



Universiteit  
Leiden  
The Netherlands

## **The effect of stress on semantic memory retrieval: a multiverse analysis**

Heyman, T.; Boere, R.; Jong, S. de; Hoogeterp, L.; Kraaijenbrink, J.; Kuipers, C.; ... ; Wijk, T. van

### **Citation**

Heyman, T., Boere, R., Jong, S. de, Hoogeterp, L., Kraaijenbrink, J., Kuipers, C., ... Wijk, T. van. (2022). The effect of stress on semantic memory retrieval: a multiverse analysis. *Collabra: Psychology*, 8(1). doi:10.1525/collabra.35745

Version: Publisher's Version


License: [Creative Commons CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/)

Downloaded from: <https://hdl.handle.net/1887/3505223>

**Note:** To cite this publication please use the final published version (if applicable).

Cognitive Psychology

# The Effect of Stress on Semantic Memory Retrieval: A Multiverse Analysis

Tom Heyman<sup>1</sup> <sup>a</sup>, Robin Boere<sup>1</sup>, Sebastiaan de Jong<sup>1</sup>, Lotte Hoogeterp<sup>1</sup>, Joyce Kraaijenbrink<sup>1</sup>, Charlotte Kuipers<sup>1</sup>, Martijn van Dijk<sup>1</sup>, Lotte van Rijn<sup>1</sup>, Twan van Wijk<sup>1</sup>

<sup>1</sup> Leiden University, Leiden, Netherlands

Keywords: multiverse analysis, stress, semantic memory, cortisol

<https://doi.org/10.1525/collabra.35745>

---

## Collabra: Psychology

Vol. 8, Issue 1, 2022

---

Stress is often associated with negative consequences, and this also applies in the context of memory retrieval. However, Smith, Hughes, et al. (2019) proposed that this relationship only holds for information stored in episodic memory, because it relies on the hippocampus. In contrast, conceptual knowledge is stored in semantic memory, which is associated with neocortical and striatal brain regions that are upregulated during stress. Indeed, Smith, Hughes, et al. (2019) found that people experiencing acute psychosocial stress performed better on a subsequent trivia questionnaire compared to a control group. Moreover, performance correlated positively with cortisol reactivity. These findings are important, novel, and perhaps somewhat surprising, hence it is important to establish their generalizability. The latter is accomplished in the present study through a multiverse analysis. The results showed that the effect is relatively robust to variations in the scoring rules for the trivia questionnaire, the type of statistical model being used, and the inclusion versus exclusion of gender as a covariate in the analyses. However, we obtained mostly null effects when using the change in psychological stress levels as a predictor variable, and/or when only considering questions that were actually answered. The latter finding in particular is worrisome as it might point to alternative explanations. That is, stress might improve performance, because participants are more engaged with the task or are more prone to guess, regardless of its (potential) impact on semantic memory retrieval. Hence, it is premature to conclude that stress enhances semantic memory.

Rivers of ink have flown on the impact of stress on human behavior. Within the popular literature, stress often has a negative connotation, yet it might also have positive effects, sometimes referred to as distress and eustress, respectively. This conceptualization of stress in terms of being both detrimental and beneficial, depending on the situation, can be traced back to the seminal work of Yerkes and Dodson (1908) on the relation between stimulus strength and habit formation in mice. They observed that in a simple discrimination task, stimulus strength (i.e., the intensity of an electric shock) and learning were positively related, whereas in a more difficult variant of the task the relation looked like an inverted-U shape. These findings have inspired theories in various fields. For instance, in the context of stress it is translated as follows: performance on simple tasks should improve as the level of stress increases, yet when the task is more complicated, too much or not enough stress is detrimental (Teigen, 1994; but see Corbett, 2015 for a critique).

The present study specifically focuses on the impact of stress on memory retrieval. The Yerkes-Dodson law also resurfaces in this domain, though stress is typically associated with negative outcomes (see Gagnon & Wagner, 2016 for a review). Recently, Smith, Hughes, et al. (2019) argued that the effect of stress depends on the nature of the task at hand: when it involves retrieving information from episodic memory (i.e., context-specific events and relations between events, Tulving, 1972), a heightened stress level indeed has a negative impact, whereas retrieval from semantic memory encompassing general knowledge and the meaning of words (Tulving, 1972), improves as stress increases. Smith, Hughes and colleagues' rationale is that stress downregulates the hippocampus - a brain structure that is considered essential for *episodic* memory retrieval - which explains why memory performance worsens under stress. However, *semantic* memory retrieval does not necessarily require the hippocampus, and can be achieved via neocortical and striatal pathways, whose activity increases under stress, thus yielding beneficial effects. Indeed, Smith, Hughes, et al.

---

a t.d.p.heyman@fsw.leidenuniv.nl

(2019) found that people's accuracy on a general knowledge test, with questions like *What Spanish city is the capital of Catalonia?*, was better when they previously underwent a stress-induction manipulation. In addition, accuracy also correlated positively with cortisol reactivity to stress.

Interestingly, a study by Merz et al. (2016) reported that higher cortisol reactivity *interferes* with the retrieval of conceptual knowledge in a sentence verification task featuring items like *Pressure produces heat*. There are of course a number of differences between both studies, which could explain the seemingly contradictory findings. For instance, it could be that the task in Merz et al. (2016) was more complicated compared to the trivia questionnaire used in Smith, Hughes, et al. (2019), hence the Yerkes-Dodson law would predict a divergent pattern of results (yet Smith, Hughes, et al. (2019) posited that Merz and colleagues' study involved a *low-demand* memory test). Furthermore, the impact of stress manifested itself as slower response times in Merz et al. (2016) and improved accuracy in Smith, Hughes, et al. (2019), so a direct comparison is difficult as it could involve a speed-accuracy tradeoff. Alternatively, the unusual positive impact of stress on memory retrieval observed by Smith, Hughes and colleagues could be a false positive. The present study explores the latter possibility in more detail. More specifically, we examined the generalizability and robustness of Smith, Hughes et al.'s conclusions by considering various alternative data-processing and analysis options.

Researchers in general need to take several steps to get from the raw data to the eventual conclusions (e.g., transforming variables, removing datapoints, and so on). In most cases, there are a number of (equally) plausible alternative pathways that remain unexplored or do not get mentioned (in some instances because they yield undesirable results). In other words, there is a plethora of outcomes of which only one or a few are considered or reported in a typical manuscript. Importantly, to get an idea about the robustness of a given finding (or lack thereof), one could investigate all reasonable alternatives imaginable, in a so-called multiverse analysis (Steege et al., 2016). For example, suppose one conducts a multiverse analysis which encompasses five plausible data exclusion approaches, and three ways to code or transform the dependent variable. This would lead to  $5 \times 3 = 15$  analysis pathways, and a distribution of results (e.g., p-values or parameter estimates), rather than a single outcome. As such, it shows how sensitive the outcome of a study is to different reasonable alternative choices. Indeed, previous studies have illustrated the importance of accounting for so-called researcher degrees of freedom (i.e., the liberty that researchers have when selecting a data exclusion approach, a way to code the dependent variable, etc.; Simmons et al., 2011). A multiverse analysis can be seen as a principled approach to this issue by (re)constructing the underlying decision tree (also referred to as the garden of forking paths, Gelman & Loken, 2014), and examining the outcome of each branch/path. Put differently, a multiverse analysis "enhances transparency by providing a detailed picture of the robustness or fragility of statistical results, and it helps identifying the key choices that conclusions hinge on" (Steege et al., 2016, p. 703).

