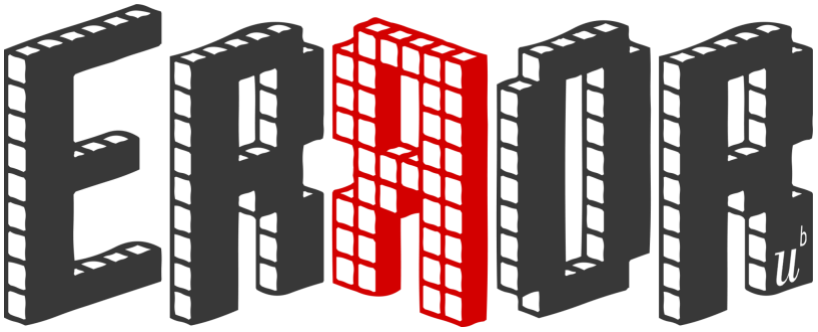


ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

AUTHOR RESPONSE

Wessel, J. R. (2018). Prepotent motor activity and inhibitory control demands in different variants of the go/no-go paradigm. *Psychophysiology*, *55*(3), e12871.

response by

Jan R. Wessel, University of Iowa

April 29th, 2024

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

[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.

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.

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.*

[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 sugests that is statistically likely that even some of the values in which both raters agreed are actually still incorrect.

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 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.*

[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.

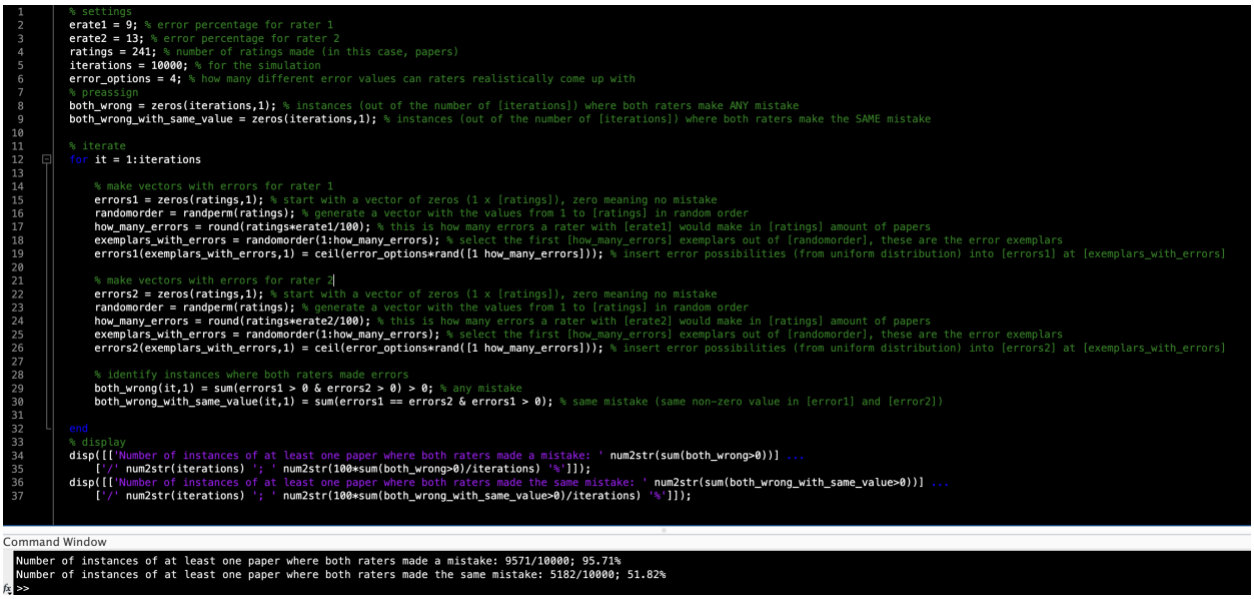
How to prevent the types of errors that occurred in this study

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.