Submitted: 01/15/20

Research Article

[1]Department of Psychology, McGill University, Montreal, QC, Canada

Keywords

Bat-and-ball problem, attribute substitution, dual process theory, replication study

Email Correspondence

alexandra.machalani@mail.mcgill.ca

Alexandra Machalani[1], Amanda Gallant[1], Victoria Orha[1], Nora Ostbo[1], Eric Hehman[1]

Substitution Sensitivity and the Bat-and-Ball Problem: A Direct Replication of De Neys *et al.* (2013)

Abstract

Background: Cognitive misers are no happy fools. Earlier findings (1) came to this conclusion by assessing people's sensitivity to attribute substitution, which they defined as the situation that occurs when we are confronted with a problem that demands greater cognitive effort, for which we rely on automatic and intuitive processes that substitute the complex situation for an easier one.

Methods: Through the exploration of the "bat-and-ball" problem, (2) De Neys, Rossi, and Houdé (1) found that participants were indeed sensitive to the substitution bias. Specifically, participants who incorrectly answered the question that gave rise to the substitution bias were significantly less confident in their answer relative to their answer on a control problem that did not give rise to the substitution. Using the same methods, we conducted a direct replication study on a sample of 264 undergraduate psychology students.

Results and Conclusion: Our results suggest that we successfully replicated the original conclusions; participants who answered by substituting the difficult question for an easier one significantly ($p<.0001$) decreased their confidence ratings on the version of the problem that gave rise to the substitution bias, relative to the problem that did not.

Limitations: Though there may have been limitations, it seems that we are sensitive to attribute substitution.

Introduction

As social creatures, humans have always been keen on finding the simplest, most quick and effective solutions. One could even say that we are cognitive misers; we not only engage in but rely on fast and intuitive processing rather than more effortful and deliberate thinking. First conceptualized by Peter Wason and Jonathan Evans, (3) this dual process theory proposes that we engage in two distinct types of processes: heuristic, in which an individual chooses which information is relevant and filters out the irrelevant information, and analytic, in which they analyze the information relevant for further processing. (4) This first theoretical proposition allowed for further research in the field of human cognition with the emergence of impactful theories such as Daniel Kahneman's (5) dual systems processes, which distinguishes processes as either intuition or reasoning based. A year later, Fritz Strack and Roland Deutsch (6) proposed their own dual process theory which puts forth the reflective and impulsive systems of the human mind. The model relies on separate systems they call the reflective and impulsive systems, where decisions are made using knowledge that comes from the situation or by using existing schemas without conscious thought. While faster cognitive processes might seem advantageous when we desire quick and effective problem-solving, they sometimes fall short when we are faced with complex situations that demand deeper reasoning. These processes thus lead us to biases in our judgment that we fail to recognize… or so it seems.

It has been theorized that the underlying problem of most cognitive biases and perceptual illusions is one of attribute substitution, a psychological process also known as the substitution bias. (1,2) Specifically, this concept is different from the decision-making strategy known as satisficing, as the latter entails "searching through the available options just long enough to find one that reaches a preset threshold of acceptability". (7p670) The substitution bias was first argued by Shane Frederick and Daniel Kahneman, (2) who based their theory on early research on the representativeness and availability heuristics. They relied on the hypothesis that "when confronted with a difficult question people often answer an easier one instead, usually without being aware of the substitution". (2p53) De Neys, Rossi, and Houdé (1) similarly define substitution bias, and revisit the 'bat-and-ball' problem first posed by Frederick and Kahneman (2):

> "A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?" (1p269)

The reasons for the high rate of errors in this easy problem lie in the fact that people are not accustomed to thinking hard, and instead trust judgment that comes to mind quickly, especially in situations in which they are not deeply committed. (2p58-59) Further explanations rely on the substitution of the critical relational statement 'more than' with an easier problem without the relational statement. Academics (2,8) who have explored the attribution bias concluded that the ultimate problem is that substitution is not detected. De Neys *et al.* however have challenged this assumption and believed that not deliberately reflecting upon one's responses does not necessarily imply that the substitution process is undetected. (1)

