al., 2017; Primbs, Holland, et al., 2022) and our own literature search (see *Figure 1* and **Supplementary Materials A**). In the next section, we report the results of an expert-consensus survey confirming the importance and applicability of our suggested pre-processing checklist. We provide readers with an easy-to-follow checklist on which data pre-processing decisions to report in a manuscript and how to provide sufficient details thereof (*Table 2*). Last, we demonstrate the necessity of a reporting checklist by using multiverse analysis to illustrate the effects of the pre-processing actions on the standardized and raw results of the Simon Task.

Table 1. Overview of Pre-processing Actions with Examples

Action	Level	Reason	Examples
Data exclusion	Participant	External events	Research assistants collected data from participants who do not match participation criteria (e.g., age, gender, visual acuity).
			Participant reported headache and being unable to concentrate.
			Fire alarm went off. Building was evacuated.
		Outlier: Fixed	Minimum number of correct trials for a participant to be included.
		Outlier: Data-dependent	Mean accuracy was assessed in each group and those participants performing two standard deviations below the group mean were excluded.
	Trial	External events	Sticky keys were activated for the first few trials. The participant proceeded normally after deactivation.
		Outlier: Fixed	Trials faster than 100ms were excluded because it is unlikely that participant indeed processed the stimulus

Action	Level	Reason	Examples
			that quickly.
		Outlier: Data-dependent	Abnormally slow trials are likely to reflect distraction on part of the participant and are thus excluded.
	Stimuli	External events	There was a mistake in the processing of a stimulus. By accident, a stimulus was not shown often enough.
		Outlier: Fixed	Stimuli which were not correctly recognised above chance level were excluded.
		Outlier: Data-dependent	Stimuli which were recognised considerably worse than other stimuli were excluded.
Data transformation			Log-transformation was employed to make the data approach the normal distribution.
			Latency-normalisation was employed to account for between-subject variability in overall reaction times.
Other scoring			D-scores were calculated from the raw IAT data in line with established procedures (Greenwald et al., 2003).
Data aggregation			Analyses were conducted on the trial-level data / on the participant means.

Note: Order - Outliers were removed before data transformation took place.

4. Expert-consensus Survey

4.1. Method

4.1.1. Participants.

The study was approved by the Institutional Review Board at Harrisburg University of Science and Technology, USA (20221206). Participants were recruited through social media, internal listservs (e.g., faculty newsletters), and personal contacts of the researchers. 141 novel observations were collected through Qualtrics, and 66 responses were analyzed after excluding participants who did not explicitly consent to the study (n = 2), who did not provide any information about their use of response time data for research (n = 54), who indicated having no

experience with response time data (n = 1), incomplete responses (n = 15), and observations corresponding to survey previews (n = 3). In order to reduce risks of non-random drop-out (i.e., participants concerned about being identified based on their demographic information being more likely to drop out of the survey), the questions on year of birth and gender were optional. Of 66 participants, 10 did not enter their year of birth and two provided implausible values (the years 1900 and 2020). The remaining participants have an average age of M = 36.20 years (SD = 8.45). Fifty-eight participants provided information about their gender: 28 selected “woman”, 26 “man”, two “non-binary”, and two “prefer not to disclose”. Participants indicated they completed or were completing their education in Western Europe (37.9%), Northern America (21.2%), Northern Europe (10.6%), Western Asia (7.6%), Southern Europe (6.1%), Eastern Europe (4.6%), Latin America and the Caribbean (3.0%), and other or multiple regions (6.1%).

Materials and Procedure. The survey in text and Qualtrics import format can be found on our Open Science Framework (OSF) repository in the materials folder (https://osf.io/reqat/). After consenting to complete the study, participants were shown three main study sections: demographics, the proposed pre-processing checklist, and a final thoughts section.

Demographics. Participants were asked to explain the types of research they performed that used RTs, and this information was used to screen participants for the appropriate sample of researchers using RT data. Participants then indicated their geographical region where they were completing or completed their higher education, using the 17 United Nation Subregions classification system “United Nations geoscheme” (2022). In an open text box, participants indicated the subdiscipline that characterizes their research, followed by indicating their current role with options (e.g., students, lecturers, professors). Two software questions were included: participants listed all software they used to measure RTs and analyze RT data. Next, they were

asked to indicate the number of years (response options: *<1, 1-3, 4-6, 7-9, and 10+*) that they were involved in collecting RT data, analyzing RT data, open science/scholarship, experiment coding (i.e., writing code to collect RTs), and analysis coding (i.e., writing code to analyze data) respectively. At the end of the survey, they were asked to indicate their gender and year of birth for reporting purposes only.

4.1.2. Procedure Checklist Creation.

After answering the demographics questions, participants were asked to think of a project that used RT data and write down how they would process the data in preparation for analysis. For the exact wording, see **Supplementary Materials B**, section "*Open Response Times*".

Next, participants answered a series of closed and open-ended questions on data exclusions, data transformations, data processing order, and reproducibility likelihood. For data exclusions, participants indicated how often they used each of three criteria to eliminate observations at each of three levels, how often they reported doing so, and how often they reported the exact number of observations excluded (*never, sometimes, about half the time, most of the time, always*). The three exclusion criteria refer to: (1) events outside the researchers' control, (2) fixed criteria (i.e., thresholds that are independent from observations in the sample), and (3) data-dependent criteria (i.e., thresholds relative to observations in the sample). The three levels refer to: participant, trial, and stimulus. For each data exclusion criterion, participants indicated how important they thought its usage to be for accuracy in analyses and interpretation (5-point between *not at all important* and *extremely important*). After answering questions about how often they use data transformations, report transformations, and report the order of pre-processing actions, participants were asked to indicate from 0 to 100 how likely it would be for

another researcher to reproduce their analyses. The participants could also indicate reasons for (not) reporting pre-processing actions in their manuscripts.

Participants were then shown the proposed checklist in the form available at the time (see https://osf.io/3qanp) and were asked to share any final thoughts and/or concerns about the proposed items to ensure feedback on potential areas missed by the proposed checklist. The list of pre-processing actions and the corresponding checklist were initially drafted by M. P. based on previous surveys of the literature (Kerr et al., 2017; Primbs, Holland, et al., 2022) and the present literature review (*Figure 1*). The draft checklist was subsequently refined in consultation with the full team before being implemented in the survey. The current version of the checklist (*Table 2*) was improved based on survey results and suggestions from the review team.

Table 2. Checklist for Reporting Pre-processing Decisions

Section	What to report?	Examples and suggestions
General	Order	"The reporting order reflects the data pre-processing order." or "Only trials followed by a correct response were incorporated in the reaction time (RT) analyses (…) (please note that error rates were arcsine transformed prior to the analysis to approximate normal distribution). (...) Subsequently, possible decade as well as five break effects were computed as follows for each participant individually: first, the logarithm (ln) of the averaged response latencies per experimental number pair (both orders collapsed; e.g. 3:5 and 5_3) was calculated. Then, a logarithmic function (…) was fitted to these individual data. (…) Afterwards, the (…) the residuals were computed by subtracting the predicted values from the actual logarithm of the RT. Finally, these residuals were standardized to a mean of 0 and a standard deviation (SD) of 1 (…)." (Domahs et al., 2010)
	Transparency (e.g., reporting discrepancies between pre-registered and actually used pre-processing pathways)	"The deviation between planned and executed pre-processing actions has been addressed in section X." or "Our confirmatory analyses do not deviate from the pre-registered procedure. All datasets and the analysis code are available for download on the OSF." (Primbs, Holland et al.,

Section	What to report?	Examples and suggestions
		2023)
	Theoretical or empirical justification for chosen pre-processing actions	"Theoretical or empirical justification for chosen pre-processing actions has been provided in the respective sections of the present manuscript." or "RTs ≤ 100 ms we removed as they reflect implausible cognitive processing of the Go signal (Gabay & Behrmann, 2014 as cited in De Pretto et al., 2021)."
Participants	Total number of participants collected	"A group of 16 participants (five women, 11 men, 18–50 years of age) participated in this experiment. They all had a normal (or corrected-to-normal) vision and gave their informed consent." (Burle et al., 2014)
	Total number of participants excluded per reason for exclusion (participant-level data exclusion)	"After application of our pre-registered exclusion criteria, a final sample size of 155 participants remained. Please note that most excluded participants (n = 102) did not actually complete the experiment – they failed the attention check presented during the instructions and were directly forwarded to the end of the experiment, skipping all experimental trials. The other participants were removed because they were too slow (3SD from the mean reaction time; n = 3) or made too many mistakes (n = 2)." (Primbs, Rinck, et al., 2022)
	Total number of participants (per condition) included in final analysis	"As pre-registered, we recruited 100 participants who fulfilled our inclusion criteria (at least 18 years old, fluent in English) and completed the online study via the research platform Prolific. Of those, 9 fulfilled our pre-registered exclusion criteria (...), leaving the data of 91 participants to be analyzed." (Rinck et al., 2022)
Materials	Number of trials (per condition, per participant)	"Eight models (...) were selected from the Radboud Faces Database on the basis of how well their emotional expressions were recognized in a validation study (RaFD; Langner et al., in press). (...) Subsequently, pictures of the three emotional expressions central to this experiment were selected per model, namely happy, sad, and angry. This resulted in a total of 24 pictures: three expressions x two ethnicities x four targets (models) per ethnicity. (...) Each experimental block consisted of sixteen pictures randomly displayed five times, resulting in 80 trials per experimental block." (Bijlstra et al., 2010)
	Stimuli-level data exclusion: how many stimuli were excluded for which reason?	"Criteria for item selection were high discriminatory power, high convergent validity with openness for experience, as well as content validity, based on expert judgment." (Mussel et al., 2011)
Analysis	Proportion of trials included in final analysis /	"For correct RTs, a mean and standard deviation were calculated for each subject within each SOA and session,

Section	What to report?	Examples and suggestions
	proportion of trials excluded for a particular reason	and any RT greater than 3 SDs above or below the mean for that subject during that SOA and session was identified as an outlier. This eliminated 1.7 % of both the lexical decision and naming RTs." (Hutchison et al., 2013)
	Type of trial-level data exclusion	"For the minimum threshold, we varied the response time cut-off from 0ms to 300ms in steps of 50ms, resulting in 7 levels. For the data-based outlier trimming method we varied the number of median absolute deviations from the median (Leys et al., 2013) from 1 to 3 in steps of 0.5 or applied no data-based trimming, resulting in 6 levels." (Primbs, Rinck, et al., 2022)
	Data transformation	"Separated repeated measures analyses of variance (rmANOVAs) were conducted to investigate one-session and two-week training effects on median RTs, arcsine-square-root-transformed error rates, and inverse efficiency scores." (Soltanlou et al., 2018)
	Data aggregation	"Most often the mean RT within each trial type is calculated (...). Researchers may opt to use the median RT instead. I included both options." (Parsons, 2022)
	Reason for choosing a specific method	"Logarithmic fitting was chosen first because of evidence for a logarithmically compressed quantity representation." (Domahs et al., 2010) or “Using the mean or the median as central tendency statistics alone may conduce to biases and increase the risk of falsely rejecting null hypotheses (Morís Fernández & Vadillo, 2020; Rousselet & Wilcox, 2020). (...) Among other alternative approaches (Ging-Jehli et al., 2021), using a theoretical distribution to describe and compare the shapes of different RT distributions has been proposed (Castellanos et al., 2006; Van Zandt, 2000). The most widely used theoretical distribution in ADHD research is the ex-Gaussian distribution.” (Bella-Fernández et al., 2023)

4.2. Results

The average number of years of experience performing research that uses RT data was 11 (median 9.5 years, distribution shown in *Figure 2*). Participants also reported the number of years of direct involvement in collecting RT, analyzing RT, open science, experiment coding, and analysis coding. *Figure 3* shows that many of the 66 participants in the analysis sample have reported extensive experience with various aspects of an RT study. Specifically, more than half

of participants have at least four years of experience setting up experiments, and collecting and analyzing RT data.