A study by Credé and Phillips (2017) nicely illustrates the relevance and usefulness of a multiverse analysis. They revisited the power pose effect - the (controversial) finding that holding a high-power body pose influences hormone levels (Carney et al., 2010) - through a multiverse analysis. The plausible alternative pathways considered by Credé and Phillips (2017) yielded mostly null effects, whereas the original single-pathway analysis revealed a statistically significant effect. Hence, it suggests that power posing does not produce a robust effect, at least in terms of hormone levels. Note that multiverse analyses have been shown to provide new insights in various other domains as well (e.g., Moors & Hesselmann, 2019; Steege et al., 2016). As Smith, Hughes and colleagues' (2019) findings are novel, highly relevant to the field, and perhaps somewhat exceptional given that stress is mostly associated with negative effects on memory retrieval, it would be important to establish their generalizability. Thus, the goal of the current study is to explore reasonable alternative pathways to those considered by Smith, Hughes, et al. (2019), using their publicly available data (Smith, 2020), thereby shedding more light on the impact of stress on semantic memory retrieval.

## Method

### Procedure of Smith, Hughes, et al. (2019)

Smith, Hughes, et al. (2019) assigned a total of 92 participants to one of two conditions: a control condition or a condition involving a stress-induction manipulation (i.e., the Trier Social Stress Test, see Von Dawans et al., 2011). Participants in the stress-induction condition had to deliver a speech within the context of a hypothetical job application, and solve math problems aloud, all while being video recorded. In the control condition, participants silently read from a textbook and solved math problems using pen and paper without being recorded. Before and after these tasks, participants in both conditions filled in the State-trait inventory for cognitive and somatic anxiety (henceforth *STICSA*; Grös et al., 2007), and provided a saliva sample from which the cortisol concentration was determined. These two measures provide an indication of participants' psychological and physiological stress levels, respectively, and can be used to determine the effect of the stress manipulation by subtracting the initial values (i.e., pre-*STICSA* and pre-cortisol) from those after the intervention (i.e., post-*STICSA* and post-cortisol; thus yielding delta-*STICSA* and delta-cortisol). Finally, participants completed a general knowledge test involving 122 open-ended trivia questions. Questions were presented one at a time, and participants had a time window of 15 seconds for each question during which they could type in their answer (see the Method section of Smith, Hughes, et al. (2019) for more details).

### Data-analysis of Smith, Hughes, et al. (2019)

Responses to the general knowledge test were classified as correct or incorrect using ad hoc scoring rules. A response was considered correct if it exactly corresponded to the intended response or a common synonym. Answers were also considered correct, if they fell into one of the

following categories: incorrectly pluralized or capitalized responses, partial answers of four letters or more that matched with the correct response, and other misspellings that still conveyed the participant knew the answer. However, whenever participants provided multiple answers to a given question, the response was scored as incorrect.

Smith, Hughes, et al. (2019) then conducted a number of statistical analyses, but we will only focus on the ones that (in our estimation) underlay the conclusion that stress enhances semantic memory. First, Smith, Hughes, et al. (2019) calculated the proportion of correct responses to the trivia questionnaire for each participant by dividing the number of correct responses by the total number of questions (i.e., 122). Next, they performed a two-way ANOVA on proportion correct with gender (male or female) and condition (control or stress-induction) as between-subjects variables. They found a significant main effect of condition  $F(1, 88) = 5.81$ ,  $MSE = 0.00$ ,  $p = .018$ ,  $\hat{\eta}_p^2 = .062$  (as well as a significant main effect of gender  $F(1, 88) = 9.65$ ,  $MSE = 0.00$ ,  $p = .003$ ,  $\hat{\eta}_p^2 = .099$ ; the interaction was not statistically significant  $F(1, 88) = 0.09$ ,  $MSE = 0.00$ ,  $p = .771$ ,  $\hat{\eta}_p^2 = .001$ ). Independent reproduction based on the original data of Smith, Hughes, et al. (2019) (i.e., using their correct/incorrect coding) confirmed these results.

A second crucial analysis involved a multiple linear regression analysis on proportion correct with gender, delta-cortisol, and a gender-by-delta-cortisol interaction as predictor variables (condition was not used as a predictor in this analysis). Smith, Hughes, et al. (2019) reported a significant main effect of delta-cortisol ( $t(85) = 2.08$ ,  $p = .040$ ), with no other effect reaching statistical significance. However, independent reproduction revealed an interesting result. Smith, Hughes, et al. (2019) used dummy coding for gender (i.e., the reference group, *male* in this case, was assigned a value of 0, whereas *female* was coded as 1). Consequently, the reported effect of delta-cortisol corresponds to the arbitrarily assigned baseline category, because the model included a gender-by-delta-cortisol interaction. If we were to switch the baseline to females, the effect of delta-cortisol is as follows:  $t(85) = 0.56$ ,  $p = .575$ . In these situations, it might be advisable to use effect coding, which was the case in the two-way ANOVA, as it presumably better reflects the hypothesis the authors wanted to test. When there are only two categories (i.e., *male* and *female*), one group gets assigned a value of 1, the other a value of -1, and the outcome no longer depends on the arbitrary assignment of values to groups. Using effect coding gave the following outcome for the coefficient of delta-cortisol:  $t(85) = 1.96$ ,  $p = .053$ . In all subsequent analyses, we used effect coding for both gender (-1 for female and 1 for male) and condition (-1 for the control group, and 1 for the stress-induction group), and mean-centered continuous independent variables (e.g., delta-cortisol). Note also that all p-values in Smith, Hughes, et al. (2019) correspond to a two-sided alternative hypothesis even though the effect of interest was directional (i.e., stress *enhances* retrieval from semantic memory). We revisit this issue at the end of the next section.

## Multiverse analysis

### Scoring rules

The scheme used by Smith, Hughes, et al. (2019) is not the only reasonable approach to processing participants' responses. For instance, one could also consider partial responses, defined as matching the first four letters, to be incorrect. After all, one can't be sure that participants indeed knew the correct answer, and simply needed more time to type it in. It is also possible that participants only knew part of the answer. Furthermore, Smith, Hughes, et al. (2019) did not always apply this criterion consistently. For example, the response *nep* to the question *Which planet in our solar system was the last to be discovered?* was coded as correct, even though it only matched the first three letters of the correct answer (i.e., *Neptune*), whereas the response *mada* to the question *Which island's wildlife is 90% unique to that island?* was coded as incorrect, even though it did match the first four letters of the correct answer (i.e., *Madagascar*).

In addition, one could eliminate any subjectivity from the coding process by only considering exact matches as correct answers. Conversely, one could use more lenient scoring rules and consider responses featuring multiple answers, including the right one, as correct. It is conceivable that in at least some of these cases, participants realized they initially made a mistake, after which they proceeded to type the correct answer.

These considerations gave rise to three alternative coding schemes. Two slight variations of Smith, Hughes et al.'s (2019) scheme: one where partial answers were considered incorrect (henceforth *alternative 1*), and one where multiple answers, including the right one, were considered correct (henceforth *alternative 2*). The third scheme only considered responses correct when they exactly matched the intended answer (henceforth *alternative 3*).

