ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

# ERROR REPORT

Wessel, J. (2018). Prepotent motor activity and inhibitory control demands in different variants of the go/no-go paradigm. *Psychophysiology*. doi: 10.1111/psyp.12871

## DECISION: Minor errors

*Reviewer:* **Russ Poldrack**, Stanford University
*Author response:* **Jan Wessel**, University of Iowa
*Recommender:* **Ian Hussey**, University of Bern

*Cite as:* ERROR (2024) Error review of Wessel (2018). Version 1. https://doi.org/10.31234/osf.io/8gtz2

# DECISION & ЯECOMMENDATION

Wessel (2018) "Prepotent motor activity and inhibitory control demands in different variants of the go/no-go paradigm" was determined to contain **Minor Errors**. That is, errors that have the benefit of being detectable thanks to the presence and sharing of research materials, but whose scope and implications on the conclusions of the article are minor. The detected errors do not rise to the level of recommending that a correction be issued. I thank Prof Wessel for nonetheless indicating his interest in pursuing such a correction. Authors can of course pursue a correction unilaterally if they choose to do so.

Following the ERROR project's emergent guidelines, the recommendations associated with a decision of 'Minor Errors' are as follows:

- The report, author response, and recommendation have been posted on the ERROR website (error.reviews/reviews/wessel-2018) and as a preprint on PsyArXiv (osf.io/preprints/psyarxiv/8gtz2). Their associated materials have been posted to OSF (see osf.io/2q3ay).
- The author is asked to recognise these errors in future discussions of the article.

# RECOMMENDER'S ЯEPORT

Ian Hussey

As the first recommendation issued by the ERROR project, I'd like to especially thank Prof Poldrack (reviewer) and Prof Wessel (author) for serving as the project's first completed pair of error review and author response. These two documents have set an exceptionally high bar for all subsequent error reviews to follow. I would like to especially acknowledge the author, Prof Wessel, for going above and beyond in his author response by reextracting, with the aid of members of his research group, the whole dataset again to comprehensively check for errors not already found by the reviewer, and by constructing a simulation study to attempt to understand the probability of there still being additional undetected errors still in the data. The substance and style of these reports embody everything we hope to see more of in academic research: acceptance of the possibility that errors occur; inspection for errors that is well-documented and verifiable; and acknowledgement and suitable correction when errors are found.

I should note that unlike editorial decision letters in pre-publication peer review, which are often private to the author and reviewers, ERROR recommendation letters – as well as the reviewer report, author response, and all materials associated materials – are public documents whose function is to (1) communicate the presence or absence of any errors detected, (2) consider their severity, and (3) provide discussion of how similar errors elsewhere might be prevented or detected.

**Summary of errors detected & how they could be prevented in future**

The original article conducted a meta-methods review of the task parameters used for go/no-go tasks and observed that only a minority of articles in the literature employ what the author argues to be the necessary parameters (i.e. rare no-go trials and short trial durations). New data collection was then used to experimentally manipulate these task parameters, showing that go/no-go tasks only elicit reliable prepotent motor activity when the task is fast paced and when no-go trials are rare.

We asked the reviewer to focus their review on the first section – the meta-methods review of go/no-go task parameters employed in the literature.

Research elsewhere has shown that meta-science studies that extract details from published studies (e.g., meta-analyses and meta-methods studies) may be more prone to errors that is currently widely acknowledged. This has been shown to be the case in meta-analyses (e.g., Maassen et al., 2020: 10.1371/journal.pone.0233107). The current review suggests that it may also be the case for meta-methods reviews. The author's simulation suggests that even after this ERROR review and his research group's careful re-extraction of data from all articles in the sample, the probability that errors remain in the dataset is 96%. From the author's data in their response, errors by individual coders seem difficult to avoid and are not clearly reduced or eliminated by reducing their individual workload or potentiating them towards finding errors.

Researchers should treat these extracted details as estimates of the original details that contain measurement error. As such, steps should be taken to quantify and reduce this error. At least two independent raters should be used to extract these details. The inter-rater reliability should be reported in the article along with the resolution strategy, and some consideration should be given to the prevalence of errors remaining in the data. Even with two raters, additional undetected errors will likely still be present (e.g. where the raters agree but are both wrong) and indeed prevalent. Naturally, the extracted data should be publicly available to allow for verifiability.

## Discussion of individual issues raised

### Paper Selection for the literature analysis

The reviewer raised the point regarding the Boolean search terms used for the review: while not erroneous, the choice of terms may not have been sufficient to identify all possibly relevant papers. Suggested alterations to the search query returned 14% more results than the original search. The reviewer noted however that it is unlikely that it is highly unlikely that additional articles would have changed the results systematically. The author agrees with the existence and scope of this issue, as do I. This is a useful question to have explored, but does not rise to the level of an error.