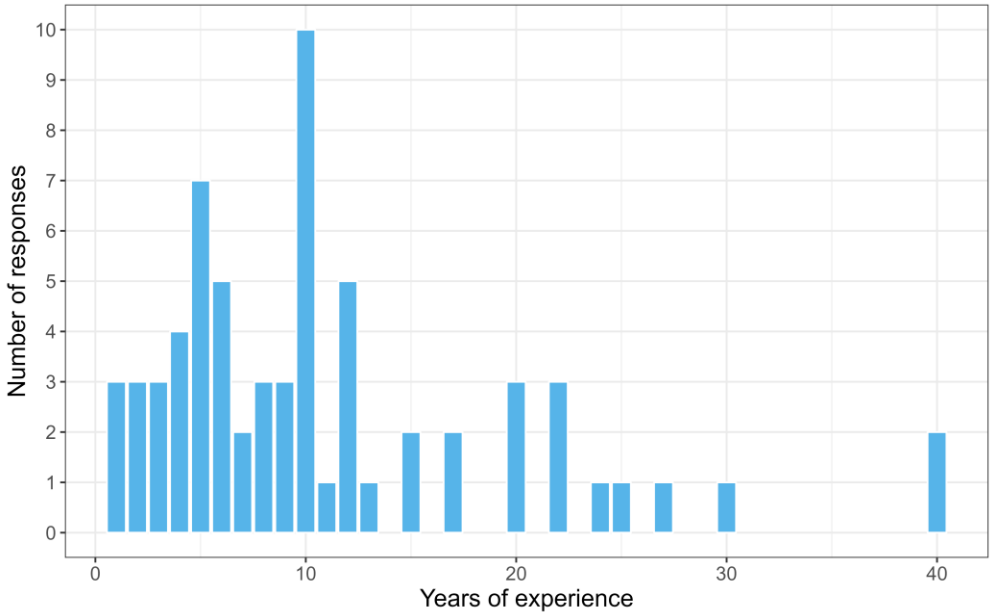

Figure 2. Reported years of experience with RT research.

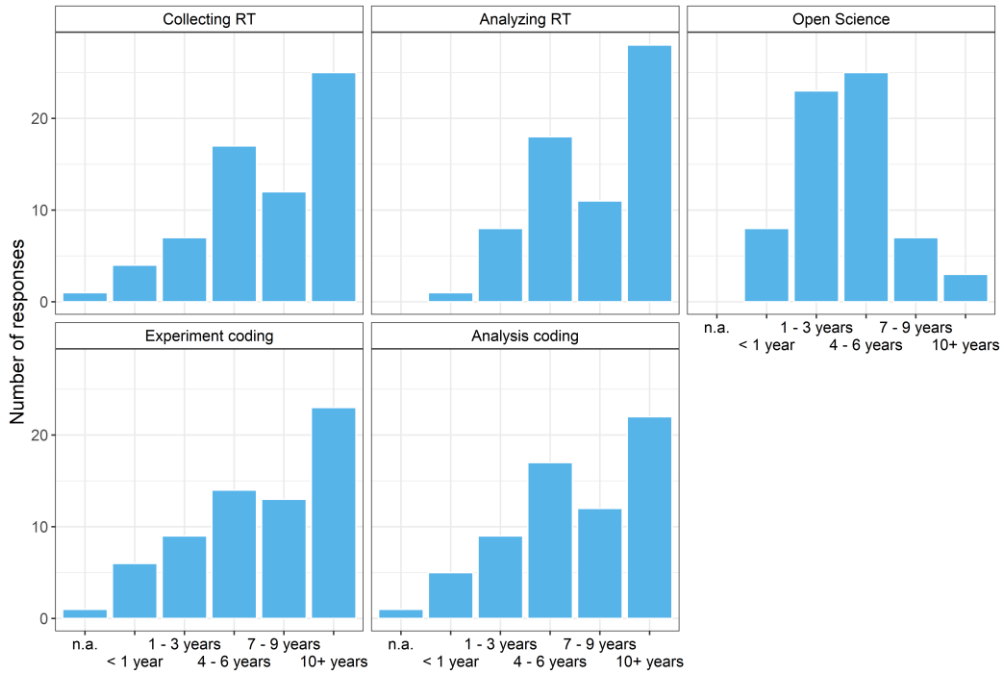

Figure 3. Number of years of direct involvement in collecting and analysing RT data, experiment and analysis coding for RT data, and open science practices. *Note*: NA responses indicate missing data.

4.2.1. Checklist Item 1: Exclusions Due to Events Beyond the Researcher's Control.

RT studies involve the use of hardware and software that present stimuli to participants and that record the time it takes for them to perform an action (e.g., press a button to indicate their response). Hardware or software malfunctions therefore impact the collection of RT data. For example, the response button can get jammed and, as a result, the recorded RT is longer than normal, or the RT cannot be recorded at all because the button press is not detected. When such an event happens during a data collection session, this issue determines how many trials within that session are impacted. When visual stimuli are presented to participants via a web interface, their loading time is impacted by the size of the image files (i.e., larger files take longer to be displayed on screen).

While there are various ways of preventing such events, if they do happen, then researchers can choose to exclude all observations of impacted participants, only those observations in trials where stimuli took longer to load, or all observations where stimuli with longer loading times were shown. While most respondents excluded participants due to events outside the researcher's control (*Figure 4*), few excluded stimuli or trials. It is important to note that there are several survey participants who report always excluding observations (participants, stimuli, or trials) due to events beyond their control, which shows that such events happen often enough. Of those survey participants who excluded observations at least sometimes, the large majority indicate that they always report doing so in the paper and that they also include the exact number of excluded observations.

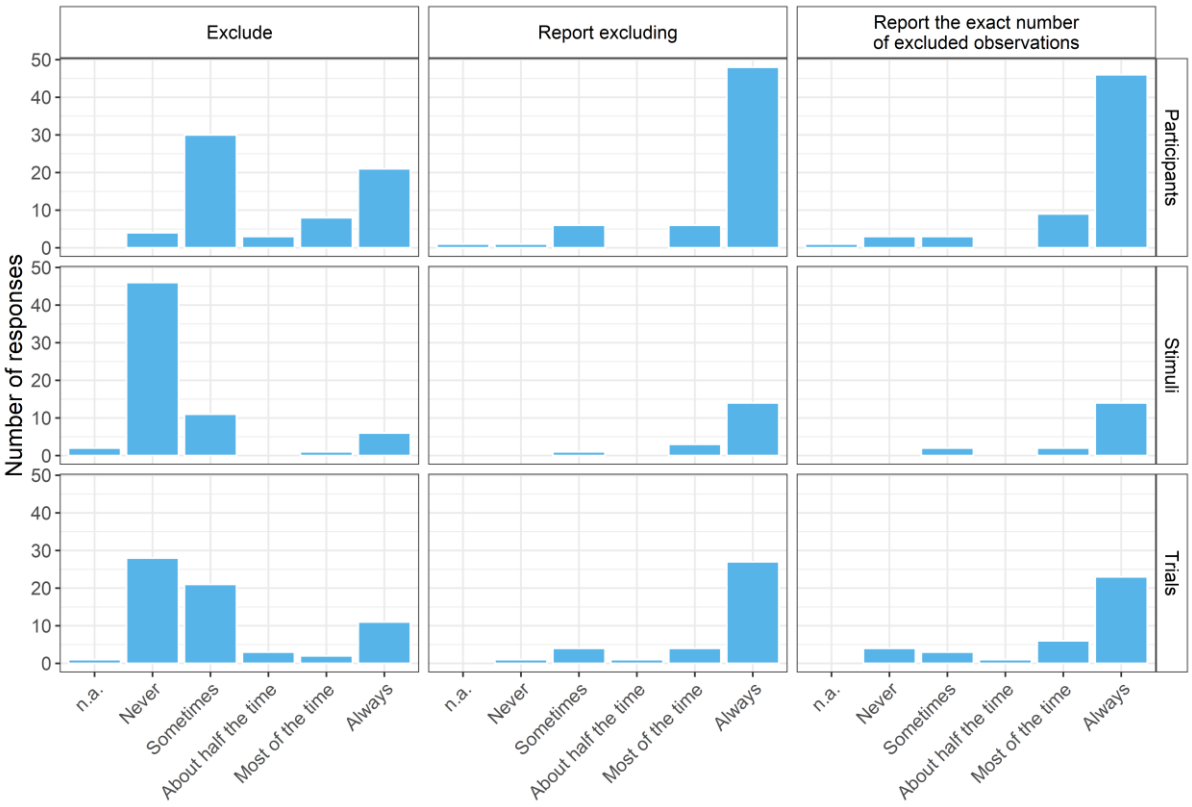

Figure 4. Frequency of excluding, reporting excluding, and reporting the exact number of participants, stimuli, and trials due to events beyond the researcher's control. *Note*: Responses in 'Report excluding' and 'Report the exact…' are from survey participants who reported excluding observations at least sometimes.

4.2.2. Checklist Item 2: Exclusions Based on Fixed Criteria.

For certain types of RT studies, it is possible for researchers to set exclusion criteria ahead of data collection, based on the literature and results in previous studies. For example, researchers can set a minimum RT based on expectations about processing time for stimuli used in the study, with the underlying assumption being that responses with an RT below this threshold are either made by accident (e.g., the participant pressed the button by mistake) or by participants who are not sufficiently engaged in the task (i.e., participants who are bored or impatient). Similar to responses for the first checklist item, most respondents used such criteria to exclude participants, but fewer excluded trials or stimuli as shown in *Figure 5*.

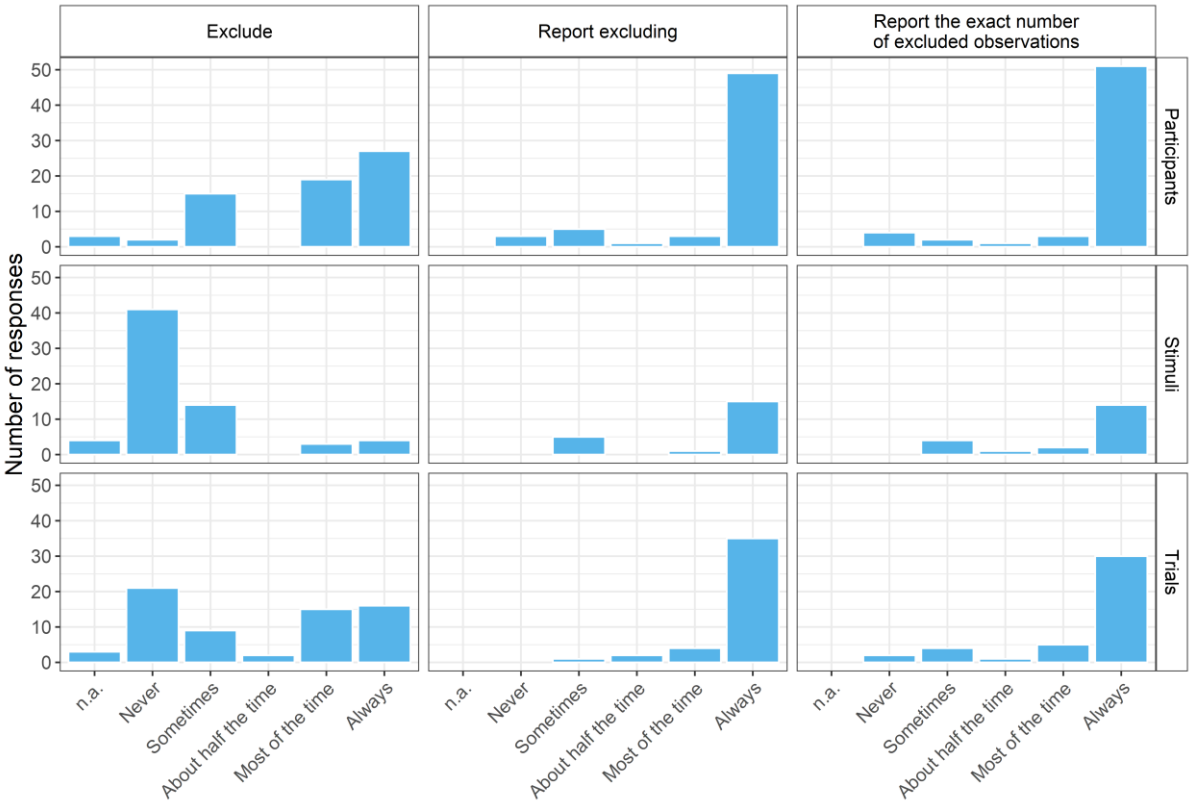

Figure 5. Frequency of excluding, reporting, and reporting the exact number of participants, stimuli, and trials due to fixed criteria. *Note*: Responses in 'Report excluding' and 'Report the exact…' are from survey participants who reported excluding observations at least sometimes.