The available data from Smith, Hughes, et al. (2019) shows the classification of each response in terms of correct/incorrect without mentioning the subcategories (i.e., exact match, synonym, partial answer, and so on). Hence, we recoded the original responses, first in terms of the various subcategories, which were then translated into correct/incorrect according to the scoring rules mentioned above. Coding was done independently by two of the co-authors. The raters assigned identical subcategories to the participants' responses in 95.95% of the cases. The first author resolved any differences. This procedure also allowed us to compare the results with Smith, Hughes et al.'s correct/incorrect classification when using their scoring rules. The rate of agreement between both was 99.24%, suggesting that the resulting scores were reliable. Nevertheless, in our multiverse analysis, we considered both the original correct/incorrect scoring by Smith, Hughes, et al. (2019), and the one applying Smith, Hughes et al.'s scheme to the recoded data (henceforth *alternative 4*). Adding these two options to the three alternative coding schemes described above, yields five different plausible choices in total.

### **Treatment of omissions**

As the general knowledge test involved open-ended questions, participants sometimes opted to not respond. This could have happened for a variety of reasons: participants didn't know the answer, they didn't want to risk making a mistake, or they were not engaging with the task due to fatigue, lack of motivation, or attentional lapses. Smith, Hughes, et al. (2019) evaluated accuracy relative to the 122 total questions, but another reasonable option would be to use the number of *answered* questions as the baseline instead. When a participant responded to all questions, nothing changes, but suppose a given participant answered 40 questions correctly, 60 incorrectly, and left 22 questions blank, then their accuracy would be 40/122 by Smith, Hughes et al.'s metric and 40/100 according to the alternative metric. The latter is more conservative in the sense that accuracy is only based on questions participants at least attempted to answer, at the expense of potentially disregarding valid information (i.e., questions left open because participants didn't know the answer). Conversely, Smith, Hughes et al.'s metric is more liberal as it equates blanks with unsuccessful attempts to retrieve the answer from semantic memory, yet it could also be the result of other factors. Put differently, the latter approach could introduce a confound in that stress might have improved performance merely because participants were more engaged with the task, or were more prone to guess if they weren't sure about the correct answer. As such, both options have their advantages and disadvantages, and seem plausible in the current context (we revisit this issue in the Discussion section).

### **Critical predictor variable**

The impact of stress on semantic memory retrieval was captured by two variables in Smith, Hughes, et al. (2019): the condition participants were assigned to, and the change in cortisol level after the stress manipulation, regardless of the condition they were in. We decided to also consider the change in psychological stress as predictor variable of interest. Smith, Hughes, et al. (2019) used both physiological and psychological stress markers to evaluate the effectiveness of the stress-induction procedure, thus recognizing the validity of both measures, yet they only used the former as a predictor variable in their analyses of interest (see the section *Memory performance under stress* in Smith, Hughes, et al. (2019) and our reproduction above). Furthermore, Merz et al. (2016) also considered subjective ratings as a potential predictor of performance besides physiological stress responses. Hence, delta-STICSA (see above) seems a reasonable alternative operationalization of the general concept stress.

### **Covariate**

Smith, Hughes, et al. (2019) included gender as a covariate in all their analyses, including the interaction between gender and the critical predictor variable. Following Simmons et al. (2011), we repeated the analyses without the covariate (i.e., omitting the main effect of gender as well as the interaction between gender and the critical predictor variable) in order to establish the generalizability of the ef-

fect. One might argue that this changes the nature of the effect in question (e.g., Del Giudice & Gangestad, 2021). However, gender was not included as a covariate in analyses of related studies conducted by the same authors (Smith, Dijkstra, et al., 2019; Smith et al., 2020). Therefore, it would appear relevant to examine this alternative pathway, if only for exploratory purposes and to increase transparency.

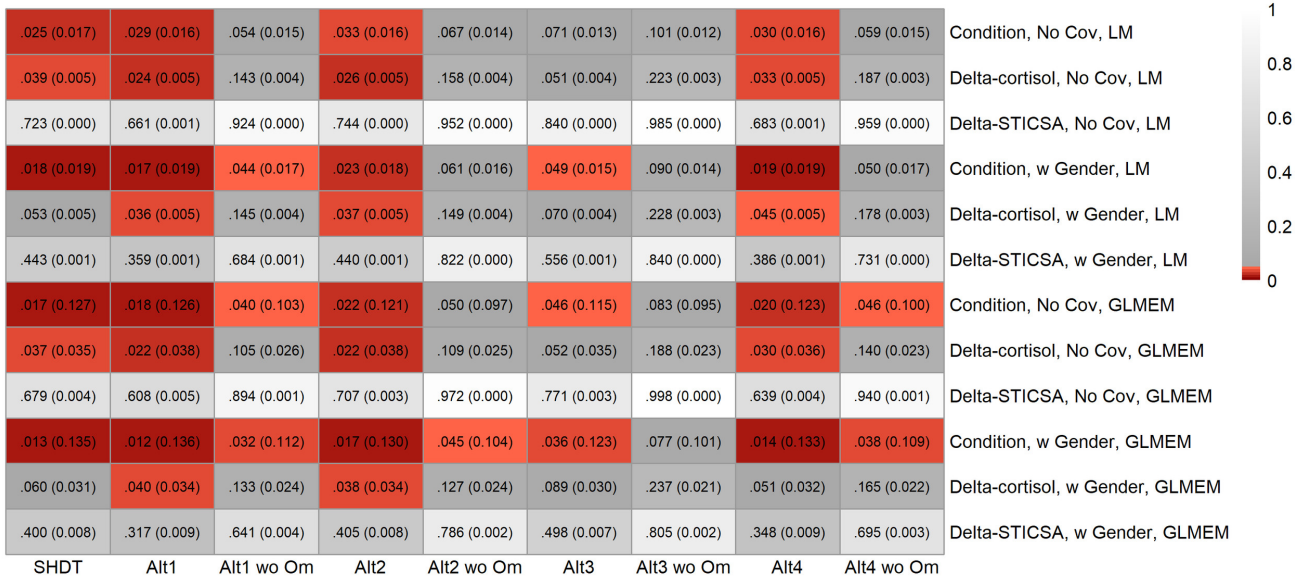
### **Statistical model**

Although it is common practice to perform ANOVAs or linear regression analyses on proportions or percentages, it violates the underlying assumptions (Jaeger, 2008). Furthermore, statisticians have pointed out that materials such as questions, words, or pictures are drawn from a population just like participants are. In other words, these are random effects and should be treated as such if one wishes to draw generalizable conclusions (Clark, 1973). To address both concerns, we decided to conduct logistic regression analyses with crossed random effects for participants and items (i.e., questions), in addition to the ANOVA and linear regression approach taken by Smith, Hughes, et al. (2019). Since ANOVA and linear regression analyses are equivalent, they are put under the same linear model umbrella. So, for each analysis, regardless of scoring rule, treatment of omissions, critical predictor variable and inclusion of gender as a covariate, an equivalent logistic regression model with the same fixed effects as well as random intercepts for participants and items (henceforth *generalized linear mixed effect models*) was fitted using the raw, trial-level data.