### Extraction and analysis of go/no-go task parameters

The reviewer re-extracted the task parameters for 24 (10%) of the articles employed in the original review. I will focus on the P(nogo) parameter here, as it is more important. Three of the 24 articles (12.5%) that the reviewer re-extracted contained values different to those reported in the article (a fourth was identified too but as the author points out, is correct when rounded).

I was very grateful that Prof Wessel and his research group were willing to take the time to re-extract data from the other 90% of the articles. This was done unprompted by us or the reviewer, out of a desire to understand the actual error rate in the data once initial errors had been found by the reviewer. This re-extraction uncovered eight (3.3%) additional errors in the values originally extracted. Additionally, 4 new errors were made where, on inspection, the original value appeared to be correct (Cohen’s Kappa between original and new ratings = .958). The author provided a useful assessment of bias in the original ratings that suggests that errors were evenly distributed above and below the correct values, therefore these errors represented noise but not systematic bias.

The author's response provides a very useful discussion of the comparable error rates found even among new raters who were potentiated towards finding errors. This is a useful finding for future research. As noted in the author's response, ratings by additional individuals appear to be key to detecting errors prior to analysis. I thank the author for taking the time to write a simulation study to estimate the probability that, even after being scored a second time, the data still contains one or more errors, and for being receptive to my feedback in finding and correcting bugs in this simulation. Like the author, I was surprised to learn that the probability of the dataset still containing at least one error is 96%. This underscores the important point that error-reviewed articles cannot now be assumed to be error-free, only that they are likely to contain fewer errors, or that the degree to which errors can be checked for has been determined. I encourage readers of this recommendation to read Prof Wessel's author response in full. The reviewer and author agree about the existence and scope of these data extraction issues. I also agree that this represents a minor error.

**Reproduction of Figures 1 and 2**

As part of attempting to reproduce the plots that illustrate the results of the meta-methods review of go/no-go task parameters, the reviewer noted that the original code was not available (although the data was linked in the article). The author's response notes that the original code was in fact available on their website. Here, we the ERROR organising team, must acknowledge our own error: we did not sufficiently communicate the fact that the reviewer could directly ask the author for such materials as part of the ERROR review process, whether they are already publicly available online or not. We have updated our instructions for future reviewers to strengthen reviewers' expectations that can and should ask the authors directly for materials, such as data and code, or aid with interpreting and understanding these materials, in the course of their review. Reviewers will be more explicitly instructed that the authors have consented to sharing such research materials beyond what is necessarily in the publication; that they can and should ask for these materials; and that they can ask appropriate questions about details where needed. We prefer that reviewers obtain and discuss materials directly with the authors (with the ERROR team in cc) rather than materials being transferred via us in order to (a) lower the internal administrative burden for the ERROR team, and (b) to make ERROR reviews a more collaborative interaction between authors and reviewers.

The reviewer's reconstruction of the plots identified that the original figures included a small offset to the observed data to add plotting, and that this offset was erroneously brought into the reported results too. That is, the offset data used for plotting was also used to report results, rather than the actual data. I agree that this does constitute an error, and was well-spotted by the reviewer. The author agrees with the existence and scope of this error, and I agree that its magnitude is minor and it does not change the interpretation of the results. I sympathise with the author's frustration that this error was nonetheless made, and yet I think it provides a relatable and accessible example of how an erroneous workflow can give rise to erroneous results. The authors response provides a clear description of the workflow and how the error was likely made.

**Unresolved issues**

It is reasonable to expect that ERROR reviews will leave some questions unresolved. It is useful to acknowledge some of these issues so that this or future ERROR recommendations do not artificially convey that they represent the final word on issues of error detection and correction for a given article. In this case, it is a difference in the results between the author's simulation study code, written as part of the author response, and my attempt to reimplement it in R.

I have never worked in MatLab, therefore I attempted to convert the author's simulation code to R with the help of ChatGPT. This was a useful exercise, as it allowed us to discover and fix an error in the original code. While my R code and the author's MatLab code produce similar estimates of the probability of there being remaining errors in the dataset (both 96%), we get different estimates for the probability that reviewers make the same error (my R code: 72%; the author's MatLab code: 51%). Neither of us could discern the reason for this difference in a reasonable amount of time, neither of us being experienced in the other language. I am happy to defer to the author's estimate here.

I sincerely thank both Prof Poldrack and Prof Wessel again for their efforts and error-acceptance here.

Ian Hussey
Chief Recommender for ERЯOR

# ЯEVIEW

Russ Poldrack

# I. METHODS, MEASUREMENT, AND DESIGN

## 1. Design No errors found

Are there errors in the conceptual design of the study? E.g., flawed randomisation technique

Note that this ERROR review only examined the literature analysis portion of the study. I did not identify any errors in the design of the literature analysis.

*Time spent: 10 minutes*

## 2. Measurement No errors found

Are there any measures, techniques, or devices that were incorrectly applied or inappropriate for the specific task described in the paper?