4.2.3. Checklist Item 3: Exclusions Based on Relative Criteria.

Relative criteria such as the mean and standard deviation or the median and median absolute deviation were used more frequently to exclude participants, rather than trials or stimuli (*Figure 6*). Of those researchers who exclude participants at least sometimes, two indicated that they never report doing so. Any such exclusions should be reported in detail if they happen, as that allows the reader to evaluate how appropriate this action was and how likely it was that it would impact the validity of the results and conclusions presented in the paper.

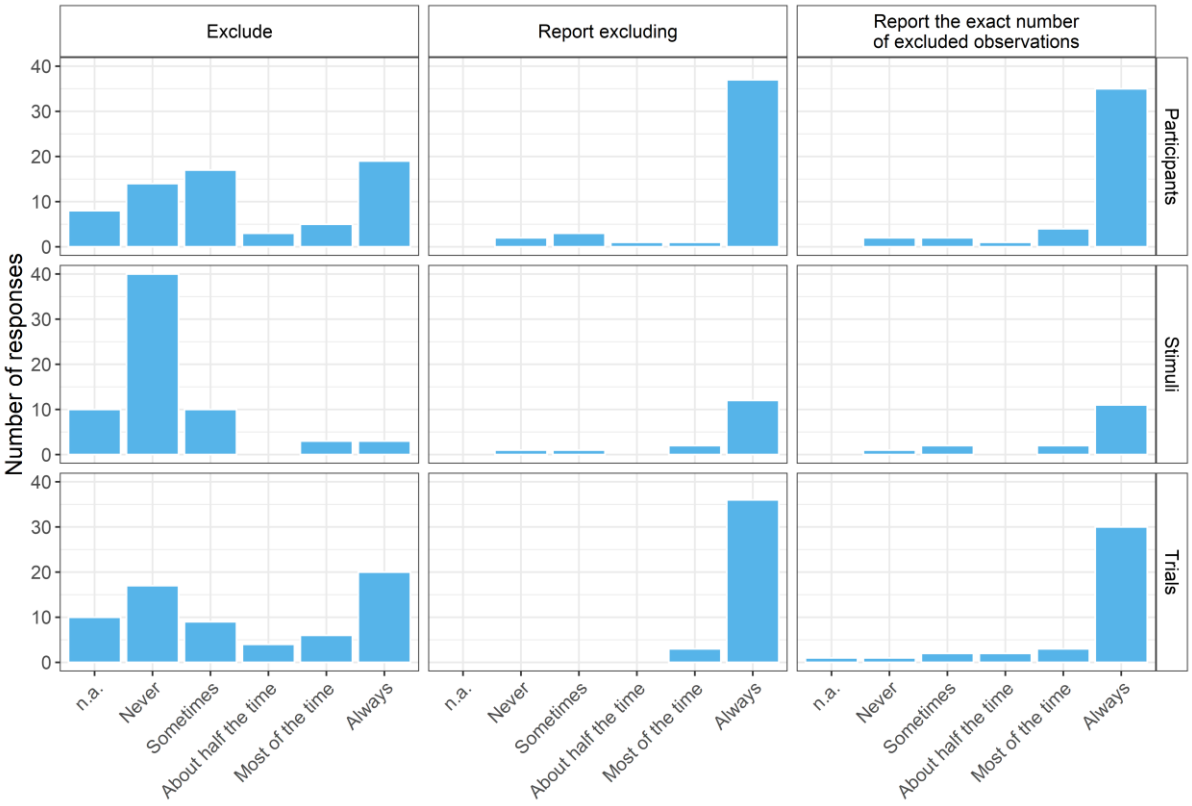

Figure 6. Frequency of excluding, reporting, and reporting the exact number of participants, stimuli, and trials due to relative criteria. *Note*: Responses in 'Report excluding' and 'Report the exact…' are from survey participants who reported excluding observations at least sometimes.

One reason why researchers might not always report excluding observations is that about one third of participants who completed our survey perceive those exclusions to be 'not at all' or only 'slightly' important for the accuracy of their analyses and interpretation (*Figure 7*). Interestingly, *Figure 7* also shows that exclusions due to events beyond the researcher's control, as well as exclusions based on fixed and relative criteria were deemed equally important for accuracy in analyses and interpretation. However, one might argue that outliers resulting from external events such as a crash of the recording system should be classified as a missing completely at random mechanism, while the question whether an observation exceeds a fixed or relative cut-off is less trivial, because we cannot truly determine whether the latter stem from a

different data generating process than the one targeted with the experimental manipulation. Perhaps the sample undertaking the survey was very aware of dangers resulting from non-transparency, which led to participants classifying more outlier exclusion procedures as critical than other researchers would.

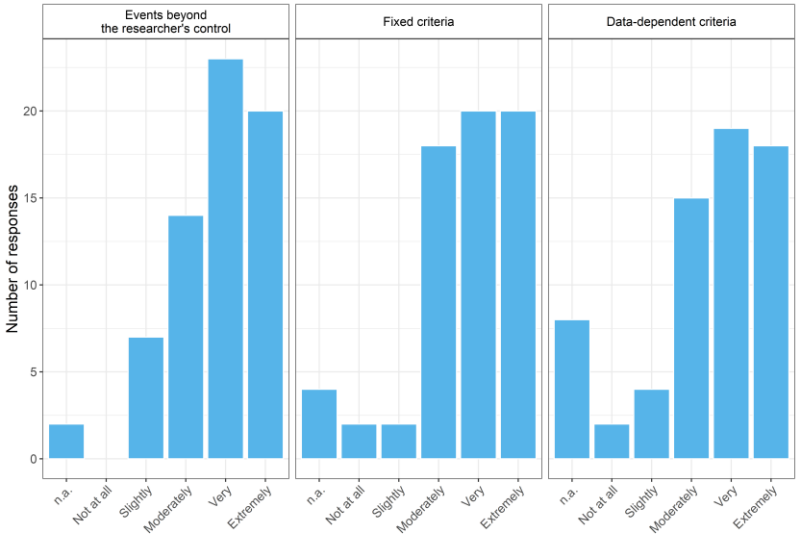

Figure 7. Responses to the data exclusion question: 'How important do you think this processing action is for accuracy in analyses and interpretation?' *Note*: Each panel represents one of the data exclusion checklist items.

4.2.4. Checklist Item 4: Data Transformations, Scoring, and Aggregation.

Besides data exclusions, an important aspect to report is the transformation, scoring, or other aggregation of the data. RT data is often modelled or analyzed with a log, or inverse transform applied to the responses, as RT data is known to have skewed distribution. Further, data may be aggregated by creating by-participant or by-item averages for conditions or groups in the study if multiple trials are used. Last, RT data may be further computed into a study specific scoring, such as difference scores between conditions in a priming or Simon task study. Each of these transformations can potentially impact the final results of a study and should be described

within the paper. In the first panel of *Figure 8*, the majority of survey participants reported using one of these methods at least sometimes, half the time, most of the time, or always in their analyses, showing that these practices are common. Nearly all survey participants indicated that they always report these data transformations, while a few participants indicated they only sometimes, half, or most of the time report data transformations (*Figure 8* middle panel). As shown in the left panel of *Figure 9*, most survey participants believe that reporting data transformations were moderately to extremely important for accuracy in analyses and interpretation.

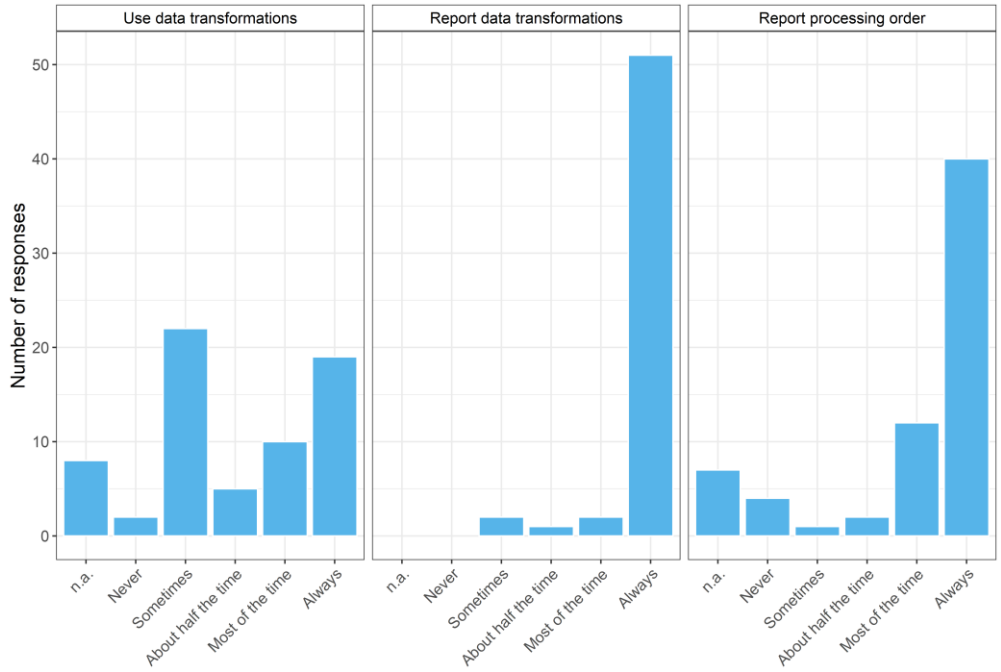

Figure 8. Responses to 'How often do you use and report transformations or processing order' from survey participants who use data transformations at least Sometimes.