The original study

The authors hypothesized that people who engage in attribute substitution must be somewhat sensitive to their error process. (1) More specifically, people should have some unconscious awareness that the substituted '10-cent' answer they gave must not be completely accurate. To study this, De Neys *et al.* (1) relied on the assumption that when people engage in attribute substitution on the bat-and-ball problem, they will most commonly incorrectly respond in a specific, pre-determined way ('10 cents'). Because they were interested in one's sensitivity, they designed an isomorphic control problem (without the "more-than" relational statement) in order to obtain base rates of people's confidence levels in a similar problem that would not give rise to substitution:

> "A magazine and a banana together cost $2.90. The magazine costs $2. How much does the banana cost?" (1p260)

A total of 248 undergraduate students who took an introductory course

in psychology were recruited to take part in the study. All participants completed both the control and standard versions of the problem in randomly assigned orders where half would start with the control version. After each problem was solved, they were asked to rate how confident they were that their answer was correct on a scale from 0% to 100%. To support their hypothesis, the authors expected their data to show a significant decrease in confidence levels when biased reasoners rated the accuracy of their answer to the standard problem, relative to that of the control problem. However, if people are completely unsensitive to the substitution, as suggested by previous research, their confidence levels following the standard problem should not differ from that of the control problem.

De Neys *et al.* (1) found that 21% (SE=2.3%) of participants successfully solved the standard bat-and-ball problem (i.e. 5-cents answer) while 99.5% of those who answered incorrectly responded in line with the substitution bias (i.e. 10-cents answer). On the control version of the problem, 98% (SE=1%) of participants gave the correct answer. This indeed supports the theory that people do not just randomly guess when they are faced with complex questions, but rather solve the problem in an automatic and intuitive way. In line with their hypothesis, the data showed that those who engaged in substitution on the standard problem also rated their confidence levels as being significantly lower ($p<.0001$, $\eta 2p=.23$) than on the control problem. These findings thus support their hypothesis in that biased reasoners are sensitive to the substitution rather than being completely blind, relative to those who resisted the substitution and gave greater confidence ratings.

These results are interpreted by the authors as a clear example of how we tend to "minimize cognitive effort and stick to mere intuitive processing" and that "cognitive misers might have more accurate intuitions about the substitution process than hitherto believed". (1p271) Additionally, they argue that this process of attribute substitution might be an explanation for other cognitive biases, such as the base-rate neglect or the conjunction fallacy. (1)

The current study

The current era of the 'replication crisis' has led to a widely felt sense of uncertainty in the scientific community, especially in experimental psychology. In his manifesto, Chris Chambers (9) offers seven sins of psychological inquiry, the third of which is the sin of unreliability. He critiques the field's reluctance to replicate and the tendency to easily dismiss different outcomes as untrue replications when replications do occur. Therefore, as part of the Collaborative Replications and Education Project (CREP), the present paper will explore and discuss a direct replication of De Neys *et al.*'s substitution sensitivity study. (1) Along with the original hypothesis, an additional hypothesis was analyzed. Ethics approval for the study was received on October 22, 2019 by McGill University's Research Ethics Board.

Additional hypothesis

To add on to De Neys *et al.*'s original hypothesis, (1) we were interested in assessing participants' self-esteem. Specifically, we looked at optimism, a correlate of self-esteem, defined as "a belief that desirable outcomes are attainable". (10p517) In fact, Hale *et al.* explain that optimists have "a favorable outlook on life, expect things to go their way, and believe that good rather than bad things will happen to them". (10p271) Because it has been found to be related to psychological well-being, such as lowering levels of depression (11) and neuroticism, (12) we were intrigued by the idea of capturing a possible correlation between optimism and substitution sensitivity. More specifically, we wanted to investigate whether optimism can influence someone's perception of their own performance on tasks, and whether it has an important effect on their confidence. As such, we have hypothesized that participants who fall victim to the substitution bias – and therefore have lower confidence ratings in the standard condition of the De Neys (1) problem – will be more inclined to rate their expectancy for success lower, relative to those who are not affected by attribute substitution. Interestingly, if a correlation were to emerge and original findings were to be replicated, we might be able to conclude that these preliminary findings suggest that only a specific population is more sensitive to the substitution bias, such as those who are less optimistic, and more neurotic.

Methods

Participants

A total of two-hundred and sixty-four (N=264) undergraduate students from McGill University completed our study, or sixteen more than the original study. Participants were recruited through SONA, an online portal that manages voluntary study participation for the McGill Psychology Human Participation Pool in exchange for course credit(s). Inclusion criteria for our study were (i) participation for psychology course credit, (ii) undergraduate student status, and (iii) being 18 years or older. In line with the De Neys *et al.* (1) study, all participants were undergraduate students who have taken psychology courses, which is important to consider when inferring the generalizability of the present and other experiments investigating substitution sensitivity.