*Time spent: 10 minutes*

## 3. Preregistration Consistency Not applicable

Are there substantial deviations from the preregistration, particularly undisclosed ones?

*Time spent: minutes/hours*

## 4. Sampling Not applicable

Is there an error in the sampling strategy? Is the power analysis reproducible? Does the model used for the power analysis match the model in the substantive analyses? Were separate power analyses conducted for all primary analyses?

*Time spent: minutes/hours*

## 5. Other Aspects Related to Methods and Measures Indeterminable

The Pubmed query used for the literature search is not incorrect per se, but may not have been sufficient to identify all possibly relevant papers. Note that the following statements are based on the Pubmed query expansion engine as of March 2024; it is likely that the query expansion has changed since the original research was completed, and it is thus impossible to know exactly how the results might have differed at that point.

The authors report using the following Pubmed query: "inhibition AND human AND go-nogo task." When terms are included in a search without quotation marks around the individual terms (which I am assuming was the case here), then the query is expanded to find related terms. As of today, Pubmed expands the authors' original query as follows (obtained by downloading the search history from the Pubmed Advanced Search Builder at https://pubmed.ncbi.nlm.nih.gov/advanced/):

```
("inhibit"[All Fields] OR "inhibitable"[All Fields] OR
"inhibite"[All Fields] OR "inhibited"[All Fields] OR
"inhibites"[All Fields] OR "inhibiting"[All Fields] OR
"inhibition, psychological"[MeSH Terms] OR ("inhibition"[All
Fields] AND "psychological"[All Fields]) OR "psychological
inhibition"[All Fields] OR "inhibition"[All Fields] OR
"inhibitions"[All Fields] OR "inhibitive"[All Fields] OR
"inhibits"[All Fields]) AND ("human s"[All Fields] OR
"humans"[MeSH Terms] OR "humans"[All Fields] OR "human"[All
Fields]) AND ("go-nogo"[All Fields] AND "task"[All Fields])
```

Because of the manual filtering that was performed by the authors, I am not concerned about the potential for irrelevant papers to be identified by this expanded search. However, there is a concern that some relevant papers may have been missed; for example, if the authors used the term “paradigm” instead of “task”. As of today, the authors’ query returns 729 matching papers in Pubmed. As an example of how the authors might have expanded their query to identify additional publications, I tested the following query:

```
inhibition AND human AND (go-nogo task OR "go-nogo paradigm" OR
"go/no-go paradigm")
```

This search returned 826 results, more than 10% more than the original search. I did not examine the additional papers found by this search to determine how many of them might have survived the authors’ filters, but this example does show that the recall of the search could have been increased.

I am not concerned about the effects of these different queries on the results of the analysis, as it seems highly unlikely that the additional papers would have differed systematically in a way that would have changed the result.

*Time spent: 20 minutes*

# II. DATA, CODE, AND STATISTICAL ANALYSES

### 1. Code Functionality Not applicable

Does the provided code run without the need to make any adjustments and without errors? If not, what steps were needed to get it to run (if it was eventually possible)?

*Time spent: minutes/hours*

### 2. Computational Reproducibility of Reported Statistics Not applicable

Is there a clear traceability of reported stats to code? Does the code output match what's reported in the paper? Are all reported statistics findable within the analysis code?

*Time spent: minutes/hours*

### 3. Data Processing Errors Errors found

Are there substantive errors during the preparation or cleaning of data (e.g. duplication of rows during a merge) prior to substantive analyses and hypothesis tests?

I checked the coding of the P(nogo) and minimum/maximum SOA values that were provided in the shared data file against the original publications. Because of the large number of papers, I selected 10% of the papers (24 papers) for checking. I first started by selecting an additional 5 papers at random and comparing the coded values to the publication, for practice. For the re-coding of the 24 papers I was blinded to the original coded values. I then compared my recoded values against the original values. Details are provided in the attached code, and my recoded values are included in a spreadsheet titled "recoded_values.xlsx".

For P(nogo), I identified errors in 4/25 of the papers checked, with a maximum absolute difference of 10%.

For SOA values I identified differences in the original coded values from my recoded values for 9 of the 24 papers. Further examination showed that in 4 of these cases, it was not possible to tell which of the values was correct. Coding errors were identified in the remaining five papers, with a maximum absolute difference of 4000 ms for minimum SOA and 1300 ms for maximum SOA.

*Time spent: 2.5 hours*

### 4. Model Misspecification Not applicable

Are there any consequential issues with the assumptions or the form of a statistical model (e.g., overfitting, wrong distribution assumption) used to describe data?

*Time spent: minutes/hours*

### 5. Erroneous/Impossible/Inconsistent Statistical Reporting Not applicable

Are there inconsistencies between test statistics, degrees of freedom, and p-values? Are there implausible degrees of freedom between compared SEM models? Are there point estimates outside the confidence interval bounds?