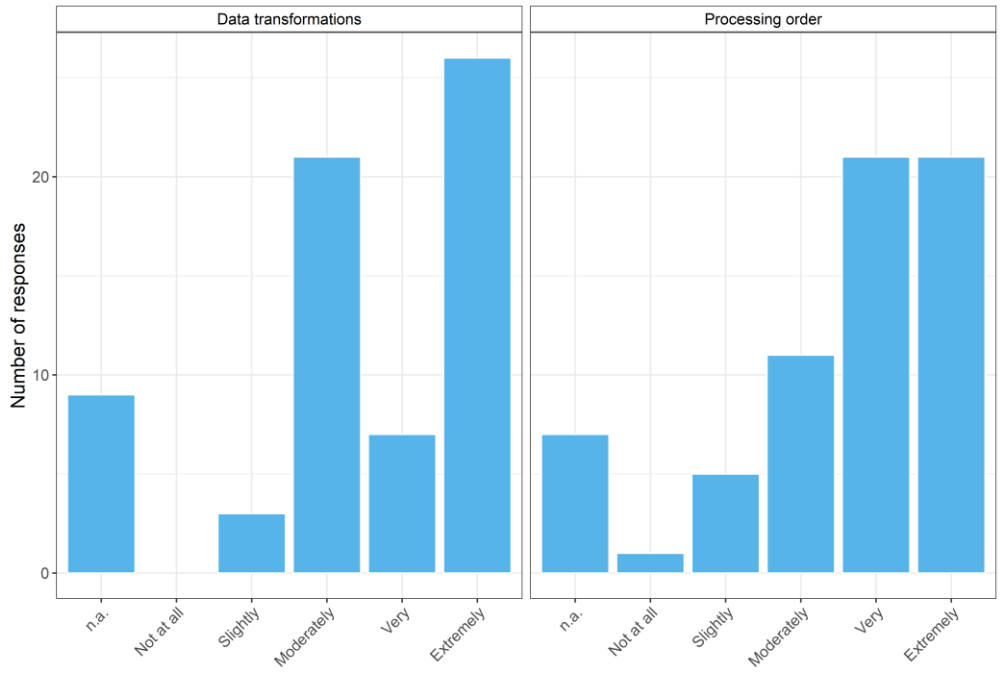

Figure 9. Answers to 'How important do you think reporting data transformations and processing order are for accuracy in analyses and interpretation'.

4.2.5. Checklist Item 5: Data Processing Order.

The last checklist item involves ensuring the data processing order is explicitly reported in the manuscript. Exclusion of outlier participants or trials could be based on raw scale data or potentially transformed data, and reporting the actions used to reach the final RT data allows for potential reproducibility of the results. *Figure 8* (right panel) indicates that survey participants nearly always report the data processing order, and *Figure 9* (right panel) shows they believe that this information is very or extremely important for analysis interpretation. This result was in stark contrast with the results from the survey of the literature, which indicated that most manuscripts do not report sufficient detail to understand the data processing order, as it could only be guessed from the order of listing actions which does not allow for a robust reproduction (see **Supplementary Materials A**).

4.2.6. Reproducibility.

As shown in *Figure 10*, survey participants were very positive about the likelihood of reproducing the analyses if a researcher used the proposed checklist (M = 83.24, SD = 14.86). However, this estimate of reproducibility, which was elicited towards the end of the survey, still shows substantial variability. One potential explanation is that among survey respondents who indicated that they exclude observations at least sometimes there is variability in how often they report the exact number of excluded observations - which is important information when trying to reproduce the study results using the original data. We find a positive association between the stated likelihood of reproducing analyses and the frequency of reporting the exact number of excluded observations (i.e., Pearson's correlation of .31, p-value = .02). The relatively high expectations regarding the likelihood of reproducibility could be biased by the sample's belief that reporting should have positive effects. Alternatively, it may be due to their firsthand experiences with the benefits of thorough reporting, given the sample's substantial experience with open science practices (4-6 years and beyond). Additionally, the sample generally reported working with coding and/or open-source software to collect RT: e-Prime (10.2%), PsychoPy (10.2%), Gorilla (3.3%), jsPsych (3.3%), and others (participants could list multiple responses, and percentages represent total reported out of total options listed). For data analyses, survey participants listed using R (21.9%), SPSS (7.7%), and JASP (4.4%) most frequently, which allow researchers to share their analysis code to facilitate reproducibility of results.

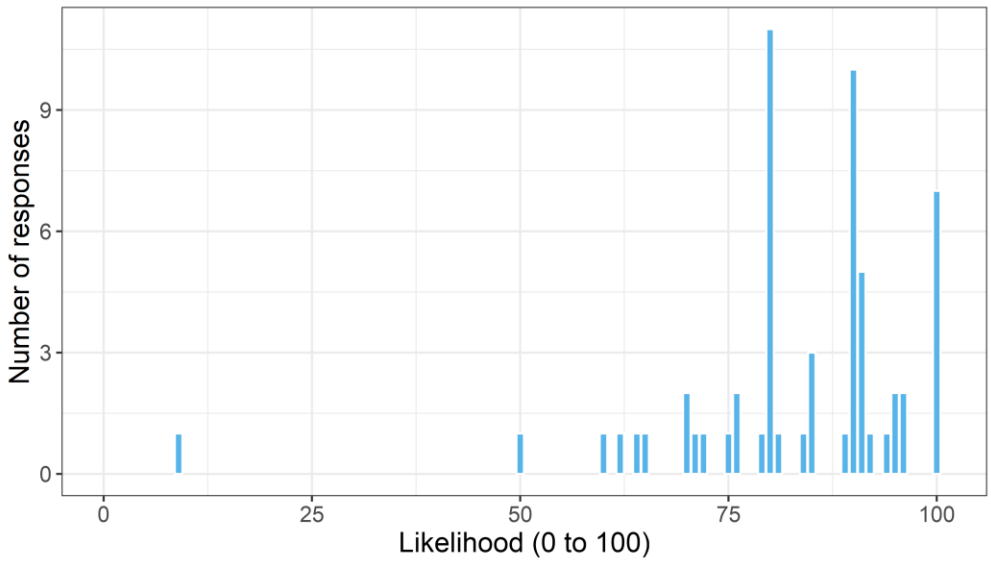

Figure 10. Responses from participants after reviewing the checklist indicating how likely they think another researcher could reproduce their analyses.

4.2.7. Attention Check.

Due to technical issues, two attention survey check items were not presented to the participants:

- *Please mark sometimes for this item.*
- *Please mark most of the time for this item.*

Therefore, we are unable to provide estimates regarding the participants' level of attentiveness. However, we can assume that all included participants responded appropriately, considering that they answered all the items, including the free text items, and open-ended responses were checked for appropriateness.

4.2.8. Final Thoughts.

We used the feedback from the expert survey and the commentary from the community (e.g., reviewers' recommendations, checklist presentation at Society for the Improvement of

Psychological Science (SIPS) 2022 and 2023 conferences) to further improve our checklist. The expert survey responses in full (corrected for spelling errors) can be found in the **Supplementary Materials C**. First, it was drawn to our attention that our survey was missing an "NA" answer, "forcing" some participants to choose "never exclude" as a response. In fact, some participants may have never encountered situations where they needed to exclude observations (e.g., due to events beyond their control). As a consequence, we added an NA option to our checklist. Second, based on the provided feedback, we added three general categories to the checklist, which are not section-specific: "order of the pre-processing actions", "transparency about planned and executed pre-processing actions", and "theoretical or empirical justification for chosen pre-processing actions". For each of the three general categories, we provided a hypothetical example, such as

- *"The reporting order reflects the data pre-processing order"*,
- *"The deviation between planned and executed pre-processing actions has been addressed in section X"*, and
- *"Theoretical or empirical justification for chosen pre-processing actions has been provided in section Y"*.

4.3. Discussion

Comparing our findings from the expert-consensus survey to those from the systematic literature search shows a contradiction between survey respondents claiming to frequently report all pre-processing actions and the number of excluded observations, and the apparent lack of reporting in the papers on the Simon effect that we have reviewed. A generous explanation of this contradiction could be that research practices and journal guidelines have significantly changed in the past couple of years in the direction of complete and transparent reporting of RT pre-processing. The results of our literature review provide very little, if any, supporting evidence for this explanation. Of the 13 papers published between 2017 and 2022, only 6 report the exact

number of excluded observations for various criteria. Further, only 5 of all reviewed articles explicitly reported the order of the pre-processing actions and these were published in 1995, 2005, 2017, 2018, and 2020. Even if journal guidelines were to require transparent reporting, we argue that there are important benefits to be gained from having a set of reporting guidelines that researchers can follow. Another reason might be the sample of researchers answering the survey and of those who conducted the original experiments included in the literature search. First of all, there might be differences in the levels of experience with RT pre-processing, as well as conventions in different working groups. Second, the sample of survey respondents might be selective in the sense that they voluntarily participated in the survey because they regard transparent reporting of RT pre-processing as an important topic.

5. Multiverse Analysis of Pre-processing Actions in the Simon Effect

We believe that researchers should report the information presented in Table 2 simply for transparency and reproducibility; however, we also demonstrate how each of the pre-processing actions may influence the results. We performed a multiverse analysis (Steegen et al., 2016) on available data from the Simon effect (Zwaan et al., 2018) using the checklist as a guide for the choice of pathways. Multiverse analyses allow a researcher to examine the influence of pre-processing, aggregation, and other choices on the results of the study, often to determine the robustness or sensitivity of a specific result. In our analysis, we examine how the application of the steps reported in our guidelines, as well as the order in which they are applied influence the size of calculated effect, power, and choice of raw or standardized effect size.

5.1. Method

5.1.1. Data.

The data from Zwaan et al. (2018)'s Simon Task was accessed from https://osf.io/x8pha. In this study, participants were recruited to complete classic cognitive psychology tasks including a replication of Craft and Simon (1970)'s task examining spatial compatibility. Participants completed two waves of the study to examine the influence of repeated participation and previous knowledge of the experiment on final results. In our multiverse analysis, we used the first wave of the data collection to most closely mimic the original Simon task. Participants were shown stimuli (i.e., red square, blue square, yellow circle, green circle) on the left or right side of the screen and asked to respond to the presentation as quickly and accurately as possible. Each stimulus was tied to a specific response using the left or right side of the keyboard. Stimuli were classified as congruent (e.g., stimulus response matched the side of the screen presented - press the left key and appeared on the left side of the screen) or incongruent (e.g., stimulus response was the opposite side of the stimulus presentation). The Simon effect occurs when the congruent responses are faster than the incongruent responses. We used the same stimulus condition (i.e., they saw the same stimuli for waves 1 and 2) for multiverse analysis which originally resulted in an effect size of d_Z = 1.30[1]. Each participant completed 92 trials for analysis, n_{trials} = 46 each in the congruent and incongruent conditions.

5.1.2. Pathways.

We used *Table 2* to create the pre-processing pathways for analysis of the Simon data from Zwaan et al. (2018). *Tables 3 and 4* summarize the different pathways.

---

[1] While the data online indicates it is the *raw* data, the original sample size tested was larger (total *N* across conditions = 172). The published sample was then reduced to *N* = 160 to create counterbalanced groups across all study conditions using the following rules: 1) participants with less than 80% accuracy were excluded, 2) participants with less than 10 percent accuracy in memory tasks were excluded, 3) participants with mean reaction times greater than the mean plus 3 standard deviations for their group were excluded, and 4) the last participants to complete the study were excluded to create equally balanced groups after exclusions 1-3 were applied.

Exclusions. One rule was used for each exclusion application, as shown in the table. Using this example data, stimuli exclusions were unlikely, but these exclusions were regarded as reasonable examples one might choose for their own data. All other rules were generated based on the data or other choices that a researcher might make for reaction time data.

Number of Exclusions. Pathways were generated using none of the above exclusions, only one exclusion, two exclusions, three exclusions, and four exclusions. We used up to four exclusions as this value represents the most that researchers generally mention in their results (see expert survey results), four exclusions allowed us to examine the impact of order on the results, and using up to four exclusions, but not up to nine exclusions, kept the number of results to 3610.

Order. The orders were created by analyzing every combination of possible order that did not repeat a specific data exclusion. Therefore, no exclusion was used twice in one pathway (e.g., 1-2-1 was not allowed); however, different orders of different exclusions could be used (e.g., 1-2-3, 3-2-1, 1-3-2 were all allowed).

Data Transformations. Analyses were conducted on both raw and log-transformed reaction times. Note that the transformation always occurred after the other data exclusions.

Data Aggregation. Data were aggregated based on the two analysis pathways chosen for this multiverse analysis. For *t*-test analysis, data were first averaged by participant and condition, then averaged by condition during the calculation of paired samples *t*-tests. For multilevel model analyses, data were not aggregated. The *nlme* (Pinheiro et al., 2022) package was used to analyze the model using the condition (i.e., congruent, incongruent) to predict raw or log-transformed

reaction times. The random-intercept of participant was included to control for correlated error of repeated measurement of the same person.