If we take all data-processing and -analysis options together, including the choices made by Smith, Hughes, et al. (2019), we get 120 unique outcomes: five scoring rules  $\times$  two treatments of omissions  $\times$  three critical predictor variables  $\times$  two covariate inclusion decisions  $\times$  two statistical models. However, because Smith, Hughes et al.'s correct/incorrect classification did not allow us to retroactively disregard omissions, we ended up with 108 unique outcomes, two of which were of course already considered by Smith, Hughes and colleagues (see above). To be consistent with Smith, Hughes, et al. (2019), we report p-values corresponding to a two-sided alternative hypothesis, which also allows us to detect effects in the opposite directions (i.e., stress inhibiting retrieval from semantic memory). However, to assess the robustness of the effect, we also calculate complementary summary statistics using a directional alternative hypothesis reflecting the expectation that stress enhances retrieval from semantic memory.

## **Results**

The results of the multiverse analysis are visualized in [Figure 1](#). It shows the p-value for the main effect of the critical predictor variable (i.e., condition, delta-cortisol, or delta-STICSA) for every analysis pathway. In addition, it also includes the corresponding point estimate of the regression weight, but note that its interpretation varies across analyses. Overall, the results showed that the choice of critical predictor variable can have a big impact on the eventual conclusion. More specifically, the condition variable often yielded a significant p-value, and so did the delta-cortisol variable, yet to a slightly lesser extent. How-



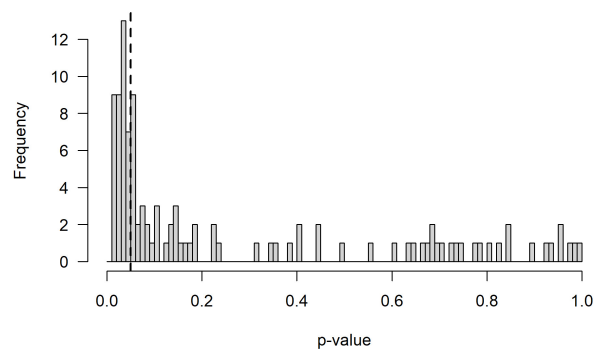
**Figure 1. Visualization of the multiverse of p-values for the main effect of the critical independent variable.**

The point estimate of the corresponding regression weight is added in brackets. Significant p-values according to an alpha level of .05 are displayed in shades of red (the darker, the smaller). Non-significant p-values are displayed in shades of grey (the darker, the smaller). All p-values correspond to a two-sided alternative hypothesis. No Cov means no covariates included, w Gender means with gender as covariate; LM means linear model, GLMEM means generalized linear mixed effects model; wo Om means without omissions (i.e., omissions were disregarded); SHDT stands for the original scoring by Smith, Hughes, et al. (2019); AltX means scoring alternative X (i.e., 1 to 4).

ever, the p-values associated with the delta-STICSA variable were never statistically significant. Furthermore, the decision to discount answers left blank by the participants also resulted in higher p-values on average, yet some still reached statistical significance. Scoring rules for responses appeared to have a lesser impact, even though the most stringent alternative, in which only exact matches were considered correct (i.e., *alternative 3*), gave rise to higher p-values and lower regression weight estimates overall. Finally, the inclusion of gender as a covariate, and the type of analysis (linear model versus generalized linear mixed effects model) did not appear to consistently affect the results. The latter did affect the estimates of the regression weights, because their meaning is rather different. In the linear models, they express expected change in proportion correct responses, whereas in the generalized linear mixed effects models they refer to changes in log odds.

To the extent that all alternatives can be considered equivalent (see Discussion section), we can also evaluate the distribution of the resulting p-values (see Figure 2). 38 out of 108 p-values (35.19%) are considered statistically significant assuming an alpha-level of .05. With a more liberal criterion of .10, that number increases to 55 (50.93%). If we use the two p-values obtained by Smith, Hughes, et al. (2019) as a reference (i.e., .018 and .053), we found that, respectively, 101 (93.52%) and 65 (60.19%) of the p-values from the multiverse analysis were larger.

To summarize the results of a multiverse analysis, Steegen et al. (2016) also suggested to take the arithmetic mean of the obtained p-values. The idea is based on the notion that a typical study would randomly select a single-pathway analysis from all the identified choices (i.e., one out of the 108 options in this case). Therefore, the average p-value based on all alternatives of the multiverse can be viewed as a proxy for the p-value one would get based on a single-



**Figure 2. Distribution of the p-values for the main effect of the critical independent variable.**

The dotted line indicates  $p = .05$ .

pathway analysis. For the current multiverse, this would give a p-value of .276 when assuming a non-directional alternative hypothesis, and .139 when assuming a directional alternative hypothesis. However, the arithmetic mean might be deceiving in that substantially different p-values, say .001 versus .0000001, can have a negligible impact on the mean. Compare the following set of hypothetical p-values from a two-pathway multiverse:  $p = .10$  and  $p = .001$  versus  $p = .10$  and  $p = .0000001$ . Both yield an arithmetic mean p-value of .05, yet the single pathway analyses tell a rather different story.

Another approach is to consider the harmonic mean of the obtained p-values. Wilson (2019) has proposed such a procedure to control the familywise error rate in the context of multiple testing. In particular, when tests are de-

pendent, as is the case in a multiverse analysis, procedures such as Bonferroni are too conservative (i.e., one would need a  $p$ -value of  $\alpha/108$  to be considered statistically significant in the present study). Instead, the asymptotically exact harmonic mean  $p$ -value procedure can be applied to a set of dependent  $p$ -values to test the null hypothesis that there is no association between stress and retrieval from semantic memory for *any* of the specifications considered in our multiverse. Running this test yielded a  $p$ -value of .091 when testing non-directional alternative hypotheses, and .036 when testing directional alternative hypotheses. Hence, in the latter case we can say that in at least one specification under consideration, we can reject the null hypothesis of no association between stress and retrieval from semantic memory. However, in terms of assessing robustness of the effect, this finding is arguably not that compelling (see Artner et al., 2021 for further discussion).

Finally, we also explored the outcome for the main effect of gender (Figure 3), and the interaction between gender and the critical predictor (Figure 4). As this only makes sense for pathways that include gender, these multiverses are half the size of the one assessing the effect of stress (analyses with gender as the only predictor were not considered relevant, given the goal of the study). In line with Smith, Hughes, et al. (2019), none of the pathways yielded a statistically significant interaction effect. The main effect of gender was statistically significant in most pathways, except in many of those that involved discounting answers left blank by the participants. This may seem to contrast with the outcome of Smith, Hughes, et al. (2019), as they also reported a null effect of gender in their analyses featuring delta-cortisol (and no discounting of blanks). Note though that we mean-centered delta-cortisol (and delta-STICSA), which was not the case in Smith, Hughes, et al. (2019). Hence, they actually assessed the effect of gender when delta-cortisol is zero, thus explaining the discrepancy. Critically, this transformation does not impact the interpretation of the regression weights involving delta-cortisol or delta-STICSA.

## Discussion

The present study re-evaluated the claim by Smith, Hughes, et al. (2019) that stress enhances semantic memory retrieval, through a multiverse analysis. The results suggested that the effect is robust to changes in the way participants' responses were scored as correct or incorrect, the inclusion versus exclusion of gender as a covariate, and the use of different analysis types (i.e., linear model versus generalized linear mixed effects model). In contrast, when the critical predictor variable was operationalized as the change in psychological stress level, none of the corresponding  $p$ -values came close to being considered statistically significant. Furthermore, pathways that did not consider questions left open also yielded mostly non-significant  $p$ -values, but less so when assuming a directional alternative hypothesis.