*Time spent: minutes/hours*

### 6. Other Aspects Related to Data or Code Errors found

I attempted to recreate the Figures 1 and 2 from the provided data; see the computational notebook provided with my review. I was able to qualitatively reproduce each of the figures; there were small differences in the details due to differences in histogram binning, which could not be exactly reproduced due to the lack of original code.

I also compared the results of my reanalysis with the statements made in the paper. I did identify two minor errors in the text. First, on page 8 (Section 3.1), the authors state that "The trial duration parameter had a wider range, spanning from 650 ms to 17,550 ms between successive stimuli." The min duration reported in the text matches the min SOA in the data (variable "SOA (ms)"). The max duration reported in the text doesn't seem to match the max SOA (variable "SOA (ms, max)"), which is 17,500 ms; it seems that the 50 ms buffer was added to these values, even though that was not a study with a fixed trial duration, which was the intended use case for adding 50 ms. This issue also appears in the statement on Page 9, "The mode for the maximum duration between two trials was 2,050 ms, which is 550 ms longer than the fast-paced condition we used in the current study.". The mode of the maximum SOA values computed from the shared data (variable "SOA (ms, max)") is 2000; it appears that the mode was computed after adding 50 ms to some studies. These are minor errors that do not change the interpretation of the results.

*Time spent: 3 hours*

# III. CLAIMS, PRESENTATION, AND INTERPRETATION

### 1. Interpretation Issues No errors found

Throughout the entire paper, is there an incorrect substantive interpretation of data or statistical tests, causal inference issues, etc.?

Note that my analyses only focused on the Literature Analysis portion of the paper. I did not identify any incorrect substantive interpretations of data, or causal inference issues.

*Time spent: 15 minutes*

### 2. Overclaiming Generalisability No errors found

Does the paper overclaim the generalisability of the findings with regards to stimuli, situations, populations, etc.? Is there hyping or overselling of the importance or relevance of findings?

*Time spent: 5 minutes*

### 3. Citation Accuracy Didn't check

Are there misrepresentations of substantive claims by cited sources? Inaccurate direct quotes? Incorrectly cited or interpreted estimates? Citations of retracted papers?

*Time spent: minutes/hours*

### 4. Other Aspects Related to Interpretation No errors found

*Time spent: minutes/hours*

# AUTHOR ЯESPONSE

Jan Wessel

## Background and Manuscript selection

I was approached in November 2023 via email by Dr. Malte Elson, who inquired whether I would be interested in contributing to the pilot of the ERROR initiative. I know Dr. Elson from college and we have been keeping sporadically in touch. I was specifically asked if I would be willing to volunteer one of my past publications to undergo an error investigation according to the ERROR guidelines. I agreed two days later and proposed several papers to Dr. Elson.

My only consideration at that initial point was for it to be a single-authored paper. This would avoid issues with error attribution, potentially adverse consequences for more junior colleagues, or the risk of coercion (i.e., current or former trainee co-authors may not be as enthusiastic about the idea of having their work scrutinized in this way, but may feel the implicit expectation to agree with me).

Dr. Elson then suggested to avoid papers that are not well cited, as well as papers whose methods may be so simple as to severely limit error likelihood.

Hence, we quickly agreed on the paper that has been subject of this review, which had been cited around 250 times according to google scholar at the time we agreed upon it. This paper features an expansive literature analysis and an event-related potential experiment. The reviewer (Dr. Russell Poldrack of Stanford University, from here on referred to as RP) chose to focus on the former component of the work.

## Identification of issues

By my reading of the ERROR report, RP identified three issues overall: one with the Paper Selection for the literature analysis, one with the extraction of the key parameters from the papers in question (Parameter Extraction), and one with the analysis of those parameters (Parameter Analysis). I will address these remarks in the following. I will focus on the two clear-cut ones first, and then discuss the third in more detail.

### Paper identification

I agree with RP that a different set of literature search parameters would have yielded a different set of papers. I will note that the stated goal of the study was not to generate a complete list of all papers published in the time period of interest (which would be very hard to obtain), but to generate a large, representative sample using a sensible database and search query. In the interest of reproducibility, the exact database (PubMed), search query (“inhibition AND human AND go-nogo task”) and search date (October 11th, 2016), were explicitly listed in the paper.

In sum, I concur with RP’s assessment that “it seems highly unlikely that [any] additional papers would have differed systematically in a way that would have changed the result”.

### Parameter analysis

On this point, I was confused about the statement “I attempted to recreate the Figures 1 and 2 […] which could not be exactly reproduced due to the lack of original code.”

All information for this work (including all analysis code, raw data for literature and EEG analysis, task code, etc.) is publicly available on the OSF, and linked on our website (https://wessel.lab.uiowa.edu/open-science). This is the case for all of our studies and has been the case since I started my lab in 2015. In RP’s defense, it is not explicitly mentioned in the paper (which predates the now-ubiquitous “Data availability” statements required by many journals). However, I also received no query from RP whether the code is openly available or not.