Standardized Effect Size. One issue with reporting in repeated measures analyses is the lack of transparency indicating *which* effect size calculation was used. Given the raw data, we were able to determine the original study calculated standardized effect size using:

$$d_z = \frac{M_D}{SD_D}$$

where M_D represents the mean of the difference scores between incongruent and congruent conditions, and SD_D represents the standard deviation of the difference scores. This effect size tends to be upwardly biased due to the reduction in variance by subtracting conditions (Dunlap et al., 1996; Lakens, 2013), and therefore, we also present results using d_{av} (Cumming, 2012) which is calculated by:

$$d_{av} = \frac{M_1 - M_2}{\frac{SD_1 + SD_2}{2}}$$

where *M* and *SD* represents the mean/standard deviation from each condition. Note that standardized effect size calculations only occur on aggregated data.

Raw Score Effect Size. The mean difference between congruent and incongruent conditions was used for the raw score effect size for *t*-test analysis aggregated data. The *b* coefficient from the multilevel model analysis was used to calculate the raw score effect size on the non-aggregated data. We included these two effect size pathways, even though not necessary in our reporting checklist, to show if the choice of other suggested reporting guidelines differentially impacts what a researcher might present as the final "result" in the study.

Table 3. Multiverse Pathways (Exclusion)

Manipulated path	Options	Criteria
Exclusion	Participant external events	To mimic data loss from random participant events or collecting data from the incorrect population, we randomly removed 3% of participants.
	Participant outlier: Fixed	Participants were excluded if they did not have 80 of the 92 total possible trials. The trial minimum included both correct and incorrectly answered trials.
	Participant outlier: Data	Participants were excluded if they did not achieve at least 85% accuracy on only included trials, and this criterion was selected by examining a histogram to find the break in accuracy between high and low performers.
	Trial external events	We randomly removed 3% of trials to mimic computer issues or other events that might affect trials individually.
	Trial outlier: Fixed	Trials were excluded that did not meet a minimum reaction time of 250 ms.
	Trial outlier: Data	Trials were excluded that were longer than the overall sample mean reaction time plus two times the standard deviation of all reaction times.
	Stimuli external events	Given the small number of stimuli each participant encountered (i.e., 2 per participant, 4 overall), we randomly excluded 3% of trials to mimic random removal due to external events.
	Stimuli outlier: Fixed	Stimuli that do not achieve at least 50% correct will be excluded, as this value represents chance performance.
	Stimuli outlier: Data - deviations	Stimuli that do not achieve at least the overall mean accuracy for stimuli plus two standard devaitions of stimuli accuracy.
Number of exclusions	0, 1, 2, 3, 4 options	No exclusions are applied, one exclusion applied, etc.
Order of exclusions	All combinations with no repeating exclusions	Participant external-participant outlier, Participant outlier-participant external, etc.

Table 4. Multiverse Pathways (Performed Actions)

Manipulated path	Options	Performed actions
Data transformation	No transformation or transformed reaction times	Log transformation.
Data aggregation	Aggregated or no aggregation	Data were first averaged by participant, then averaged by condition when aggregated.
Standardized effect size	Aggregated data	Effect size with standard deviation of the differences (dz), effect size with average standard deviation (dav).
Raw score effect size	Aggregated data or non-aggregated data	Mean difference between congruent versus incongruent, congruent versus incongruent coefficient from multilevel model.

5.2. Exploratory Questions

1) Does the order of data exclusions impact the results of the analysis?

To answer this question, we will present all results for the two calculations of standardized effect sizes (Question 3) and raw effect sizes (Question 4). If order of processing is not important, we would expect to find the same effect size for processing orders that included the same applied exclusions.

2) How do differences in exclusions applied and analysis choice influence power?

The original effect size found for this study (i.e., d_Z = 1.30) indicates that likely all effects will be significant using $\alpha < .05$. Therefore, we will comment on the effect of exclusions on power by displaying the overall sample size (for aggregated data and *t*-tests) and overall trials included (for raw data and multilevel models) to denote differences in change in sample size used to calculate degrees of freedom for the statistical test.

3) What is the impact of the data exclusions, transformations, and choice of effect size on the standardized effect sizes?

In this question, we will compare the standardized effect sizes d_Z and d_{av} for the impact of data exclusion and transformation choice on the final reported effect size. We expected that d_Z would be upwardly biased in comparison to d_{av}, but it was unclear how processing and transformation would impact these effects.

4) ***What is the impact of the data exclusions, transformations, and analysis choice on the raw effect size?***

For the last question, we will visualize raw effect size (incongruent - congruent) by exclusions and transformation for each analysis type to determine the impact of each on the final raw score effect sizes in the same manner as Question 3.

5.3. Results

5.3.1. Question 1 - Order of Exclusions.

For this question, we first excluded analysis pathways that did not use exclusions or only used one exclusion. We then matched pathways that included the same exclusions with different orders (total pathways = 3610; total matches = 246). Within each matching set of exclusions, the results of the effects were subtracted (i.e., exclusion 1 order 1 minus exclusion 1 order 2, separately for standardized and unstandardized effects), and the absolute value was taken. This procedure created 36072 combinations of two to four pre-processing eliminations. *Figure 11* displays the results. If the order of processing combinations did not affect the results, the results should show zero differences for combinations. Standardized effect sizes indicated that the difference in processing order could be up to d_Z = 0.20 or d_{av} = 0.14 (M_{d_Z} = 0.02, SD_{d_Z} = 0.02; $M_{d_{av}}$ = 0.01, $SD_{d_{av}}$ = 0.01). For unstandardized effect sizes, the log RT could be up to 0.01 (i.e., 1.01 ms) while raw reaction times could be 4.49 ms different (M_{log} = 0.00, SD_{log} = 0.00; M_{raw} =

0.55, SD_{raw} = 0.52). Given these results are not zero, the order of pre-processing actions appears to influence the final results.

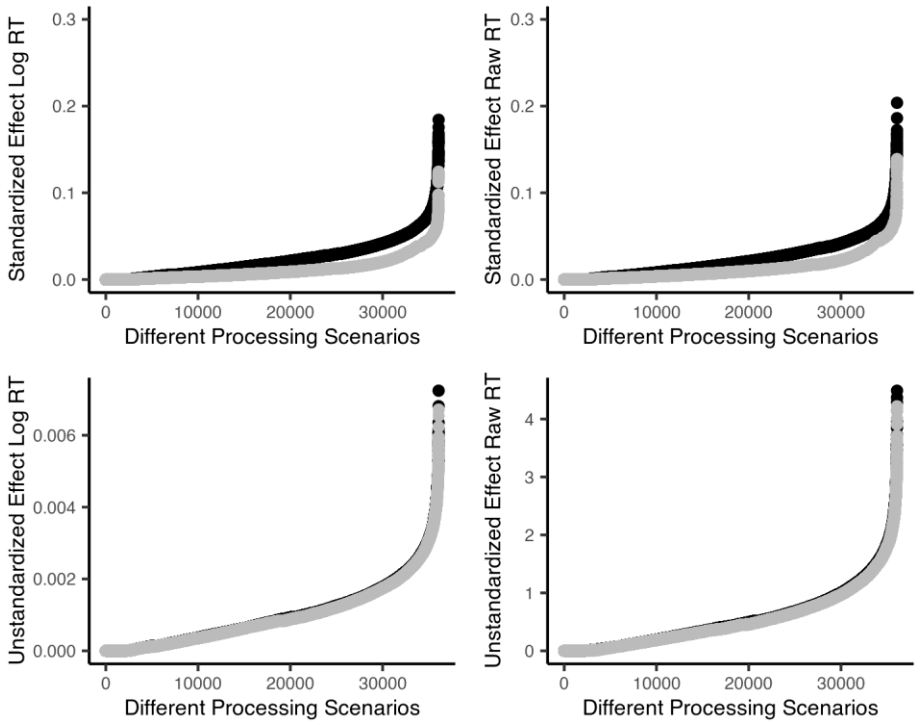

Figure 11. This figure demonstrates the effects of order on both raw and standardized effect sizes by examining the difference scores for the same exclusion criteria in different orders. These difference scores are arranged from smallest to largest to demonstrate a sensitivity plot of the effect of order. If order did not affect the results, all dots will align at zero. However, we find that different orders create different effect sizes, and thus, the difference scores are not zero. The first row represents standardized effect sizes, and the second row represents unstandardized effect sizes. The left side represents a log-score transformation on RT, and the right side indicates no transformation of RT. In the first row, the black dots indicate d_Z effect sizes and gray dots indicate d_{av} effect sizes. In the second row, the black dots indicate mean differences calculated from *t*-tests, and the gray dots indicate mean differences calculated from multilevel models.

We found the same pattern of effect size differences across standardized and raw score effect sizes, wherein the order of processing can create non-zero differences between different processing pathways. In *Figure 12*, the difference in d_Z is highlighted to determine if one particular exclusion was the reason for these differences. For example, if the exclusion of poorly

performing participants was the main improvement in effect size — because of reduced error and noise in performance — we might expect to see this exclusion have a higher standardized effect size difference than the rest of the exclusions. By examining the box plots in *Figure 12*, the largest differences appear to occur with *more* processing actions, as there are more opportunities to remove participant/trial/stimuli outliers. However, within each number of processing actions, we do not seem to find a consistent pattern of one exclusion that creates the largest or least differences. Therefore, the order of processing does change results, but it cannot be necessarily attributed to a single exclusion type suggested in our checklist.

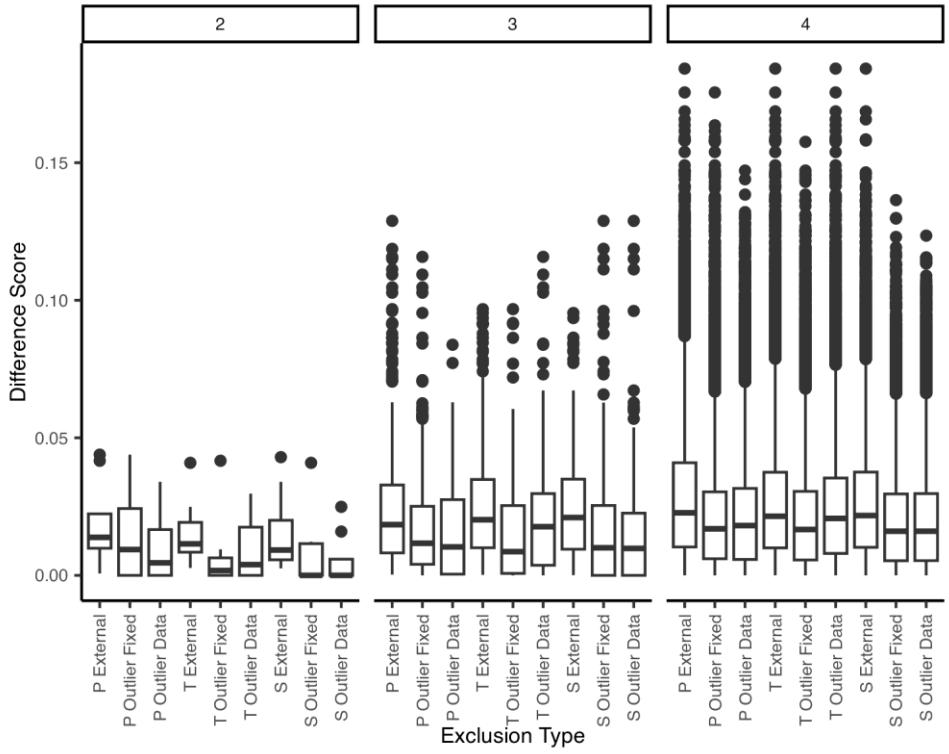

Figure 12. Difference in effect size scores plotted by exclusion type suggested in our checklist. Each panel represents the number of exclusions applied on the data before effect size calculation. P = Participant, T = Trial, S = Stimuli.