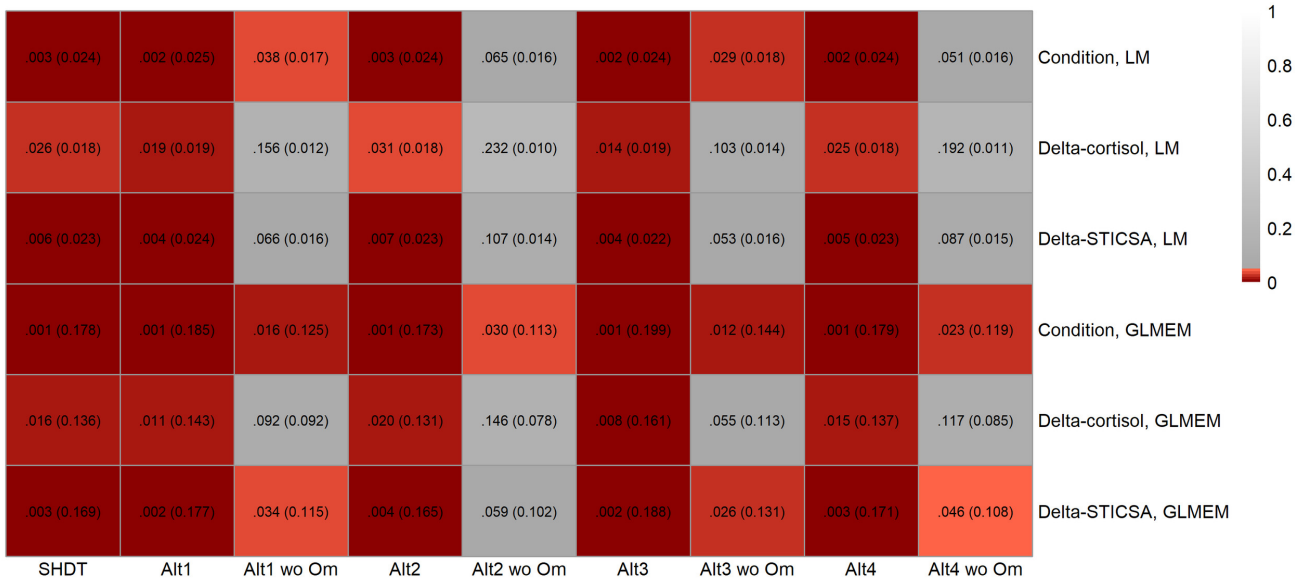
Taken together, the outcome of this analysis is somewhat mixed. On the one hand, the effect does not appear to be a fluke, nor does it seem to arise as a result of biased data-analytic decisions, though a multiverse analysis is no sub-

stitute for an independent replication study. On the other hand, the results may point towards potential boundary conditions of the effect, yet it is important to note that non-significant  $p$ -values should not be considered as evidence for the null hypothesis (Dienes, 2014), and differences in "statistical significance" do not necessarily translate to significant differences (Gelman & Stern, 2006). In the remainder of the discussion, we focus on the implications of the results for the different measures of stress, and the different treatments of omissions. We end with some general considerations regarding multiverse analyses.

## Measures of stress

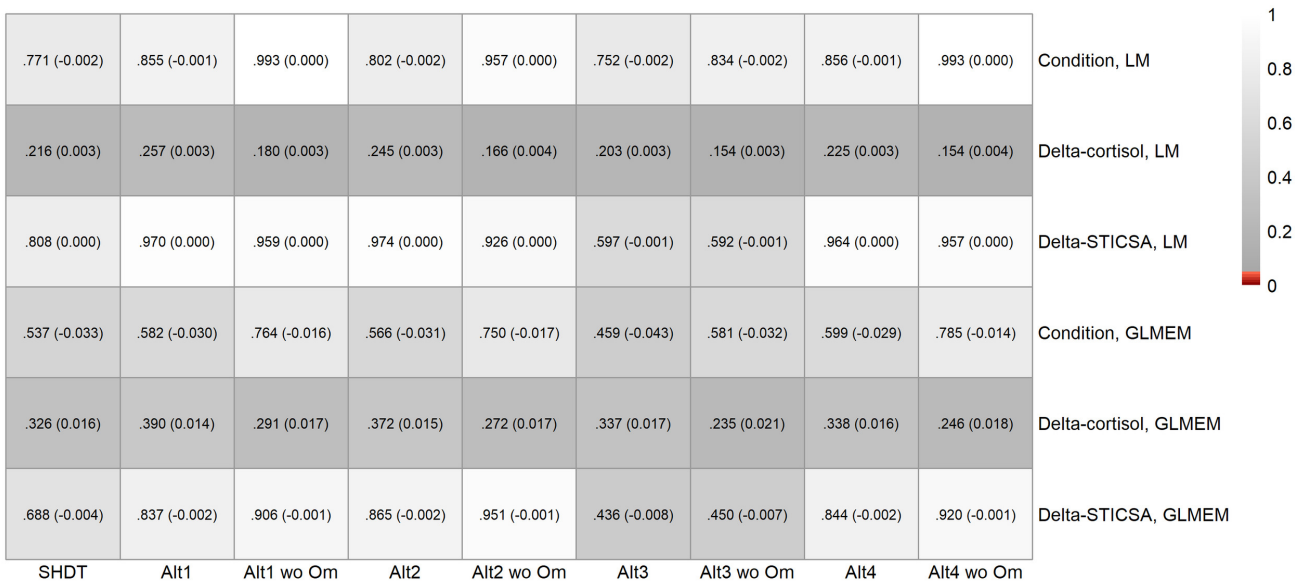
The operationalization of the concept stress in terms of delta-STICSA (psychological stress reactivity) consistently yielded null effects in terms of explaining performance on the general knowledge questionnaire. These could be Type 2 errors, but it is noteworthy that Merz et al. (2016) also reported no significant correlation between subjective stress appraisals and performance on a sentence verification task tapping into semantic memory. Furthermore, the correlation between STICSA scores and cortisol levels in Smith, Hughes et al.'s (2019) study turned out to be very low (i.e., -.04 when assessed before the Trier Social Stress Test, and .08 afterwards), which is in line with research from other domains suggesting that there is only a weak association, if any, between cortisol levels and self-report measures of stress (Gidlow et al., 2016; Hjortskov et al., 2004; Maina et al., 2008). A possible explanation is that measures of psychological stress depend on introspection, and people may differ in their ability to accurately estimate their stress level. Consequently, the employed assessment of psychological stress might be less valid and/or reliable compared to that of physiological stress, for example, which could explain our findings. Indeed, as pointed out by an anonymous reviewer, STICSA was designed to measure state and trait anxiety. Even though research has shown that measures of anxiety and stress are closely related (e.g., Lovibond & Lovibond, 1995), one might view them as distinct constructs. For example, Tindall et al. (2021) assessed the *divergent* validity of STICSA (in part) by comparing it to the DASS stress scale (Lovibond & Lovibond, 1995).

From that perspective, analyses involving the psychological stress variable can be considered nonequivalent pathways, and should be considered separately (Del Giudice & Gangestad, 2021). That being said, this explanation is post hoc and might be driven by the results we obtained. Suppose that the psychological stress variable correlated significantly with performance and the physiological stress variable did not. In that scenario, we would have perhaps looked for signs that the assessment of physiological stress yielded less reliable or valid outcomes. Moreover, if the roles were reversed, would it have changed the conclusion that stress enhances semantic memory? Possibly not, which illustrates the flexibility one has to find support for such general notions. In sum, this discussion highlights the relevance of multiverse analyses and the importance of establishing the reliability and validity of the different measures.