The actually-identified error pertains to the presentation of the SOA results. Specifically, it is highlighted that the in-text reported modal values for this parameter appear to be exactly 50ms off between the data table and the paper. **That is correct.**

Looking at the analysis code, the source of the error is obvious. As mentioned in the paper, I had to address the following issue: the way MATLAB plots the x-axis values in Figure 1 makes them impossible to see if a study only had a single SOA value (i.e., if there was no range between the shortest and longest possible SOA value in a given experiment). Therefore, I added an artificial “buffer” of 50ms around the single-SOA value for those papers to make them visible. This process is described in the legend for Figure 1 of the original manuscript. So far, so good. The problem is that I then did not extract the modal values stated in the text from *the original Excel spreadsheet* into which I entered the values from the literature analysis. Instead, I erroneously extracted those modal values from *the buffered MATLAB matrix* that is underlying Figure 1 in the manuscript.

In sum, this is a clear mistake on my part. While I concur with RP that “These are minor errors that do not change the interpretation of the results”, it is still frustrating that this happened. I will talk about ways to avoid this type of error in the final section of this text.

### Parameter extraction

This was the most interesting ‘error’ to me. RP took a subsample of the 241 papers sampled for this research (24 papers) and extracted the parameters for SOA and p(nogo) himself. He then identified several cases in which the values he extracted from the methods section differed from the ones I originally extracted in 2016.

I was frankly astonished by the degree of this mismatch. RP identified 4 instances in which the p(nogo) parameter did not match (though I am taking the liberty to dispute one of them as an error, because I rounded from 31.7% to 32% and he did not). Either way, this still leaves an error rate of 12.5% (3/24). For the SOA value, the mismatch was even more substantial: RP identified 5 cases with purported errors, resulting in a 20.8% error rate. (Note that in this text, I will only focus on the SOA(ms) value, and not the maximum SOA, which was less important for the argument put forward in the original article). RP's subsequent impression that in four additional cases, "it was not possible to tell which of the values was correct" actually presages one of the factors that I will examine in more detail in the following.

Like I said, I was very surprised by this high rate of discord between the two raters during the extraction of these parameters from the papers in question, as well as the high general error rate (both on my part, and, as we will see, on the part of RP). [1]As such, I decided to investigate further.

First, I wanted to identify the *actual* error rate across the complete sample, not just the 24 papers RP re-coded. To do so, I enlisted the help of everyone in my lab that had at least a B.Sc. degree in Psychology (or a related discipline), was a full-time paid researcher, and had at least 3 years of experience working in an academic lab. I assigned each of those six researchers 36 (and one of them 37) of the remaining 217 papers in the sample (241 minus the 24 RP re-coded) and asked them to identify the p(nogo) and SOA parameters from the respective papers, similar to what RP had done. In case of mismatches between the original and newly-identified values, we then re-read the original articles together and attempted to identify what the actually-correct value was. This resulted in one of three possible outcomes: Original value was incorrect, Original value was correct and the value identified by the new rater was incorrect, and Unclear. The latter category includes any instances in which the description of the respective parameter was incomplete or ambiguous. I also applied the same procedure to RP's newly-identified values with one of these new raters – i.e., we checked every instance of discord between RP and my ratings, and attempted to identify the actually-correct parameter value.

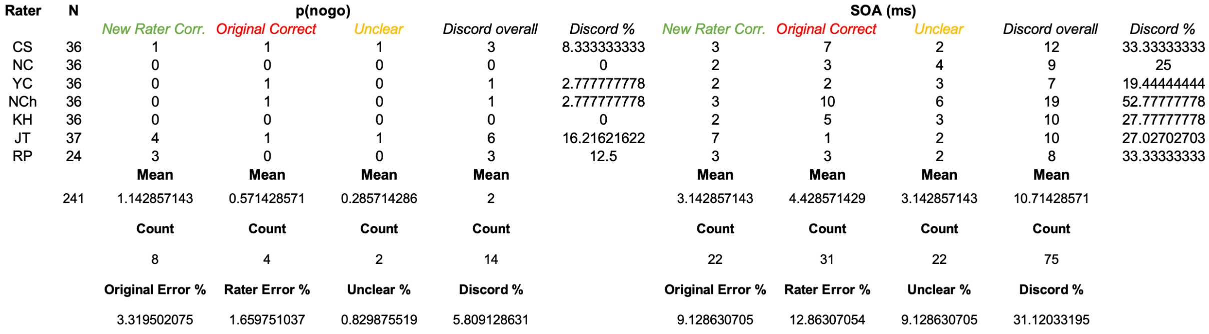
Doing this provided two outcomes. First, it provided a (more) correct version of the original literature analysis spread sheet (i.e., the 'raw' data).[2] Second, this analysis allowed me to dig more into the error rates of the procedure itself.

[1] My entire PhD thesis hinged on a) getting people to make a substantial number of errors in simple, but highly repetitive tasks, and b) getting them to not notice some of those errors. As such, I was not expecting a near-zero error rate for this procedure. But 12 and 20% seemed extremely high.