5.3.2. Question 2 - Power.

All statistical tests indicated a significant non-zero effect using $\alpha < .05$. The number of participants and trials can/did decrease with increasing number of exclusions; thus, resulting in lower power via degrees of freedom when more exclusions were applied. However, exclusions are often applied to reduce the amount of error and noise in the study, and as shown above, these inclusions may increase power/effect size when applied due to the reduction in the error term. With this in mind we will further investigate the impact of pre-processing steps on the power of the study. Therefore, we demonstrate the differences in sample size in aggregated analyses (*t*-test) versus differences in trial level analyses (multilevel modeling). In *Figure 13*, the results on sample size are displayed, while *Figure 14* shows the results on trial level analysis. In these graphs, we demonstrate the pairwise-combinations of exclusions to explore the effect of exclusion pairs on the data. The heatmap cells represent the maximum change across all combinations for that step (i.e., when I apply X exclusion and then Y exclusion, what is the change in the number of trials/participants between those two pre-processing actions?).

In participant level analyses, the exclusions related to participants will have the strongest impact on power, given that these exclusions will remove an entire participant from the data. Participant level exclusions have the largest impact on power when trial-level exclusions are first applied because in this scenario, trials are excluded for fixed or data focused reasons, and then participants do not have enough data to be included in the study or become an outlier for other reasons. The reverse processing order (i.e., excluding participants and then trials) does not affect the data in the same way. For trial level analyses, we see a similar pattern that participant level exclusions have a larger impact on power — generally because removing an entire participant means the removal of all their trials. However, the trial level and stimuli level exclusions also

show changes in the number of trials across pairwise combinations, but again, we find that the order of exclusion does not show the same exclusion numbers (i.e., 1-2 versus 2-1 exclusions do not show the same number of trials removed or the top and bottom half the figure would be colored in the same pattern). Overall, trial level analyses are likely to have more power, due to larger degrees of freedom for the focal statistical test, but it is important to note that the exclusions show different effects on the overall power based on their application and order.

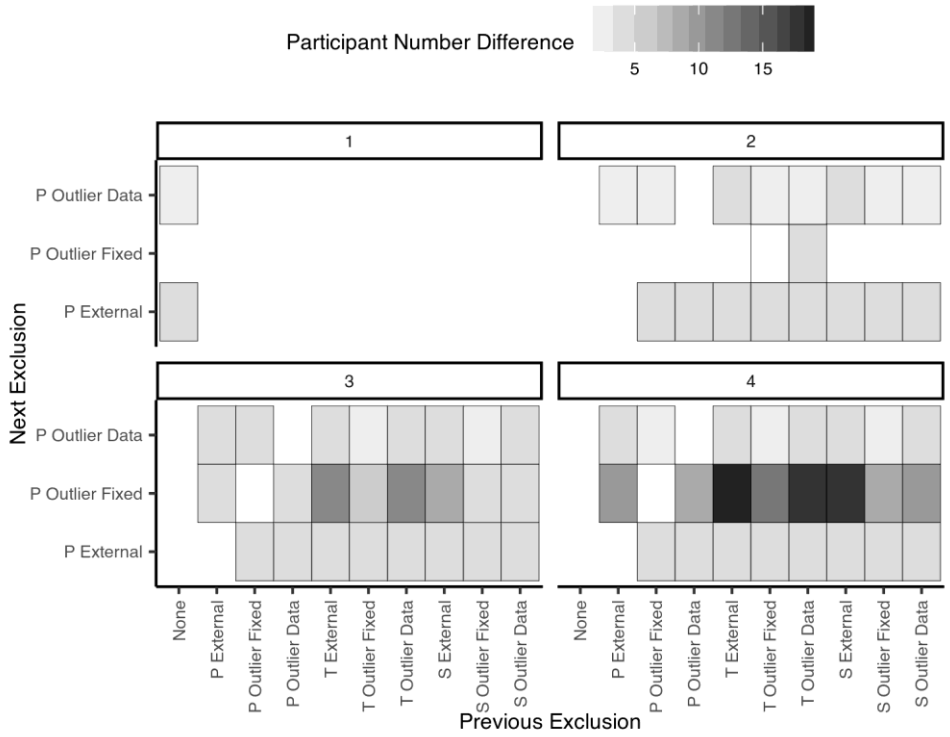

Figure 13. Number of participants included in the participant-level analysis (*t*-test) based on the order and pairwise-combination of exclusions. The panels represent the number of exclusions present. The x-axis represents the previous exclusion applied, while the y-axis represents the next exclusion to be applied. The heat color represents the change in number of participants for that combination. For example, one tile may show that after participant external exclusions were used, then participant outlier fixed may show a change of up to 10 participants. These cells represent the maximum change across all combinations of pathways. P = Participant, T = Trial, S = Stimuli. All zero values have been excluded.

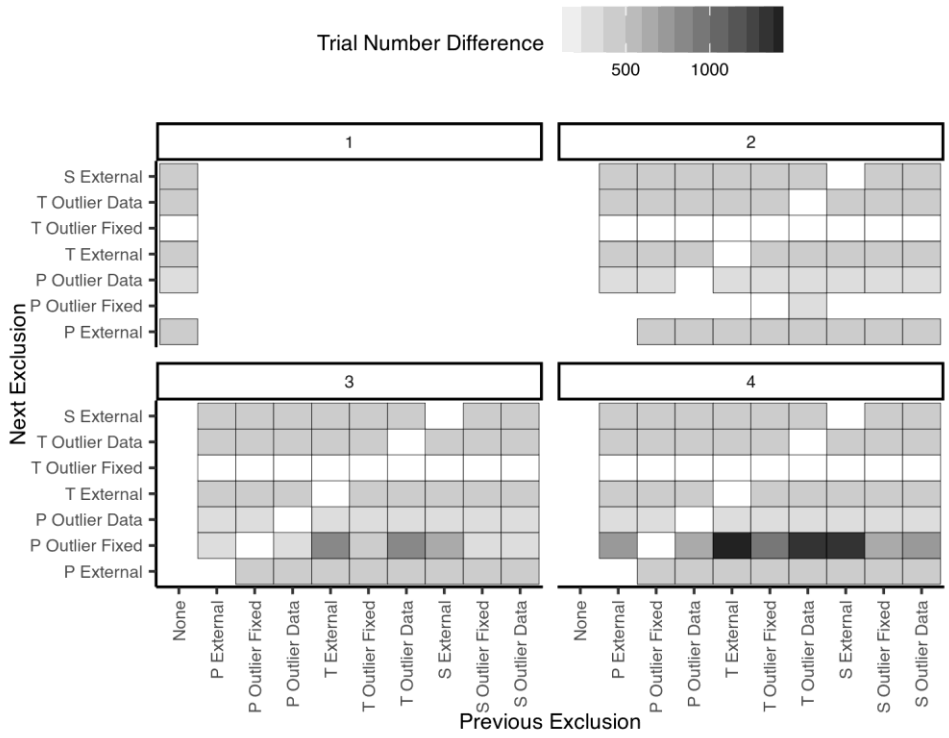

Figure 14. Number of trials included in the trial-level analysis (multilevel modeling) based on the order and pairwise-combination of exclusions. The panels represent the number of exclusions present. The x-axis represents the previous exclusion applied, while the y-axis represents the next exclusion to be applied. The heat color represents the maximum change seen in number of trials for that combination. P = Participant, T = Trial, S = Stimuli. All zero values have been excluded.

5.3.3. Question 3 - Standardized Effect Size.

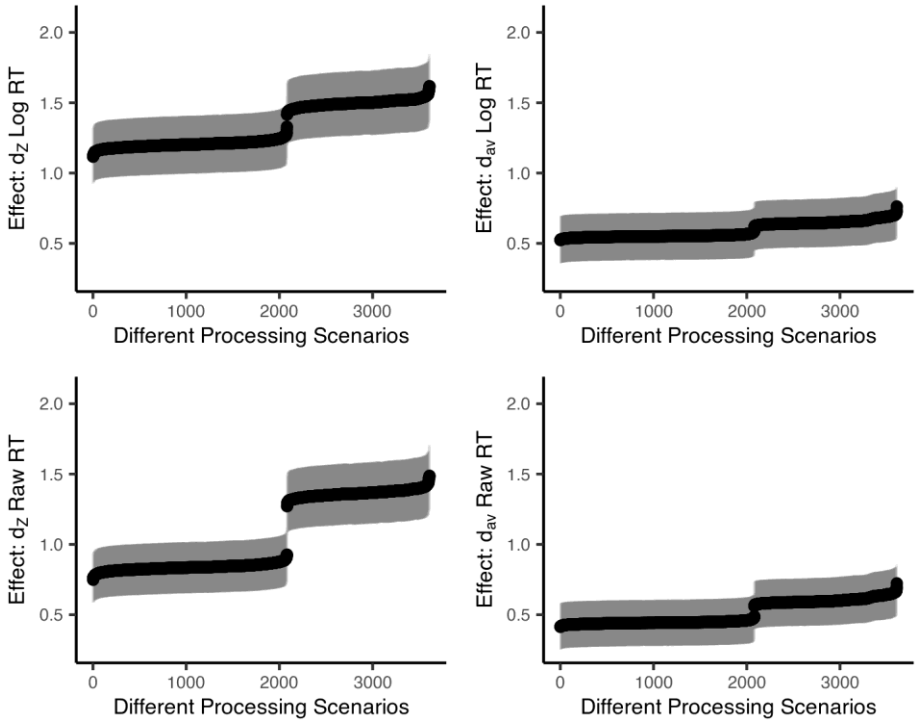

Figure 15. Effect sizes for d_Z (column 1) and d_{av} (column 2) for log transforms (row 1) and no transformation (row 2). These effect sizes are arranged from smallest to largest across the x-axis to demonstrate a sensitivity plot of the range of possible effects found.

Figure 15 portrays the differences in choice of effect size split by the choice of data transformation. First, it may have a small (i.e., r < .20), but non-zero relationship with sample size as the correlation between sample size and log RT d_Z was $r = -.04$, 95% CI $[-.08, -.01]$ with corresponding correlations for each of the other effect sizes: log RT d_{av} $r = -.14$, 95% CI $[-.17, -.11]$, raw RT d_Z $r = -.04$, 95% CI $[-.07, .00]$, and raw RT d_{av} $r = -.11$, 95% CI $[-.14, -.08]$. The results indicate that d_{av} may be more influenced by final sample size; however, the patterns are the same for both data aggregation methods. Second, the difference in effect size bias is clear — d_Z returns an effect size that is double or more the size of d_{av} on the

same data. They appear to show the same pattern of effects when the pre-processing actions are applied, but the effect is more pronounced in d_Z.

5.3.4. Question 4 - Raw Effect Size.

As shown in *Figure 16*, the effect of transformation and aggregation choice does not appear to affect the final raw score effect size; however, the size of the transformed effect size difference = 1.08 - 1.12 ms is smaller than the overall raw score difference: between 30 - 50 ms. The effect does vary by pre-processing scenario with larger confidence intervals in various scenarios due to sample or trial level size.