**Figure 3. Visualization of the multiverse of p-values for the main effect of gender.**

The point estimate of the corresponding regression weight is added in brackets. Significant p-values according to an alpha level of .05 are displayed in shades of red (the darker, the smaller). Non-significant p-values are displayed in shades of grey (the darker, the smaller). All p-values correspond to a two-sided alternative hypothesis. LM means linear model, GLMEM means generalized linear mixed effects model; wo Om means without omissions (i.e., omissions were disregarded); SHDT stands for the original scoring by Smith, Hughes, et al. (2019); AltX means scoring alternative X (i.e., 1 to 4).



**Figure 4. Visualization of the multiverse of p-values for the interaction effect of the critical independent variable and gender.**

The point estimate of the corresponding regression weight is added in brackets. Significant p-values according to an alpha level of .05 are displayed in shades of red (the darker, the smaller), yet in this case there are none. Non-significant p-values are displayed in shades of grey (the darker, the smaller). All p-values correspond to a two-sided alternative hypothesis. LM means linear model, GLMEM means generalized linear mixed effects model; wo Om means without omissions (i.e., omissions were disregarded); SHDT stands for the original scoring by Smith, Hughes, et al. (2019); AltX means scoring alternative X (i.e., 1 to 4).

### Treatment of omissions

Another outcome of the current study is that discounting questions left open invariably resulted in higher, often non-significant p-values. As mentioned before, one might argue that this measure essentially throws away useful information (i.e., participants might leave questions open because they fail to retrieve the answer from semantic memory).

For example, if one were to implement this in the context of exams, it would systematically bias the resulting scores. Students could leave all but one question open, and get a perfect score, presuming they answer the one question correctly. However, the latter would imply prior knowledge about how answers would be scored, which did not apply to the participants in Smith, Hughes et al.'s (2019) study. In addition, (most) students are motivated to do well on exams



or at least obtain a passing grade, whereas participants in Smith, Hughes et al.'s study did not have much of an incentive (i.e., they merely fulfilled a research participation requirement). Hence, one could also flip the argument in that equating non-responses to a failure to retrieve certain information from semantic memory, might induce confounds.

The latter notion receives support if we break down the data from Smith, Hughes, et al. (2019) by response type (i.e., omissions, correct responses, and incorrect responses excluding omissions). Looking at the sample means, the percentage of incorrect responses is about the same in both conditions; if anything it is slightly lower in the control group (65.45%) compared to the stress-induction group (65.24%). In contrast, participants in the control group tended to leave more questions blank relative to the stress-induction group (11.64% and 15.25%, respectively). Following Smith, Hughes, et al. (2019), we also broke this down by gender (see Figure 5). When comparing the stress-induction condition with the control condition, female participants followed the general trend (which is not surprising given that 63 out of 92 participants were female): the accuracy boost goes hand in hand with a decrease in questions left blank. In addition, female participants in the stress-induction condition gave slightly more incorrect answers too. Male participants overall showed a slightly smaller accuracy boost, which was reflected in a small decrease in both incorrect responses and questions left open. Taken together, the beneficial effect of stress on accuracy (i.e., respectively 22.92% and 19.51% correct on average in the control and stress-induction condition across genders) primarily manifests itself as the reduction of the item non-response rate. Of course, this could be because stress enhances semantic memory retrieval as proposed by Smith, Hughes and colleagues (2019). However, there are several alternative explanations for the effect.

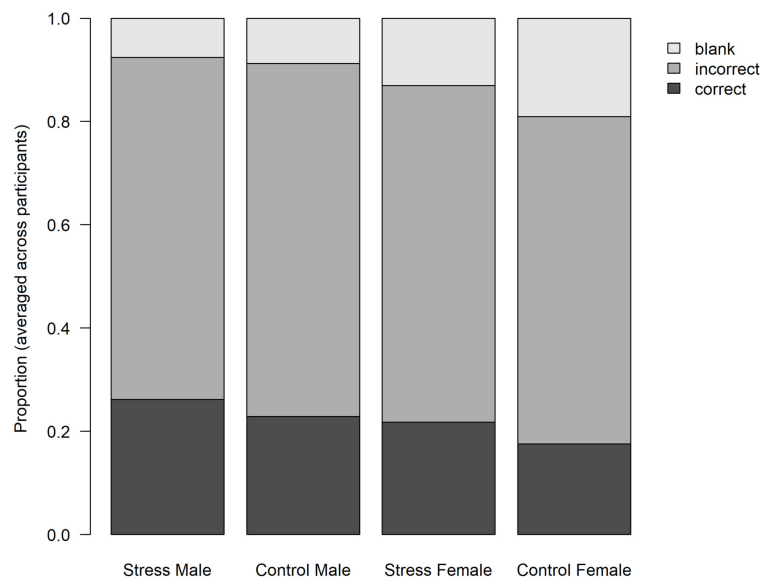
Stress might, for instance, improve participants' focus. Indeed, increased stress levels are in general associated with better performance on relatively easy tasks according to the Yerkes-Dodson principle. In other words, stress might keep participants engaged with the task rather than enhancing semantic memory per se, which can explain why their accuracy is only significantly better when omissions are treated as an incorrect answer. Another possibility is that participants under stress might be less inclined to leave questions open when in doubt. As stress has been argued to increase risk taking (e.g., Buckert et al., 2014), it could translate into a higher propensity to follow their gut or guess when participants were not sure about the correct answer. If this hypothesis were true, one would expect to find a similar pattern of results as obtained by Smith, Hughes, et al. (2019), particularly that of the female participants. In sum, because there are several plausible alternative explanations for the current findings, it seems premature to conclude that stress enhances semantic memory retrieval.

As an aside, Smith, Hughes, et al. (2019) claimed that stress "did not influence the number of incorrect responses that participants offered", and, importantly, that "stress did not influence participants' tendency to leave questions blank" (p. 41). These assertions were based on additional analyses with condition and gender as predictor variables. They found that condition was not significantly associated

with proportion omissions, relative to the 122 total questions, nor with proportion incorrect responses excluding omissions, also relative to the 122 total questions. However, it is a common mistake to treat non-significant p-values as evidence for the null hypothesis of no effect (Dienes, 2014). Furthermore, from Smith, Hughes and colleagues' claim that stress increases correct responding follows logically that it must decrease incorrect responding and/or errors of omission as these variables are communicating vessels (i.e., percentage correct responses + percentage incorrect responses not counting omissions + percentage omissions = 100%; this is also illustrated by Figure 5). Consequently, one can not conclude from these results that stress does not impact participants' tendency to leave questions blank.

Considering the break-down by condition reported above, the conclusions from Smith, Hughes, et al. (2019) may need to be re-assessed. However, it is important to point out, as we did in the discussion about psychological stress, that the nature of these analyses is exploratory. The alternative explanations offered here, were derived post hoc, and the two approaches of dealing with errors of omission are arguably not equivalent (Del Giudice & Gangestad, 2021). Critically, neither operationalization of the dependent variable appears superior, though it is possible, even likely, that other researchers have a different perspective on this (Simonsohn et al., 2020). Furthermore, future studies should aim to *deflate* the multiverse as much as possible by carefully considering the study's design and materials as well as the underlying theoretical framework (Steege et al., 2016).