All participants read through the consent form explaining the purpose and contents of the study and indicated that they were informed and were voluntarily participating in our study. It was also made clear that participants could end the study at any time and would still be receiving compensation. would fit the background models, fit the source parameters and obtain their resulting "test statistics". The test statistic value is a way to quantify the quality of the maximum likelihood fit, and it roughly represents $\sigma 2$ significance for a normal ("Gaussian") distribution - so a larger value implies a higher likelihood of a gamma-ray signal. The formula for the test statistic is

Procedure

As this was a direct replication study, the detailed procedure that was made available in the original paper was closely followed. However, due to time constraints, an important difference should be noted: consistently across all testing sessions, participants had to complete our study following another replication study by Griskevicious, Tybur, and Bergh. (13) The experiment took the form of a computer-based survey through Qualtrics, completed in-person and in the presence of other participants. Sessions ranged from testing 5 participants to 20 participants at a time. Because the original study had not specified data collection methods, this was done to maximize sample size.

All testing sessions were completed in the same computer laboratory at McGill University's downtown campus and were in the presence of two experimenters (although a few sessions were conducted by one experimenter).

After students had completed the Griskevicious, Tybur, and Bergh (13) study and had read and agreed to our consent form, Qualtrics automatically randomly assigned them to one of four conditions. Following the original De Neys repeated-measures design, (1) all participants completed the control version of the bat-and-ball problem as well as a variation of the standard problem. Here, the superficial item content of the original bat-and-ball problem was modified to minimize surface similarity. The standard problem that was presented to participants was as follows:

> "A pencil and an eraser together cost $1.10.
> The pencil costs $1 more than the eraser.
> How much does the eraser cost?"

The four possible conditions included a control and a standard version using either the pencil/eraser combination or the magazine/banana combination. Consequently, about half of the participants completed the control version first, and the other half completed the standard question first. After answering each problem, participants were asked to type in how confident they were that their answer was correct on a scale of 0% (totally not sure) to 100% (totally sure). After the replication portion of the study was

completed, participants were then redirected to our additional measure, the GESS-R (see section 2.3), followed by basic demographic questions. Finally, participants were debriefed and were informed of the right answer to the standard problem.

Materials and measures

As this was a computer-based study, participants were not provided with any additional materials other than the computers available to them in the laboratory. A URL was presented to them on a whiteboard and students were asked to access the survey by themselves.

In addition to the original study's materials, the Revised Generalized Expectancy for Success Scale (GESS-R) (10) was used to measure optimism. Originally developed in 1978 by Fibel & Hale, (11) we used the 1992 revised scale published in the Journal of Clinical Psychology. The GESS-R is designed to measure optimism specifically in one's expectations for successful outcomes. Psychometric measures from the 1992 report suggests an acceptable level of reliability over time as well as a high level of internal consistency. Importantly, the GESS-R was not found to be correlated to neuroticism, giving way for a better interpretation of the results as attributable to optimism. Permission for the use and reproduction of the GESS-R was granted on October 27, 2019 by Dr. Daniel Hale.

To recall our main hypothesis, we expect biased reasoners on the standard problem to rate their confidence levels as significantly lower than their confidence levels on the control problem. In other words, the confidence rating in the correctness of one's answer to the standard problem was the main dependent variable that was being measured. Indirectly, confidence levels are meant to measure the latent construct of substitution sensitivity, where those who engage in substitution are shown to be sensitive by feeling less confident in their answer to a complex problem, relative to an easier problem. Additionally, we are also assessing optimism as a potential moderator of the relationship between confidence ratings and sensitivity to substitution. If data were to support our hypothesis that those who are less optimistic would perceive their answers as less accurate – thus being more sensitive to substitution – we would expect to see a greater decrease in confidence ratings in participants who score lower than normal on the GESS-R.