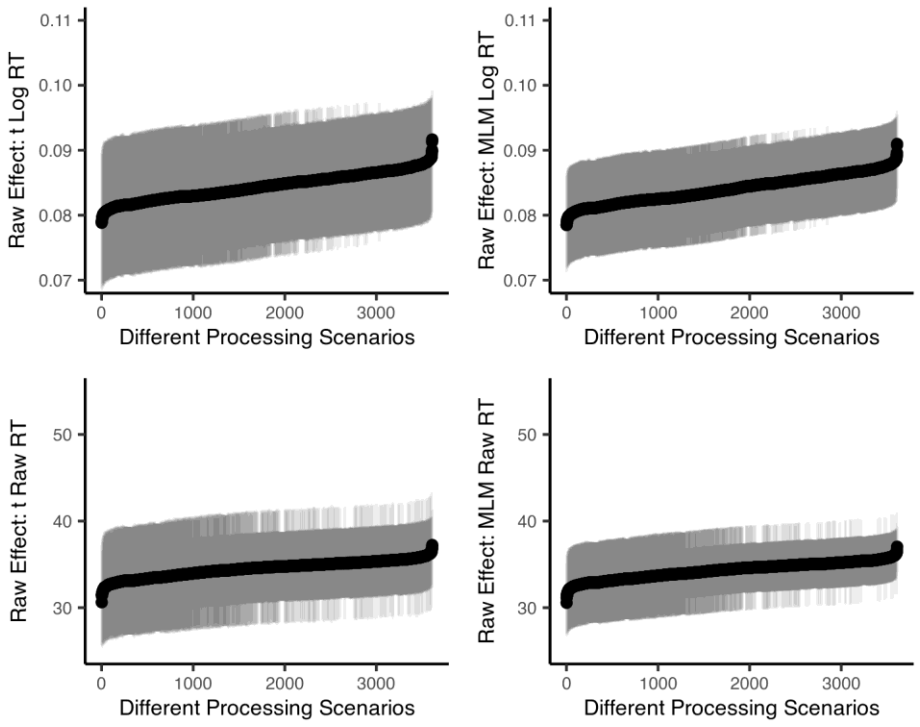

Figure 16. Raw score effect sizes for *t*-test and multilevel models using both raw RTs and log RTs. These effect sizes are arranged from smallest to largest across the x-axis to demonstrate a sensitivity plot of the range of possible effects found.

5.4. Discussion

The multiverse results portray that all facets of the pre-processing choices and checklist items proposed in this manuscript likely have impacts on the result presented. First, the order and number of pre-processing actions changed effect size results, with larger impacts on the traditional choice for effect size, d_Z over the now recommended choice d_{av}. The number of participants or trials will affect the power of the study, and the choice of pre-processing action and order also impacted these results. Finally, data transformation and aggregation choices influenced the results of the study. Therefore, we conclude that all aspects of the checklist are likely necessary for sufficient understanding of the results presented in a manuscript.

We would like to stress the fact that we could not use raw data, as the original authors excluded some participants without enough documentation for us to simulate the missing participants. These circumstances are another example for our results from the literature search showing why a checklist like ours can be helpful for understanding the methods of a study and re-using available data. We assume that the multiverse analysis with actual raw data should have yielded larger effect size differences between the pipelines, as the data would have included more noise, while already conveying differences of 0.14 to 0.20 in the current set-up. Of course, the Simon effect is a relatively stable and large effect in a straightforward two-by-two design, so that smaller and less reliable cognitive effects might be even more prone to changes in reaction time preprocessing (as indicated by large false positive rates identified by Berger & Kiefer, 2021, and Morís Fernández & Vadillo, 2020). However, we used the Simon effect only as a well-known example of cognitive processing, so that we have an empirical basis for our suggestions about reporting RT preprocessing. It will be instructive to see the preprocessing actions and their order as manipulated in the current study applied to other effects and research domains.

6. Recommendations and Conclusion

Data processing is a multi-action process - from the initial data cleaning, to outlier detection and variable transformation. For each action, there are many possible options used in the scientific literature, and these data pre-processing choices can systematically affect statistical outcomes and theoretical conclusions (see present multiverse analysis and Kerr et al., 2017; Primbs, Holland, et al., 2022). Therefore, it is important to document each of these actions and transparently report all pre-processing decisions. To that end, the present research provides an overview of frequent pre-processing decisions (*Table 1*) and a checklist for reporting pre-processing actions (*Table 2*). The multiverse analysis performed for this project indicates the necessity of reporting all forms of data pre-processing as the choice of order, aggregation, and effect size type influenced the final size of raw/standardized effects and power. Our checklist offers concrete guidance on what to report and where to report it in order to facilitate the accurate reporting of pre-processing decisions. This checklist does not give advice on how to pre-process reaction time data, but on how to report these actions, thus, our recommendations can be generalized to any kind of experimental reaction time paradigm.

As a result of the present research, we provide the scientific community with two available versions of the checklist: a short and a detailed one. Both can be found in the **Supplementary Materials D and E** as well as downloaded as pdf documents from our OSF project page [longer version https://osf.io/q3cxb; shorter version https://osf.io/32hpy]. We give brief instructions on how to use the checklist, without the need to read the entire manuscript. We foresee that our checklist will be a living document, that integrates within itself changes of a scientific mindset, and most importantly used by researchers. Note, it is not a goal of our checklist to tell which pre-processing actions should be applied (if any), as responses to the survey clearly indicated a large

variety of approaches being applied, each with a legitimate rationale. The heart of the checklist is the easy-to-grasp transparency of RT data pre-processing.

Sharing the code is the gold standard of open-data practices and enables access to exact analysis pipelines. However, for several reasons, access to the code is not itself enough to provide the public with the information they can use. On the one hand, some researchers use non-open-source programs with expensive licences, thus making the shared code useless for the researchers who cannot afford such software. On the other hand, researchers use different programming languages for statistical analyses, with few researchers being fluent in multiple programming languages. Consequently, literacy reading and understanding code is not self-evident. Additionally, going through code from an unknown researcher can be time-consuming. Therefore, we deem our checklist to improve inclusiveness by ensuring that all relevant information is provided as a verbal (programming language/ software-agnostic) description using common methodological wording.

We envision that journals should require the checklist to be uploaded as a Supplementary Material to the published manuscript so that it does not take up valuable space in the main text (comparable to policies of medical journals). This will ensure sufficient reporting by authors and enable reviewers to quickly get an overview on how the data has been processed. Furthermore, it will be easier for a potential data editor to double-check the correctness and appropriateness of the code if the authors verbally formulate their intention. The headers we suggested for the different actions could then be used in any shared code, as well to structure the code, and make code easier to relate the respective pre-processing action. Therefore, we provide an additional file with checkboxes on OSF which can be downloaded and filled in by the authors.

We also see advantages in the educational domain. First of all, supervisors can give their students guidance in writing theses and conducting their own experiments by introducing common pre-processing approaches and reflect on their applicability to the respective research question. Second, the checklist allows for researchers who are new to the field to faster summarize best practice of RT pre-processing in their specific domain.

For cases where no or only some pre-processing has been applied, it does not make sense to upload an empty checklist. Therefore, we would like to recommend the following sentences in the analysis section of the manuscript:

- *In our analysis, we have used unpre-processed (raw) data,* where no pre-processing was done.
- If the authors conducted only parts of the pipelines, they should explicitly state: *First, we excluded erroneous data, then RT > 2000 ms or < 150 ms. No other pre-processing actions were performed.*

In addition, the use of checklists is widespread in other research areas such as health research, leading to development of entire frameworks such as the EQUATOR framework (Altman et al., 2008). Their main aim is to increase transparency and standardize the reporting of various aspects of papers. This transparency is also especially important when it comes to reaction times. The necessity for data transformations due to a typical experiment producing skewed distributions to meet statistical assumptions, and lack of reporting on outlier exclusion may lead to a variety of outcomes, as shown in our multiverse analysis. Therefore, clear information on how the data were processed prior to statistical analyses seems crucial. In a way, this transparency can be achieved by sharing analysis codes, in addition to data. However, analysis codes can vary in programming languages, styles of coding, and clarity of the comments available to the secondary user. Also, analysis codes are not subject to mandatory peer review. Bearing that in mind, standardization attempts currently need to focus on the main paper text and

journal published supplementary information. An important point that needs to be made in addition to being available for use and improvement, checklists need to be actively promoted by both authors and editorial boards (Altman et al., 2008). Finally, this checklist may be used in order to structure parts of methods or results sections of a paper, or to be included as authors instructions or recommended by reviewers. In other words, all parties in the publishing process may have some use from this checklist. As an example, the ARRIVE 2.0 guidelines (Percie du Sert et al., 2020) are recommended to be used with the intent to facilitate joint effort to increase reporting transparency; however, they are not intended to be a replacement for journal requirements.

Our research aligns with recent demands for greater transparency in psychological research (Wicherts et al., 2016). Wicherts and colleagues (2016) provide researchers with a list of decisions that allow for researchers' degrees of freedom including data pre-processing decisions and argue for the importance of pre-registration. Our checklist extends this work as a detailed account of all planned pre-processing actions should be part of a good pre-registration. As such, our checklist facilitates the evaluation of the severity of a test (Lakens, 2019). Multiverse approaches that include analyses based on multiple pre-processing pathways (Steegen et al., 2016) are not exempt from this transparency and should also be pre-registered (Primbs, Rinck, et al., 2022). However, the complete and accurate reporting of data pre-processing decisions is also important for a different reason: differences in data pre-processing decisions have been shown to considerably influence results and conclusions (Bastiaansen et al., 2020; Botvinik-Nezer et al., 2020; Kerr et al., 2017; Primbs, Holland, et al., 2022; Silberzahn et al., 2018). Ultimately, our checklist allows researchers to gauge whether different results may be due to differences in data pre-processing decisions, which is crucial for designing replication projects (Bokhove, 2022).

Importantly, while the examples we provide in *Table 2* show how to report pre-processing actions within the text of a paper, we recognize that it can be difficult to include very detailed information while complying with word/page limits and that too much detail can sometimes negatively impact the readability of a paper. Therefore, we strongly recommend that researchers share the code and data needed to reproduce the numerical results presented in their paper. This information makes it more likely that pre-processing actions, and the sequence in which they are implemented, are completely and transparently reported.

Our checklist also facilitates novel meta-scientific endeavors: If researchers report all decisions they make, this information allows for the development of updated and improved checklists, enables the creation of inventories of common practices, and facilitates comparisons between different tasks and research fields. Overall, the present research contributes to the scientific literature by providing a checklist for the complete and accurate reporting of reaction-time-based experiments which is not only accessible and easily implementable, but also achieved through interaction among the scientific community members.

Transparency and Openness Promotion (TOP)

In the present paper, we reported how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study (Simmons et al., 2012). **Preregistration**: no part of the study procedures and analyses was pre-registered prior to the research being conducted.

Study materials and corresponding outputs can be found on the project's OSF page: https://osf.io/reqat/. **Materials**: *Expert-consensus Survey* (pdf) https://osf.io/pa3g8, (qsf)

https://osf.io/buc9d; *Checklist for Reporting Reaction Time Pre-Processing Decisions*, longer version https://osf.io/q3cxb, shorter version https://osf.io/32hpy; **Data**: *Data from the expert-consensus survey* https://osf.io/avhqr; *complete free-text responses in the expert-consensus survey* https://osf.io/7ubv6; **Analysis**: *Meta-analysis of the literature search* https://osf.io/kj3yx; *Analysis of the expert-consensus survey responses* https://osf.io/q9ge2, *Multiverse analysis* https://osf.io/d29ck.

CRediT Author Statement

Hannah Dorothea Loenneker: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration. **Erin M. Buchanan**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **Ana Martinovici**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **Maximilian A. Primbs**: Conceptualization, Methodology, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Project administration. **Mahmoud Medhat Elsherif**: Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing, Project administration. **Bradley J. Baker**: Validation, Investigation. **Leonie A. Dudda**: Validation, Investigation. **Dušica Filipović Đurđević**: Validation, Investigation, Writing - Review & Editing. **Ksenija Mišić**: Validation, Investigation, Writing - Review & Editing. **Hannah K. Peetz**: Validation, Investigation, Writing - Review & Editing. **Jan Philipp Röer**: Validation, Investigation, Writing - Review & Editing. **Lars Schulze**: Validation, Investigation. **Lisa Wagner**: Validation, Investigation. **Julia Katharina Wolska**: Conceptualization,

Methodology, Validation, Investigation, Writing - Review & Editing. **Corinna Kührt**: Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing, Project administration. **Ekaterina Pronizius**: Conceptualization, Methodology, Validation, Investigation, Writing - Review & Editing, Visualization, Project administration, Funding acquisition.

