Replication of "The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information" by E Hopkins, D Weisberg, J Taylor (2016, *Cognition*)

Skyé Lowe
[redacted]
Haruka [redacted]
[redacted]
Diana [redacted]
[redacted]
Sophia [redacted]
[redacted]
Uma [redacted]
[redacted]

Introduction

The original study by Hopkins, Weisberg, and Taylor (2016) investigated whether the 'seductive allure effect' extends beyond psychology to other scientific disciplines. The seductive allure effect refers to the tendency to judge explanations of psychological phenomena as "more satisfying" when they contain irrelevant neuroscientific information. Previous work by Weisberg et al. (2008) demonstrated this effect initially through the specific examination of how irrelevant neuroscientific information and images increase lay quality evaluations of explanations for psychological phenomena.

Hopkins and colleagues hypothesized that this effect is representative of a general preference for reductive explanations across sciences, not just psychology. Reductive explanations are those that describe more fundamental processes or lower-level scientific domains to better explain a phenomenon. They predicted that participants would give reductive explanations higher ratings compared to other horizontal explanations in the reductive scientific hierarchy (in descending order: physics, chemistry, biology, neuroscience, psychology, social science).

The original study found a significant main effect of explanation level ($\beta=0.21$, $t=2.23$, $p < .05$), such that reductive explanations ($M=1.26, SD=1.71$) were rated significantly higher than

horizontal explanations (M=1.04,SD=1.88). These results supported their initial hypothesis: Participants generally preferred explanations that contained reductive information, even when that information was logically irrelevant.

This replication attempts to replicate the main Explanation Level effect through a reproduction of their primary explanation-rating task containing the same information. We used one set of participants recruited via Prolific. We hypothesized that explanations which contain logically irrelevant reductive information will again be judged as more satisfying compared to matched horizontal explanations.

Methods

Participants

We had a planned sample size of 50 (23 male, 26 female, 1 prefer not to say; mean age = 44.8 years, *SD* = 11.14). The participants were recruited from Prolific and were limited to those living in the United States. The only exclusion criteria applied was for those who failed the attention check (n = 0). No participants were removed from the study, for a final sample size of 50. Participants received USD$3.00 for completing the study, calculated by an estimated completion time of 15–20 minutes and an hourly wage of $10.

Materials

The materials were identical to those used in the original study. As described by Hopkins et al. (2016):

"The Rating Explanations task used 24 different phenomena (four per science). All phenomena and their corresponding explanations are available in the online supplemental materials and via the Open Science Framework [...] The phenomena described concepts, principles, or research findings from each of the six sciences" (Hopkins et al., 2016).

For each phenomenon within its respective science, there were four corresponding explanations constructed: reductive-good, reductive-bad, horizontal-good, and horizontal-bad. For the horizontal condition, explanations used only terminology and concepts from the same discipline as the phenomenon. In the reductive condition, explanations included additional logically irrelevant terminology from the immediately lower science in the hierarchy, while maintaining the same structure as horizontal explanations.

In both conditions, Good explanations consisted of clear, accurate causal accounts of the phenomenon as verified by experts. Bad explanations were constructed by removing key explanatory information and replacing it with circular restatements or irrelevant details. As

reiterated by Hopkins et al. (2016): “The added reductive text was identical for good and bad versions of the explanation”

The experiment was administered via a Qualtrics survey. Participants, after being shown an explanation, were prompted to rate explanations on a multi-point scale from -3 (very poor) to +3 (very good).

Procedure

Hopkins et al. (2016) reported the following procedure:

> “All participants completed the Rating Explanations task first. For this task, participants used a sliding scale ranging from 3 to 3 to indicate their responses. They were first given instructions on how to use the slider; this also served as a check that participants were reading instructions. They were told to use the slider to select 0 on the first page in order to proceed with the survey. If they selected anything other than 0, they were directed to another page asking them again to select 0. Participants who did not select the correct response on this second page (three MTurk workers and nine undergraduates) were excluded from analyses.
>
> After these general instructions on using the slider, participants were given instructions for the explanations task (modified from Fernandez-Duque et al., 2015):
>
> > You will now be presented with descriptions of various scientific findings. All the findings come from solid, replicable research; they are the kind of material you would encounter in a textbook. You will also read an explanation of each finding. Unlike the findings themselves, the explanations of the findings range in quality. Some explanations are better than others: They are more logically sound.
> >
> > Your job is to judge the quality of such explanations, which could range from very poor (3) to very good (+3).
>
> On each trial, participants were presented with a description of a scientific phenomenon, which was displayed for 10 s before participants could advance to the next screen. On the next screen, an explanation was displayed below the phenomenon, and participants were instructed to rate the quality of the explanation. Participants rated twelve explanations, with an attention check trial administered after the first six (Oppenheimer, Meyvis, & Davidenko, 2009).”