Data processing and analysis

In line with the original study and the CREP-provided step-by-step methods, separate analyses were conducted. To assess the possible differences in confidence levels, participants were grouped according to accuracy in their response on the standard problem. Specifically, accuracy was coded as "biased" (i.e. '10-cents' answer) and "correct" (i.e. '5-cents' answer). From this, the data was analyzed using a repeated-measures ANOVA, as this experiment was a two by two within-subjects design, in order to assess the differences in confidence levels of biased and unbiased reasoners on both control and standard problems.
For our additional hypothesis, a bivariate Pearson correlation analysis was performed in order to determine if there is a correlation between confidence levels in the standard problem and GESS-R score. Alpha values to determine statistical significance were set at 5% for both the original and additional hypotheses. Analyses were carried out by EH on IBM SPSS Statistics software version 26.0. (14)

Results

Participant characteristics

The majority of participants (84.47%) were cisgender females and 12.88% were cisgender males, where 'cisgender' was defined as identifying to one's biological sex assigned at birth. It is important to note that upon request, students were provided with a brief explanation of the term 'cisgender'. Among the sample, 50.8% were Caucasian, 20.5% were Asian, 10. 6% were of mixed ethnicities, and, the remaining 18.1% were Arab or Middle Eastern, Hispanic or Latin American, or Black or African American (Table 1).

Accuracy of responses

In agreement with the original study's results, 97.72% of participants answered the control problem correctly (Fig. 1). The other seven people who answered the control problem incorrectly were subsequently excluded from further analyses.

In the standard problem, results were not as dramatically distinct: 50.38% of participants answered correctly, while 43.94% answered incorrectly by engaging in a substitution (Fig. 2). The other fifteen people who answered incorrectly to the standard problem in a way that was not in line with the substitution bias were subsequently excluded from further analyses. Although the data suggests that it is not the majority who fall into the trap of attribute substitution, it does however show that the standard question is much trickier than the control. If the standard question would not give rise to some sort of confusion, results would be much more heterogeneous in which there would be an unequal split between correct and incorrect answers among participants.

	Number of participants *(N)*	Percent *(%)*	Mean *(M)*
Age			20.42
Ethnicity			
Asian	20	7.6	
Arab or Middle-Eastern	54	20.5	
Black or African American	8	3.0	
Caucasian or White	134	50.8	
Hispanic or Latin American	12	4.5	
Mixed	28	10.6	
Other	8	3.0	
Gender			
Cisgender male	34	12.88	
Academic Year			
U0	10	3.79	
U1	83	31.44	
U2	80	30.30	
U3	75	28.41	
U4	16	6.06	
First Language			
Arabic	12	4.55	
English	122	46.21	
French	50	18.94	
Mandarin	25	9.47	
Spanish	11	4.17	

Table 1. Demographics of participants.

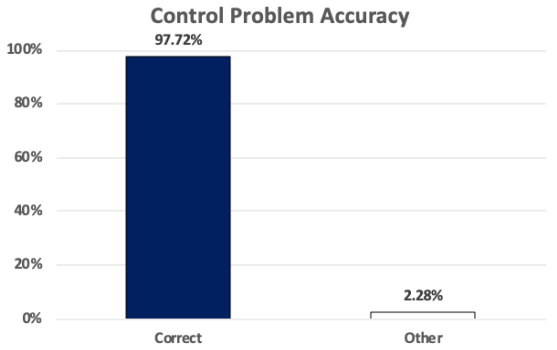

Figure 1. Accuracy of participants' answers on the control problem.

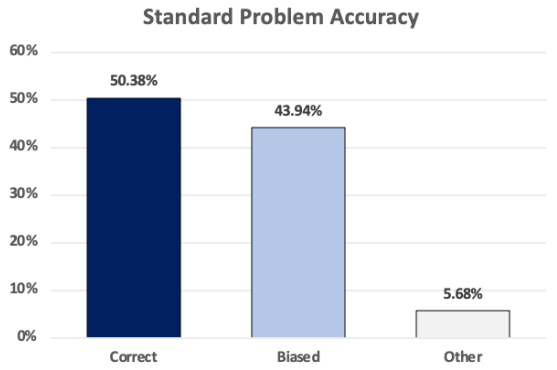

Figure 2. Accuracy of participants' answers on the standard problem.

Confidence ratings

Consistent with the central hypothesis, results show that biased reasoners (participants who engaged in the substitution bias) rated their confidence levels as significantly lower on the standard problem compared to the control problem, $F(1, 114)=31.62$, $p<0.0001$. In participants who correctly answered the standard problem, there was a negligible confidence discrepancy between the two conditions (Fig. 3).