[2] I am saying "(more) correct" on purpose, because the section "**How to prevent the types of errors that occurred in this study"** contains an analysis that suggests that is statistically likely that even some of the values in which both raters agreed are actually still incorrect.

### *P(nogo) parameter*

Overall, there were 14 instances of discord for the p(nogo) parameter, or 5.79%. In 8 of those cases, the original value was incorrect (3.3%). In 4 cases, the original value was correct and the value identified by the new rater was incorrect (1.7%). 2 cases were unclear. The Cohen's kappa comparing the original ratings and the corrected set of ratings (after the rejoining procedure) was .958. This was calculated using a function from the MATLAB file exchange (https://www.mathworks.com/matlabcentral/fileexchange/69943-simple-cohen-s-kappa). The mean error was .02%, in that the original ratings overestimated the nogo-trial percentage by .02% (or the p(nogo) by .0002). This fortunately suggests little to no systematic bias in the procedure – i.e., the error was evenly distributed around 0 (***Table 1***).



| Rater | N | | | p(nogo) | | | | | SOA (ms) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *New Rater Corr.* | *Original Correct* | *Unclear* | *Discord overall* | *Discord %* | *New Rater Corr.* | *Original Correct* | *Unclear* | *Discord overall* | *Discord %* |
| CS | 36 | 1 | 1 | 1 | 3 | 8.333333333 | 3 | 7 | 2 | 12 | 33.33333333 |
| NC | 36 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 4 | 9 | 25 |
| YC | 36 | 0 | 1 | 0 | 1 | 2.777777778 | 2 | 2 | 3 | 7 | 19.44444444 |
| NCh | 36 | 0 | 1 | 0 | 1 | 2.777777778 | 3 | 10 | 6 | 19 | 52.77777778 |
| KH | 36 | 0 | 0 | 0 | 0 | 0 | 2 | 5 | 3 | 10 | 27.77777778 |
| JT | 37 | 4 | 1 | 1 | 6 | 16.21621622 | 7 | 1 | 2 | 10 | 27.02702703 |
| RP | 24 | 3 | 0 | 0 | 3 | 12.5 | 3 | 3 | 2 | 8 | 33.33333333 |
| | | **Mean** | **Mean** | **Mean** | **Mean** | | **Mean** | **Mean** | **Mean** | **Mean** | |
| | 241 | 1.142857143 | 0.571428571 | 0.285714286 | 2 | | 3.142857143 | 4.428571429 | 3.142857143 | 10.71428571 | |
| | | **Count** | **Count** | **Count** | **Count** | | **Count** | **Count** | **Count** | **Count** | |
| | | 8 | 4 | 2 | 14 | | 22 | 31 | 22 | 75 | |
| | | **Original Error %** | **Rater Error %** | **Unclear %** | **Discord %** | | **Original Error %** | **Rater Error %** | **Unclear %** | **Discord %** | |
| | | 3.319502075 | 1.659751037 | 0.829875519 | 5.809128631 | | 9.128630705 | 12.86307054 | 9.128630705 | 31.12033195 | |

**Table 1.** Result of the complete recoding of the original spreadsheet.

### *SOA parameter*

The discord for the SOA parameter was substantially higher. Overall, there were **75** instances of discord, or **31%**. In 22 of those cases, the original value was incorrect (9.1%). In 31 of those cases, the original value was correct and the value identified by the new rater was incorrect (12.9%). **22** cases were unclear (9.1%). The Cohen's kappa comparing the original ratings and the corrected set of ratings (after the rejoining procedure) was still nominally high, at .908. The mean error was 12.75ms, in that the original ratings underestimated the SOA duration by 12.75ms.[3]

## Summary of the parameter extraction investigation

As mentioned, it is remarkable how low the inter-rater reliability was (despite the nominally high kappa values), especially for a procedure that should have a clear and obvious true answer and little noise in the source signal (i.e., the parameters reported in the

[3] Since the argument in the original manuscript was that go-nogo task SOAs are often too long, what little empirical bias there may have been in the procedure for SOA identification was luckily in the opposite direction of the argument put forward in the original paper.

papers). However, the data also indicate that there actually is a substantial amount of ‘noise’, as obvious from the rate of unclear descriptions in the methods sections of the papers (9.1% even after the rejoinder). This is particularly remarkable since I had already removed 21 additional papers from the original sample (as described in the paper) because the methods were unclear to me at that point. As such, one of the lessons is that the description of methods in published research may require improvement, and / or that methods descriptions should not be perceived as a noiseless signal. (My apologies if that was already obvious to experts in metascience. It wasn’t to me).