The procedure for the primary confirmatory analysis (as described) was followed precisely except for the following: Participants in the original study answered additional questions for secondary analysis that we did not replicate. The original study also used a sliding scale as the response tool; our study used a multiple choice for better interface visualization on Qualtrics.

Finally, we included both the original phenomenon description and the explanation on the rating page, rather than only displaying the description before the explanation.

In general, a session started with instructions on how to use the horizontal multiple choice selection, which served as the first attention check, with participants being instructed to select "0." On each trial, participants were presented with a description of a scientific phenomenon which was displayed for 10 s before auto-advancing to the next screen. The next screen displayed both the original phenomenon description *and* the explanation and asked the participant to rate the quality of the explanation. Participants rated twelve explanations with an attention check administered after the first six.

Analysis Plan

To examine our main hypothesis, we conducted a **linear mixed-effects regression model** to predict participants' explanation ratings on each trial (ranging from –3 to +3). The model included three fixed effects:

- ***Explanation Level*** *(between-subjects: horizontal vs. reductive)*
- ***Quality*** *(within-subjects: good vs. bad)*
- ***Science Discipline*** *(within-subjects: physics, chemistry, biology, neuroscience, psychology, social science)*

We tested all possible interactions between these variables. We used likelihood ratio tests to evaluate whether higher-order interactions significantly improved model fit. To determine whether the reductive effect was dependent on domain or explanation quality, we examined the following two-way interactions:

- ***Explanation Level*** **×** ***Science Discipline***
- ***Explanation Level*** **×** ***Quality***

We focused primarily on the ***Explanation Level*** **×** ***Science Discipline*** interaction. After verifying that the two-way interaction proved significant, we conducted a simple-effects analysis to determine where the reductive allure effect occurred.

The dependent variable is the participant's rating of the explanation, on a scale from -3 to 3. The independent variables are the Explanation Level, Quality, and Science Discipline.

Simple-Effects Analyses:

The simple-effects analysis was conducted by fitting separate mixed-effects models within each scientific discipline. Each model predicted explanation ratings from the ***Explanation Level*** and included random intercepts for participants and items.

Repeated Measurements & Random Effects:

To account for repeated measurements across participants and items, we included random intercepts for both participants and items. We also included a random slope for the *Quality* variable by item, following the best-fitting random-effects structure reported in the original study. The Science Discipline variable was deviation-coded, with physics serving as the reference level.

Data Exclusion and Violations of Statistical Assumptions

Participants failing either attention check were excluded, though none failed in our dataset. To address violations of statistical assumptions, we reported bootstrapped confidence intervals for all fixed-effect coefficients at 90%, 95%, and 99%, as in the original study. The mixed-effects framework and bootstrapping provided robustness without requiring data transformation.

Differences from the original study

Participant Population

The primary difference in participant population concerns sample size: our replication included 50 participants, whereas the original study analyzed 259 participants after exclusions. Although both studies recruited adult online participants, there were differences in platform (e.g., Prolific versus MTurk/undergraduates). The markedly smaller sample size in the replication reduces statistical power, increases uncertainty in parameter estimates, and increases the likelihood of singular random-effects structures. Additionally, our replication observed no attention-check failures, whereas the original excluded nearly 19% of participants. This discrepancy may reflect participant differences, platform effects, or differences in how the attention checks were implemented or interpreted. Such differences could meaningfully influence data quality, though in this case, no evidence of inattention was detected.

Equipment or Software

Both studies relied on online data collection, but our replication used a modern version of Qualtrics and R packages that have evolved since the original analysis. Differences in software versions, browser behavior, or slider-response exports could influence how attention-check items were recorded or how participants interacted with stimuli. However, these differences are not expected to alter the psychological processes under study meaningfully. The presence of warnings regarding model convergence and boundary fits in the replication appears attributable to a small sample size rather than software malfunction.