Analyses of optimism

Following a bivariate correlation analysis between GESS-R score and confidence levels, it appeared that there was no significant correlation between variables ($r=0.006$, $p=0.929$). Hence, optimism as we have defined it has not been found to be correlated with confidence levels on either the standard or control problem.

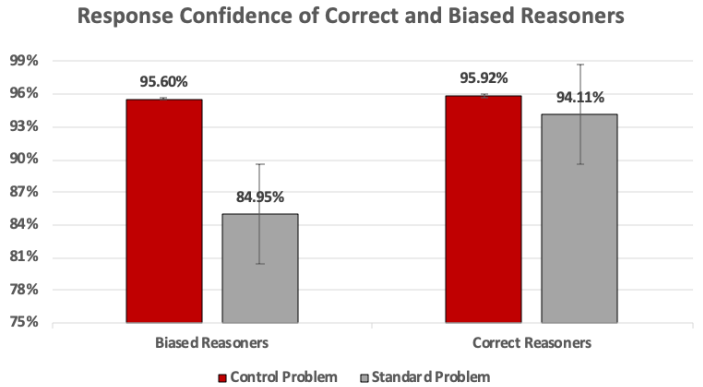

Figure 3. Response confidence on the standard and control versions of the problem for participants who answered the standard problem incorrectly (biased) and correctly (correct). Error bars are standard errors.

Discussion

The present data suggests that we have successfully replicated the results of De Neys *et al.*'s original paper. (1) First, we have replicated the way in which we obtained significant results in the same direction by using the same confidence levels as the original study ($\alpha=95\%$). Second, we have replicated the observed significant decrease in biased reasoners' confidence ratings on the standard problem relative to the control. In line with this, we have also replicated the narrow gap in confidence ratings of correct reasoners in both conditions. Finally, we have replicated the magnitude of answers in the control problem, where the majority of participants answered correctly.

What we have not replicated, however, is the percentage of people who incorrectly answered the standard problem. De Neys *et al.* (1) found that only 21% got the standard question correct, while we found that 50.38% got the standard question correct. Possible confounds to this could be explored, such as the possibility that students who took part in the study were from different cultural backgrounds, as well as in different academic levels (Table 1). It may be that some participants had already been exposed to this type of question through preparing for the Graduate Record Examination (GRE), for instance.

It is therefore safe to conclude that we have successfully replicated De Neys *et al.*'s original Bats, balls, and substitution sensitivity study, (1) but with certain conditions. Our results support the theory that we are not completely blind to questionable cognitive processes, and that we might actually be sensitive to the fact that we are giving unreliable answers to complex questions. What this could suggest is that when we are faced with difficult situations, we tend to prefer less effortful ways of reasoning. In other words, we prefer minimizing our cognitive effort and resort to intuitive processing instead.

The results for our additional hypothesis demonstrated that optimism did not have a significant effect on the confidence ratings. The negative findings regarding the construct imply that it does not have a significant impact on the way people perceive their confidence regarding their performance. In the current literature, however, optimism has been known to impact future performance positively. (15) It is possible that our results may therefore indicate that when individuals assess their performance immediately after completing a task, they tend to be more realistic than optimistic. Future research should further investigate the relationship between optimism and both immediate and future performance on tasks.

Limitations

Although most of our data supports our hypothesis, there are a few limitations to consider. First, the way confidence was measured might not be entirely appropriate. Because confidence is a latent construct, we cannot be sure that it is confidence levels in participants' answers that we are measuring. Uncertainty in a measure's construct validity poses a threat to the interpretation of our results, especially since this is a self-reported measure. Further, confidence levels in one's answer are meant to measure substitution sensitivity. It may be that being more or less confident in one's answer does not represent sensitivity to attribute substitution, but rather confidence in one's answer due to mathematical skills, complex wording of the problem, or any other confound not measured. Specifically, it is interesting to think about how the wording of the question might impact a participant's confidence ratings, since the only difference between the standard and the control problem is the use of the relational statement "more than". The former is used to elicit a specific incorrect response from participants, which in turn should decrease their confidence in their answer. A second limitation is that external validity in this study is not substantial. Because this sample was a convenience sample, it may be that these results are only of interest for young psychology undergraduates. In addition, the time at which the study was offered poses a threat to external validity. In fact, the Fall semester is often very stressful when undergraduates can be overwhelmed with work, study, and applications to graduate school.