Moreover, it is notable that even the *new* raters, who dealt with a substantially smaller sample of papers (24 to 37 instead of 241) and were sensitized to the stated focus of this investigation (viz., scientific error and accuracy), did not show systematically lower error rates compared to (my) original rating. While the new raters did show half the error rate (3.3% vs. 1.7%) for the p(nogo) parameter, they actually showed a higher error rate on the SOA parameter (9.1% vs. 12.9%). *As such, it is probably safe to say that manually extracting methodological parameters from peer-reviewed, published work is a much more ‘noisy’ technique than one may have assumed.*

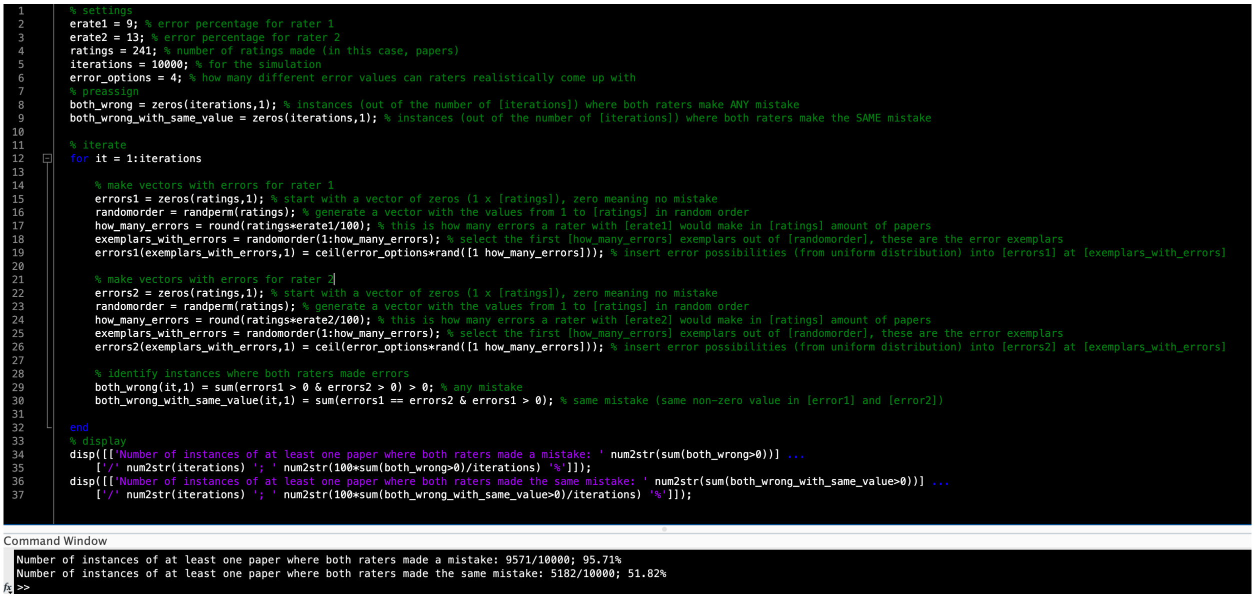
**How to prevent the types of errors that occurred in this study**

Once again, I will touch on this question with respect to every individual aspect highlighted in RP’s ERROR report.

With regard to the topic of Paper Identification, I think the assessment of “Indeterminable” is unjustified. The search terms, search period, database, and date of search were all explicitly listed in the manuscript. I am unsure what could have been done to warrant an assessment of “No errors found”. I assume “Indeterminable” refers to the fact that it is hard (or perhaps impossible) to extract the exact set of results PubMed returned on October 11th, 2016 using PubMed’s online interface.

With regard to the topic of Parameter Extraction, the results of the above analysis suggest that – surprisingly – it may be very hard to avoid errors in this procedure, and that these errors should be treated essentially as measurement noise. In fact, given the base rate of errors across both parameters, it is likely that even the current, corrected sheet of rejoined raw parameter data still contains errors. While I’m sure the exact probability has an analytic solution and can be quantified exactly, I am not a skilled enough mathematician / statistician to figure this out quickly. However, I did write a short snipped of (hopefully error-free) code to run a monte-carlo simulation. This analysis numerically identifies the empirical probability of a scenario in which two raters with error rates of 13% and 9% (see above) both identify the wrong parameter in at least one out of 241 papers. As it turns out, that probability is around **96%** (***Figure 1***). This, of course, assumes that there are only two options to choose from (one correct one and an erroneous one). Of course, with task parameters like SOA (in ms), there are infinite possible responses, and hence, and an infinite number of possible error values. However, after working through the whole spreadsheet and investigating all instances of discord with my lab, it becomes obvious (at least to me) that the errors are not randomly distributed in reality. Instead, there are probably anywhere between 1 and 4 realistic 'error' options for, e.g., the SOA parameter in most papers (indeed, almost all errors resulted from one of the raters not counting one aspect that contributes to SOA, such as the duration of a specific stimulus or inter-trial interval). Therefore, I also simulated the likelihood that both raters not only make any mistake on a given paper, but pick the same wrong value out of 4 possible erroneous values. Even that likelihood was still **greater than 50%** (***Figure 1***).