Acknowledgements

This project is the result of a collaboration started at SIPS 2021 in the following unconference and hackathon: Cipora, K., Loenneker, H.D. (2021, June). (Too) many shades of reaction time data pre-processing [Unconference session]. Society for the Improvement of Psychological Science (SIPS) 2021 meeting, virtual conference, originally Padua: Italy. https://osf.io/mnv7w/; Cipora, K., Loenneker, H.D. (2021, June). (Too) many shades of reaction time data pre-processing [Hackathon]. Society for the Improvement of Psychological Science (SIPS) 2021 meeting, virtual conference, originally Padua: Italy. https://osf.io/mnv7w/. We thank the attendees of both sessions.

Moreover, we would like to thank Gijsbert Bijlstra for comments on earlier versions of the checklist.

We also would like to express our gratitude to the following open-source resources, which greatly improved the project workflow: *GitHub* (github, 2020), *R Markdown: The Definitive Guide* (Xie et al., 2018), *papaja* (Aust et al., 2022), *ggplot2* (Wickham, Chang, et al., 2022), *flextable* (Gohel et al., 2021), *officedown* (Gohel & Ross, 2022), *rio* (Chan et al., 2021), *dplyr* (Wickham, François, et al., 2022), *tidyverse* (Wickham & RStudio, 2022), *Cairo* (Urbanek & Horner, 2022), *here* (Müller & Bryan, 2020), *gmodels* (Warnes et al., 2022), *metafor* (Viechtbauer, 2022), *data.table* (Dowle et al., 2022), *nlme* (Pinheiro et al., 2022), *patchwork*

(Pedersen, 2020), *RColorBrewer* (Neuwirth, 2022), *latex2exp* (Meschiari, 2022), and *psych* (Revelle, 2022).

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Altman, D. G., Simera, I., Hoey, J., Moher, D., & Schulz, K. (2008). EQUATOR: Reporting guidelines for health research. *The Lancet*, *371*(9619), 1149–1150.

André, Q. (2022). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General*, *151*(1), 213–223. https://doi.org/10.1037/xge0001069

Aust, F., Barth, M., Diedenhofen, B., Stahl, C., Casillas, J. V., & Siegel, R. (2022). *Papaja: Prepare American Psychological Association Journal Articles with R Markdown*. https://CRAN.R-project.org/package=papaja

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28. https://doi.org/10.21500/20112084.807

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., Jonge, P. de, Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., … Bringmann, L. F. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Bella-Fernández, M., Martin-Moratinos, M., Li, C., Wang, P., & Blasco-Fontecilla, H. (2023). Differences in ex-gaussian parameters from response time distributions between individuals with and without attention deficit/hyperactivity disorder: A meta-analysis. *Neuropsychology Review*, 1–18. https://doi.org/10.1007/s11065-023-09587-2

Berger, A., & Kiefer, M. (2021). Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in Psychology*, *12*, 675558. https://doi.org/10.3389/fpsyg.2021.675558

Bijlstra, G., Holland, R. W., & Wigboldus, D. H. (2010). The social face of emotion recognition: Evaluations versus stereotypes. *Journal of Experimental Social Psychology*, *46*(4), 657-663. https://doi.org/10.1016/j.jesp.2010.03.006

Bokhove, C. (2022). The role of analytical variability in secondary data replications: A replication of Kim et al. (2014). *Educational Research and Evaluation*, *27*(1-2), 141–163. https://doi.org/10.1080/13803611.2021.2022319

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Burle, B., Spieser, L., Servant, M., & Hasbroucq, T. (2014). Distributional reaction time properties in the Eriksen task: marked differences or hidden similarities with the Simon task?. *Psychonomic bulletin & review*, *21*, 1003-1010. https://doi.org/10.3758/s13423-013-0561-6

Cespón, J., Hommel, B., Korsch, M., & Galashan, D. (2020). The neurocognitive underpinnings of the Simon effect: An integrative review of current research. *Cognitive, Affective, & Behavioral Neuroscience*, *20*(6), 1133–1172. https://doi.org/10.3758/s13415-020-00836-y

Chan, C., Chan, G. C., Leeper, T. J., & Becker, J. (2021). *Rio: A Swiss-army knife for data file I/O*.

Christensen, G., Wang, Z., Levy Paluck, E., Swanson, N., Birke, D., Miguel, E., & Littman, R. (2020). *Open science practices are on the rise: The state of social science (3s) survey*. https://escholarship.org/uc/item/0hx0207r

Craft, J. L., & Simon, J. R. (1970). Processing symbolic information from a visual display: Interference from an irrelevant directional cue. *Journal of Experimental Psychology*, *83*(3, Pt.1), 415–420. https://doi.org/10.1037/h0028843

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

De Pretto, M., Mouthon, M., Debove, I., Pollo, C., Schüpbach, M., Spierer, L., & Accolla, E. A. (2021). Proactive inhibition is not modified by deep brain stimulation for parkinson's disease: An electrical neuroimaging study. *Human Brain Mapping*, *42*(12), 3934–3949.

Domahs, F., Moeller, K., Huber, S., Willmes, K., & Nuerk, H.-C. (2010). Embodied numerosity: Implicit hand-based representations influence symbolic number processing across cultures. *Cognition*, *116*(2), 251–266. https://doi.org/10.1016/j.cognition.2010.05.007

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R.,

Seiskari, O., Dong, X., Lang, M., Iwasaki, W., Wenchel, S., … Schwen, B. (2022). *Data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170

Gabay, S., & Behrmann, M. (2014). Attentional dynamics mediated by subcortical mechanisms. *Attention, Perception, & Psychophysics*, *76*, 2375–2388.

github. (2020). *GitHub*. https://github.com/

Gohel, D., Fazilleau, Q., Nazarov, M., Robert, T., Barrowman, M., & Yasumoto, A. (2021). Flextable: Functions for tabular reporting. *R Package*, *8*. https://CRAN.R-project.org/package=flextable

Gohel, D., & Ross, N. (2022). *Officedown: Enhanced 'r markdown' format for 'word' and 'PowerPoint'*. CRAN. https://CRAN.R-project.org/package=officedown

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., ... & Buchanan, E. (2013). The semantic priming project. *Behavior research methods*, *45*, 1099-1114. https://doi.org/10.3758/s13428-012-0304-z

Kerr, J. A., Hesselmann, G., Räling, R., Wartenburger, I., & Sterzer, P. (2017). Choice of analysis pathway dramatically affects statistical outcomes in breaking continuous flash suppression. *Scientific Reports*, *7*(1), 3002. https://doi.org/10.1038/s41598-017-03396-3

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2019). *The value of preregistration for psychological science: A conceptual analysis*. 心理学評論刊行会. https://doi.org/10.24602/sjpr.62.3_221

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Meschiari, S. (2022). *latex2exp: Use LaTeX expressions in plots*. https://cran.r-project.org/web/packages/latex2exp/index.html

Morís Fernández, L., & Vadillo, M. A. (2020). Flexibility in reaction time analysis: Many roads to a false positive? *Royal Society Open Science*, *7*(2), 190831. https://doi.org/10.1098/rsos.190831

Müller, K., & Bryan, J. (2020). *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here

Mussel, P., Spengler, M., Litman, J. A., & Schuler, H. (2011). Development and validation of the german work-related curiosity scale. *European Journal of Psychological Assessment*.

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. https://cran.r-project.org/web/packages/RColorBrewer/index.html

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748.

Pain, M. T. G., & Hibbs, A. (2007). Sprint starts and the minimum auditory reaction time. *Journal of Sports Sciences*, *25*(1), 79–86. https://doi.org/10.1080/02640410600718004

Parsons, S. (2022). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *Meta-Psychology*, *6*. https://doi.org/10.15626/MP.2020.2577

Pedersen, T. L. (2020). *Patchwork: The composer of plots*. https://CRAN.R-project.org/package=patchwork

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227.

Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., et al. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *Journal of Cerebral Blood Flow & Metabolism*, *40*(9), 1769–1777.

Pinheiro, J., Bates, D., & R Core Team. (2022). *Nlme: Linear and nonlinear mixed effects models*. https://CRAN.R-project.org/package=nlme

Primbs, M. A., Holland, R., Quandt, J., & Bijlstra, G. (2022). *The effects of data pre-processing and analysis pathway on statistical outcomes*. OSF. https://osf.io/9gj8v/

Primbs, M. A., Rinck, M., Holland, R., Knol, W., Nies, A., & Bijlstra, G. (2022). The effect of face masks on the stereotype effect in emotion perception. *Journal of Experimental Social Psychology, 103*, 104394. https://doi.org/10.1016/j.jesp.2022.104394

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research*. https://cran.r-project.org/web/packages/psych/

Rinck, M., Primbs, M. A., Verpaalen, I. A., & Bijlstra, G. (2022). Face masks impair facial emotion recognition and induce specific emotion confusions. *Cognitive Research: Principles and Implications, 7*(1), 1-15. https://doi.org/10.1186/s41235-022-00430-5

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. (2012) A 21-word solution. SSRN Electronic Journal. doi: 10.2139/ssrn.2160588.

Soltanlou, M., Artemenko, C., Ehlis, A.-C., Huber, S., Fallgatter, A. J., Dresler, T., & Nuerk, H.-C. (2018). Reduction but no shift in brain activation after arithmetic learning in children: A simultaneous fNIRS-EEG study. *Scientific Reports*, *8*, 1707. https://doi.org/10.1038/s41598-018-20007-x

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67. https://www.jstor.org/stable/2237638

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*. https://doi.org/10.1037/0096-3445.123.1.34

United Nations geoscheme. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=United_Nations_geoscheme&oldid=1099506198

Urbanek, S., & Horner, J. (2022). *Cairo: R Graphics Device using Cairo Graphics Library for Creating High-Quality Bitmap (PNG, JPEG, TIFF), Vector (PDF, SVG, PostScript) and Display (X11 and Win32) Output*. https://CRAN.R-project.org/package=Cairo

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631–650. https://doi.org/10.1080/14640749408401131

Viechtbauer, W. (2022). *Metafor: Meta-Analysis Package for R*. https://CRAN.R-project.org/package=metafor

Warnes, G. R., Bolker, B., Lumley, T., & Johnson, R. C. (2018). *Gmodels: Various R Programming Tools for Model Fitting*. https://CRAN.R-project.org/package=gmodels

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-Hacking. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & RStudio. (2022). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., Müller, K., & RStudio. (2022). *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr

Wickham, H., & RStudio. (2022). *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H., Huff, K. D., Mitchell, I. M., Plumbley, M. D., et al. (2014). Best practices for scientific computing. *PLoS Biology*, *12*(1), e1001745.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Computational Biology*, *13*(6), e1005510.

Woods, A. D., Davis-Kean, P., Halvorson, M. A., King, K. M., Logan, J. A. R., Xu, M., Bainter, S., Brown, D. M. Y., Clay, J. M., Cruz, R. A., Elsherif, M. M., Gerasimova, D., Joyal-Desmarais, K., Moreau, D., Nissen, J., Schmidt, K., Uzdavines, A., Van Dusen, B., & Vasilev, M. R. (2021). *Missing data and multiple imputation decision tree*. PsyArXiv. https://doi.org/10.31234/osf.io/mdw5r

Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P., Halvorson, M. A., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K., & Elsherif, M. M. (2023). Best practices for addressing missing data through multiple imputation. *Infant and Child Development*, e2407. https://doi.org/10.1002/icd.2407

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC.

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant Nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, *25*(5), 1968–1972. https://doi.org/10.3758/s13423-017-1348-y