As for our additional hypothesis using the GESS-R to assess optimism, we

expected a pattern of results showing that higher scores on the GESS-R would have influenced confidence ratings in both conditions in a positive direction, as well as a high confidence levels in biased reasoners. Although we could easily conclude that optimism is unrelated to confidence levels and substitution sensitivity, it is important to note that we had no base rates to compare participants' scores to. It would have been more powerful to assess their optimism score before exposing them to potentially threatening situations, which might decrease confidence in one's abilities for a short period of time, causing optimism levels to also decrease.

Implications and future directions

Future research wanting to explore possible covariates affecting confidence levels and substitution bias may want to include certain personality characteristics such as impulsivity or neuroticism. It may be that those who are more impulsive or neurotic tend to rate themselves as more or less confident respectively regardless of the type of condition they are in. Other possible variables that could be tested is simple test anxiety, where examiners can send a scale assessing people's anxiety when taking tests, or specifically arithmetic examinations to see if these people are more likely to get answers wrong or rate themselves as less confident.

The implications of this type of research for the field of human cognition is important to consider. Recent work has shown that biased reasoners on the bat-and-ball problem do not lazily monitor their intuitive reasoning. (16) Rather, they do evaluate quite automatically their substituted answers by answering the mathematical and the confidence questions much more slowly than on the control question. The authors suggest that it all boils down to the relational term "more than", which causes conflict between our automatic, linguistic operations and our quick, intuitive reasoning processes, thus providing this semantic awareness that we have not fully complied with the relational terms of the sentence. (16) Thus, future research in this interdisciplinary area of cognitive awareness and decision-making will give deeper insight into the ways we are not entirely "happy fools".

Acknowledgements

Dr. Eric Hehman carried out statistical analyses, along with Psychology graduate student Eugene Ofosu. We would like to thank the Collaborative Replications and Education Project and the McGill Research Ethics Board for approving our replication project. We would also like to thank anonymous reviewers for their helpful and constructive feedback.

References

1. De Neys W, Rossi S, Houdé O. Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. Psychonomic Bulletin & Review. 2013;20(2):269–73.

2. Kahneman D, Frederick S. Representativeness Revisited: Attribute Substitution in Intuitive Judgment. Heuristics and Biases. 2002, Aug;49–81.

3. Wason P, Evans J. Dual processes in reasoning? Cognition. 1974;3(2):141–54.

4. Evans JSBT. Heuristic and analytic processes in reasoning*. British Journal of Psychology. 1984;75(4):451–68.

5. Kahneman D. A perspective on judgment and choice: Mapping bounded rationality. American Psychologist. 2003Sep;58(9):697–720.

6. Strack F, Deutsch R. Reflective and Impulsive Determinants of Social Behavior. Personality and Social Psychology Review. 2004Aug1;8(3):220–47.

7. Colman AM. A dictionary of psychology [Internet]. 2nd ed. Oxford: Oxford University Press; 2006 [cited 2020Mar2]. Available from: archive.org

8. Tversky A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases. Science. 1974;185(4157):1124–31.

9. Chambers C. The seven deadly sins of psychology: a manifesto for reforming the culture of scientific practice. Princeton, NJ: Princeton University Press; 2017.

10. Hale WD, Fiedler LR, Cochran CD. The revised Generalized Expectancy for Success Scale: A validity and reliability study. Journal of Clinical Psychology. 1992;48(4):517–21.

11. Fibel B, Hale WD. The Generalized Expectancy for Success Scale: A new measure. Journal of Consulting and Clinical Psychology. 1978;46(5):924–31.

12. Smith TW, Pope MK, Rhodewalt F, Poulton JL. Optimism, neuroticism, coping, and symptom reports: An alternative interpretation of the Life Orientation Test. Journal of Personality and Social Psychology. 1989;56(4):640–8.

13. Griskevicious V, Tybur JM, Bergh BVD. Going green to be seen: Status, reputation, and conspicuous conservation. Journal of Personality and Social Psychology. 2010;98(3):392–404.

14. IBM. SPSS statistics for Windows. Version 26.0. Armonk, NY: IBM; 2019.

15. Loftus TJ, Filiberto AC, Rosenthal MD, Arnaoutakis GJ, Sarosi GA, Dimick JB, et al. Performance advantages for grit and optimism. The American Journal of Surgery. 2020.

16. Johnson ED, Tubau E, Neys WD. The Doubting System 1: Evidence for automatic substitution sensitivity. Acta Psychologica. 2016Dec23;164:56–64.