For example, it might make sense to mix in different types of questions that do not tap into semantic memory (e.g., questions assessing working memory capacity or perceptual reasoning, or questions about the Planet Earth video all participants saw as part of the stress induction manipulation). If stressed participants only experience a benefit for the items tapping into semantic memory, or if the benefit is greater for those items, then one could be more confident about the claim that stress enhances semantic memory retrieval. Moreover, one might wonder whether a general knowledge test is the best tool to assess semantic memory retrieval, even if we can account for factors such as motivation, attention, tendency to guess, and so forth. An incorrect response or the lack of a response to a question might reflect failure to retrieve information from semantic memory, or it might simply be due to ignorance: someone might never be exposed to the notion that, for example, the polar bear is the largest bear on earth, or they might have erroneously learned that the brown bear is the largest. As such, it might make sense to consider different paradigms, such as a forced choice semantic categorization task wherein participants need to decide as fast as possible whether a picture or word represents, say, an animate or inanimate concept. One would presumably expect faster (and more accurate) decisions when participants are stressed. Critically, one would again need a negative control to rule out alternative explanations (e.g., detecting a letter or visual feature in the same stimuli). However, note that Smith, Hughes, et al. (2019) speculated that the beneficial effect of stress only emerges when using more effortful



**Figure 5. Distribution of responses types grouped by condition (control or stress) and gender (male or female).**

Note that if one only considers the correct responses, the figure is identical to Figure 3 of Smith, Hughes, et al. (2019), barring cosmetic changes.

memory tasks, so a semantic categorization task might not be ideal in this sense. In sum, the outcomes of the current study suggest avenues for future research, and, in the meantime, the claim that stress enhances retrieval from semantic memory should be treated with due caution.

### Concluding thoughts regarding multiverse analyses

The present multiverse is extensive, yet it does not exhaust all reasonable alternative pathways imaginable. One could, for instance, envision even other scoring rules for the trivia questionnaire than the ones we considered here. Furthermore, we measured performance only in terms of accuracy while ignoring response times, we didn't inspect the data for any potential outliers, nor did we attempt to replace missing values via imputation procedures. Indeed, there is typically no single, definitive multiverse analysis, but rather a multiverse of multiverse analyses. As such, a multiverse analysis is susceptible to biases, just like single-pathway analyses are, though to a lesser degree. Given that the interpretation of multiverse analyses often come down to eyeballing histograms like [Figure 2](#), it is possible to add or remove outcomes in order to obtain a more desirable outcome. For example, omitting the different covariate and/or statistical model pathways from the current analysis would have fueled the perception that the effect is fragile, whereas adding many more slight modifications of the scoring rules could have given the impression that the effect is very robust. That being said, assessing the result of a typical single-pathway analysis is even more difficult as there is no way of telling whether the outcome hinges on that one particular, potentially hand-picked, constellation of data processing and modelling choices.

Pre-registration of the analysis plan is often put forth as a solution to prevent the exploitation of such researcher degrees of freedom (Nosek et al., 2018; Wicherts et al., 2016). Pre-registration preferably occurs before initiating the data collection, but many multiverse analyses, including the present one, involve existing data gathered by a different team of researchers. Even solutions like blind analyses, or splitting data into training and test sets, are not straightforward to apply, because one ideally needs to establish analytic reproducibility first, which requires access to the entire, unblinded dataset. Though it is not a necessary condition for undertaking a multiverse analysis, being able to reproduce the results reported in the original article instills confidence that one has correctly interpreted the data. As such, pre-registration or blinding might not be practically feasible, or may fail to restrict researcher degrees of freedom to a certain extent in these kind of situations.

Nevertheless, we mitigated the risk of bias by using a type of many-analysts procedure (Silberzahn et al., 2018). That is, the eight co-authors of this paper came up with their own multiverse analysis as part of their Bachelor's thesis. The choices that were well-motivated and most in line with the existing literature in the eyes of the first author, got selected for the eventual multiverse analysis presented here. Taken together, the outcome of all multiverse analyses were relatively similar (see Heyman & Vanpaemel, 2021 for a visualization of all individual multiverse analyses), hence the impact of cognitive biases on the outcome of this study is presumably minimal. One may object that the employed procedure is difficult to replicate (e.g., when is a choice well-motivated). However, this is by no means unique to a multiverse analysis. How does one define an outlier criterion, scoring rules, and the like in a typical single-pathway analysis? Usually, based on a discussion be-

tween the authors, consultation of the literature, and/or post-hoc rationalizations after seeing the data, but this is rarely reported. The current approach at least made that process explicit and more transparent.

Finally, one might wonder about how large the sample should be when conducting a multiverse analysis. By definition, a multiverse analysis does not produce a single outcome, so in that sense there is no straightforward answer to that question. One could, for example, consider the statistical power associated with each pathway within the multiverse. Two of the 108 pathways considered in our study were outlined by Smith, Hughes, et al. (2019), hence the corresponding statistical power is identical. As for the other pathways, it is often (implicitly) assumed, also in the current case, that the a priori statistical power is approximately equal. If particular alternatives would result in a substantially lower statistical power (e.g., because an exclusion criterion greatly reduces sample size without improving reliability), one could argue that they are inferior and should not be included in the multiverse to begin with, or that they should be considered separately from the other pathways (Del Giudice & Gangestad, 2021). When it comes to summarizing the outcomes of a multiverse analysis by, for instance, calculating the harmonic mean p-value (see also Simonsohn et al., 2020 for other approaches), one would need to take many factors into account (i.a., level of dependency between the various test, potential variability in the effect across alternative specifications, etc.). Synthesizing the outcomes of a multiverse analysis (when desirable) is a challenging endeavor (see e.g. Artner et al., 2021), which requires more attention in future research, also in terms of sample size considerations. Finally, when there is uncertainty about whether pathways in a multiverse are equivalent, like in the current study, one should be careful when interpreting summary statistics or plots like [Figure 2](#). In such cases, the focus lies (more) on exploration and hypothesis-generation, hence tests involving all pathways under consideration, and their statistical power, might only be an afterthought.

## Conclusion

A recent study by Smith, Hughes, et al. (2019) seemed to show that people experiencing stress performed better on a trivia questionnaire, which was taken to mean that stress enhances semantic memory retrieval. However, this conclusion was based on two particular data-processing and

-analysis pathways, and the present study sought to examine whether other options would yield a similar outcome. It turned out that the results were relatively robust to alternative coding schemes of participants' responses, the inclusion or exclusion of gender as a covariate, and the use of different analysis types. In contrast, using psychological stress as the critical predictor variable yielded null effects, and the same happened, though to a lesser extent, when discounting questions that were left open by participants. The latter inspired some follow-up data-exploration, which suggested that the beneficial effect of stress primarily manifests itself as a reduction in the non-response rate: participants left fewer questions unanswered when stressed. Importantly, this opens the door for alternative explanations of Smith, Hughes, et al.'s results (e.g., stress increases risk taking and/or improves people's focus). Thus, we deem it premature to conclude that stress enhances semantic memory retrieval.