Hence, it is more likely than not that even the corrected spreadsheet still contains an error.



```
% settings
erate1 = 9; % error percentage for rater 1
erate2 = 13; % error percentage for rater 2
ratings = 241; % number of ratings made (in this case, papers)
iterations = 10000; % for the simulation
error_options = 4; % how many different error values can raters realistically come up with
% preassign
both_wrong = zeros(iterations,1); % instances (out of the number of [iterations]) where both raters make ANY mistake
both_wrong_with_same_value = zeros(iterations,1); % instances (out of the number of [iterations]) where both raters make the SAME mistake

% iterate
for it = 1:iterations

    % make vectors with errors for rater 1
    errors1 = zeros(ratings,1); % start with a vector of zeros (1 x [ratings]), zero meaning no mistake
    randomorder = randperm(ratings); % generate a vector with the values from 1 to [ratings] in random order
    how_many_errors = round(ratings*erate1/100); % this is how many errors a rater with [erate1] would make in [ratings] amount of papers
    exemplars_with_errors = randomorder(1:how_many_errors); % select the first [how_many_errors] exemplars out of [randomorder], these are the error exemplars
    errors1(exemplars_with_errors,1) = ceil(error_options*rand([1 how_many_errors])); % insert error possibilities (from uniform distribution) into [errors1] at [exemplars_with_errors]

    % make vectors with errors for rater 2
    errors2 = zeros(ratings,1); % start with a vector of zeros (1 x [ratings]), zero meaning no mistake
    randomorder = randperm(ratings); % generate a vector with the values from 1 to [ratings] in random order
    how_many_errors = round(ratings*erate2/100); % this is how many errors a rater with [erate2] would make in [ratings] amount of papers
    exemplars_with_errors = randomorder(1:how_many_errors); % select the first [how_many_errors] exemplars out of [randomorder], these are the error exemplars
    errors2(exemplars_with_errors,1) = ceil(error_options*rand([1 how_many_errors])); % insert error possibilities (from uniform distribution) into [errors2] at [exemplars_with_errors]

    % identify instances where both raters made errors
    both_wrong(it,1) = sum(errors1 > 0 & errors2 > 0) > 0; % any mistake
    both_wrong_with_same_value(it,1) = sum(errors1 == errors2 & errors1 > 0); % same mistake (same non-zero value in [error1] and [error2])

end
% display
disp([['Number of instances of at least one paper where both raters made a mistake: ' num2str(sum(both_wrong>0))] ...
    ['/' num2str(iterations) '; ' num2str(100*sum(both_wrong>0)/iterations) '%']]);
disp([['Number of instances of at least one paper where both raters made the same mistake: ' num2str(sum(both_wrong_with_same_value>0))] ...
    ['/' num2str(iterations) '; ' num2str(100*sum(both_wrong_with_same_value>0)/iterations) '%']]);
```

Command Window

```
Number of instances of at least one paper where both raters made a mistake: 9571/10000; 95.71%
Number of instances of at least one paper where both raters made the same mistake: 5182/10000; 51.82%
>>
```

**Figure 1.** MATLAB code for Monte Carlo analysis and output (MATLAB 2023b).

Even if this is true, however – i.e., if even with multiple raters, errors in the final spreadsheet are still likely – it is also clear that the error probability would have been lower had I used another rater beyond myself in 2016. Indeed, we have adopted this approach in my laboratory several years ago: every hand-entered parameter (e.g., a digitally logged value taken from a pen-and-paper form) has to be double-checked by at least one other person. The simple reason why I did not do this back in October 2016 is that my lab (est. 2015) essentially only consisted of myself at that point. That was, of course, a mistake. I did have my first graduate student start in August 2016, and I could have asked her to also rate these papers.

Finally, on the topic of Parameter Analysis, the source of the error was very clearly a mistake I made, resulting in modal values that were off by 50ms. While the consequences of that mistake are minor (according to RP and my own estimation), its likelihood could have been reduced by the same two-eye principle for code – i.e., any segment of code that ends up producing results for a paper should be checked by at least one other person. We have also instituted this process in my lab several years ago.

**Overall impression; Actions to be taken**

I found this to be an enormously enlightening (and honestly somewhat fun) exercise. I am grateful to Malte for approaching me with this idea and to Russ for taking the time to review my work.

I will leave it to the ERROR team to outline the greater conclusions and implications regarding the likelihood of errors in published scientific work.

In regards to this particular study concretely, I will wait for this process to conclude and all the documentation to be publicly available. At that point, I will approach the editor of *Psychophysiology* to suggest a corrigendum to get the reported in-text modal values changed, which will also point out the general issues regarding the 'noise' in the parameter extraction procedure and include explicit reference to the ERROR report. I will defer to the editor's judgment on whether she thinks publishing such a corrigendum is appropriate.

I will also share the corrected spreadsheet in the same OSF folder as the original raw data and code, and will include mention of this ERROR report.