## Data, code and materials

The manuscript was written in R (R Core Team, 2016) using the packages *papaja* (Aust & Barth, 2017) and *rmarkdown* (Allaire et al., 2016). On the project's OSF page (<https://osf.io/rh54b/>) one can find the data and .Rmd file. The analysis code is available in a Code Ocean container (<https://doi.org/10.24433/CO.3539342.v1>).

## Author Contributions

TH developed the design of the study. RB, SDJ, LH, JK, CK, MVD, LVR, and TVW independently developed a multiverse analysis as part of their Bachelor's thesis. TH compiled the final multiverse analysis presented here, and drafted the manuscript, with RB, SDJ, LH, JK, CK, MVD, LVR, and TVW providing feedback. All authors approved the final version for submission.

## Competing interests

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Submitted: December 07, 2021 PDT, Accepted: May 03, 2022 PDT



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

## References

- Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., & Hyndman, R. (2016). *rmarkdown: Dynamic Documents for R*. <http://CRAN.R-project.org/package=rmarkdown>
- Artner, R., Lafit, G., Vanpaemel, W., & Tuerlinckx, F. (2021). *A statistical investigation of the specification curve analysis procedure* [Manuscript submitted for publication].
- Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. <https://github.com/crs/papaja>
- Buckert, M., Schwieren, C., Kudielka, B. M., & Fiebach, C. J. (2014). Acute stress affects risk taking but not ambiguity aversion. *Frontiers in Neuroscience*, *8*, 82. <https://doi.org/10.3389/fnins.2014.00082>
- Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*(10), 1363–1368. <https://doi.org/10.1177/0956797610383437>
- Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. [https://doi.org/10.1016/s0022-5371\(73\)80014-3](https://doi.org/10.1016/s0022-5371(73)80014-3)
- Corbett, M. (2015). From law to folklore: Work stress and the Yerkes-Dodson Law. *Journal of Managerial Psychology*, *30*(6), 741–752. <https://doi.org/10.1108/jmp-03-2013-0085>
- Credé, M., & Phillips, L. A. (2017). Revisiting the power pose effect: How robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological and Personality Science*, *8*(5), 493–499. <https://doi.org/10.1177/1948550617714584>
- Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–15. <https://doi.org/10.1177/2515245920954925>
- Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. <https://doi.org/10.3389/fpsyg.2014.00781>
- Gagnon, S. A., & Wagner, A. D. (2016). Acute stress and episodic memory retrieval: Neurobiological mechanisms and behavioral consequences. *Annals of the New York Academy of Sciences*, *1369*(1), 55–75. <https://doi.org/10.1111/nyas.12996>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465. <https://doi.org/10.1511/2014.111.460>
- Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. <https://doi.org/10.1198/000313006x152649>
- Gidlow, C. J., Randall, J., Gillman, J., Silk, S., & Jones, M. V. (2016). Hair cortisol and self-reported stress in healthy, working adults. *Psychoneuroendocrinology*, *63*, 163–169. <https://doi.org/10.1016/j.psyneuen.2015.09.022>
- Grös, D. F., Antony, M. M., Simms, L. J., & McCabe, R. E. (2007). Psychometric properties of the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA): Comparison to the State-Trait Anxiety Inventory (STAI). *Psychological Assessment*, *19*(4), 369–381. <https://doi.org/10.1037/1040-3590.19.4.369>
- Heyman, T., & Vanpaemel, W. (2021). *Multiverse analyses in the classroom*. <https://doi.org/10.31234/osf.io/4eh6b>
- Hjortskov, N., Garde, A. H., Ørbæk, P., & Hansen, Å. M. (2004). Evaluation of salivary cortisol as a biomarker of self-reported mental stress in field studies. *Stress and Health*, *20*(2), 91–98. <https://doi.org/10.1002/smi.1000>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. <https://doi.org/10.1016/j.jml.2007.1.007>
- Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour Research and Therapy*, *33*(3), 335–343. [https://doi.org/10.1016/0005-7967\(94\)00075-u](https://doi.org/10.1016/0005-7967(94)00075-u)
- Maina, G., Palmas, A., & Filon, F. L. (2008). Relationship between self-reported mental stressors at the workplace and salivary cortisol. *International Archives of Occupational and Environmental Health*, *81*(4), 391–400. <https://doi.org/10.1007/s00420-007-0224-x>
- Merz, C. J., Dietsch, F., & Schneider, M. (2016). The impact of psychosocial stress on conceptual knowledge retrieval. *Neurobiology of Learning and Memory*, *134*, 392–399. <https://doi.org/10.1016/j.nlm.2016.08.020>
- Moors, P., & Hesselmann, G. (2019). Unconscious arithmetic: Assessing the robustness of the results reported by Karpinski, Briggs, and Yale (2018). *Consciousness and Cognition*, *68*, 97–106. <https://doi.org/10.1016/j.concog.2019.01.003>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

- 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. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11), 1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>
- Smith, A. M. (2020). *Acute stress enhances general-knowledge semantic memory*. Open Science Framework. <https://doi.org/10.17605/OSF.IO/EQ8SY>
- Smith, A. M., Dijkstra, K., Gordon, L. T., Romero, L. M., & Thomas, A. K. (2019). An investigation into the impact of acute stress on encoding in older adults. *Aging, Neuropsychology, and Cognition*, 26(5), 749–766. <https://doi.org/10.1080/13825585.2018.1524438>
- Smith, A. M., Elliott, G., Hughes, G. I., Feinn, R. S., & Brunyé, T. T. (2020). Acute stress improves analogical reasoning: Examining the roles of stress hormones and long-term memory. *Thinking & Reasoning*, 27(2), 294–318. <https://doi.org/10.1080/13546783.2020.1819416>
- Smith, A. M., Hughes, G. I., Davis, F. C., & Thomas, A. K. (2019). Acute stress enhances general-knowledge semantic memory. *Hormones and Behavior*, 109, 38–43. <https://doi.org/10.1016/j.yhbeh.2019.02.003>
- Steege, 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>
- Teigen, K. H. (1994). Yerkes-Dodson: A law for all seasons. *Theory & Psychology*, 4(4), 525–547. <https://doi.org/10.1177/0959354394044004>
- Tindall, I. K., Curtis, G. J., & Locke, V. (2021). Dimensionality and measurement invariance of the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA) and validity comparison With measures of negative emotionality. *Frontiers in Psychology*, 12, 644889. <https://doi.org/10.3389/fpsyg.2021.644889>
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). Academic Press.
- Von Dawans, B., Kirschbaum, C., & Heinrichs, M. (2011). The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format. *Psychoneuroendocrinology*, 36(4), 514–522. <https://doi.org/10.1016/j.psyneuen.2010.08.004>
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Van Aert, R. C. M., & Van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7, 1832. <https://doi.org/10.3389/fpsyg.2016.01832>
- Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4), 1195–1200. <https://doi.org/10.1073/pnas.1814092116>
- Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5), 459–482. <https://doi.org/10.1002/cne.920180503>

## Supplementary Materials

### Peer Review History

Download: [https://collabra.scholasticahq.com/article/35745-the-effect-of-stress-on-semantic-memory-retrieval-a-multiverse-analysis/attachment/90406.docx?auth\\_token=vEG6f5pErEkQPzuS86eR](https://collabra.scholasticahq.com/article/35745-the-effect-of-stress-on-semantic-memory-retrieval-a-multiverse-analysis/attachment/90406.docx?auth_token=vEG6f5pErEkQPzuS86eR)

---