Stimuli or Materials

Our replication recreated the original stimuli via an item mapping system, assigning each explanation to its scientific discipline and quality manipulation based on item-set counterbalancing. The exact text presented to participants (including the placement of reductive sentences, length matching, and contextual details) was directly verified against the original supplemental materials. Small discrepancies in formatting could influence effect sizes, though

there is no direct evidence of such differences. If any mismatch exists, it's unlikely that it could weaken the observed effects, given that the changes in formatting don't impact manipulations of explanatory content. Additionally, in dividing the stimuli into block A and block B, the original study did not detail their process of selecting the predetermined set, so our replication randomly assigned stimuli into these blocks.

Procedure

Procedurally, our replication followed the structure of the original experiment, including the 2 × 2 × 6 factorial design, randomization across item sets, and rating each explanation on the same –3 to +3 scale. The inconsistency in exclusion rates between studies suggests that the replication's attention-check procedure may have functioned differently, though this is not expected to bias results unless inattentive participants were included—which the replication found no evidence for. No procedural failures or interruptions occurred, and no exploratory analyses beyond model diagnostics (e.g., bootstrapped confidence intervals, inspection of convergence warnings) were conducted.

Results

Data Preparation

Cleaning dataframe

We removed the Qualtrics metadata (recipient information and geographic data) and system rows. Anonymous participant IDs were created in the format "sub-##" to protect privacy. Age was converted to numeric format for analysis.

Final sample cleaning: removing attention check failures

There were a set of two attention checks occurring throughout the experiment. One before the experiment begins, and one after six trials have taken place.

> First attention check: participants were instructed to click the number "0" from the following options: -3, -2, -1, 0, +1, +2, +3.
>
> Second attention check: participants were instructed to click the number "+3" from the following options: -3, -2, -1, 0, +1, +2, +3

Participants who failed either attention checks (selecting any number other than the one given in the instruction) were excluded from the final sample; participants who failed either attention checks were to be manually removed from our final sample.

None of the participants were dropped because every participant passed both attention checks.

Confirmatory Analysis

Data from the rating tasks were analyzed with a mixed-effects linear regression model (using the lme4 package in R) predicting the rating given on each trial from the sample, the quality of the explanation (good, bad), the explanation level (horizontal, reductive), and the science from which the phenomenon was drawn (physics, chemistry, biology, neuroscience, psychology, and social science). Explanation levels were between-participant variables and quality and science were within-participants. All possible interactions between variables were tested. The three-way interaction and most of the two-way interaction did not significantly improve model fit (see below), as assessed by likelihood ratio testing, and these were dropped.

Three-way interaction (Explanation Level × Quality × Science):
$\chi^2(5) = 6.60$, $p = .252$ (not significant)
Two-way interaction (Explanation Level × Science):
$\chi^2(10) = 9.92$, $p = .447$ (not significant)
Two-way interaction (Explanation Level × Quality):
$\chi^2(6) = 6.72$, $p = .348$ (not significant)

We did not find a significant main effect of explanation level with $t(48.0) = -1.21$, $p = .232$: Reductive explanations were rated higher than horizontal explanations, but this was not statistically significant. There were no significant Explanation Level x Science interactions, meaning the magnitude of the effect did not significantly differ across the sciences. However, we did report that the effect of the explanation level was non-significantly larger for psychology compared to other sciences ($t = 0.25$, $p = .062$).

We did not find a significant two-way interaction between explanation level x science ($p = .447$), nor between explanation level x quality ($p = .348$). We also did not find a significant three-way interaction between explanation level x science x quality ($p = .252$).

Table 1. Comparison of Statistical Results: Hopkins et al. (2016) vs. Current Replication

Effect	Hopkins et al. (2016)	Current Replication
Sample Size (Before Exclusions)	N = 319	N = 50
Sample Size (After Exclusions)	N = 259	N = 50
Main Effect: Explanation Level	t = 2.23, p = .026* d = 0.18, n^2p = .004	t(48.0) = -1.21, p = .232 (ns) d = -0.35, n^2p = .030
Main Effect: Quality	t = 12.20, p < ".001" d = 0.96, n^2p = .361	t(14.4) = 0.06, p = .953 (ns) d = 0.03, n^2p = .000
Two-way: Explanation Level x	Significant (p = .001)	x^2(5) = 9.92, p = .447 (ns)

Science		
Strongest contrast (Social Science)	t = -3.19, p < ".001"**	See simple effects analysis
Two-way: Explanation Level x Quality	t = -1.33, p = .18 (ns)	x^2(1) = 6.72, p = .348 (ns)
Three-way: Explanation Level x Quality x Science	Not significant; dropped from model	x^2(5) = 6.60, p = .252 (ns)

Table 2. Original Study vs. Replication Study: Mixed-effects regression predicting ratings of explanations.

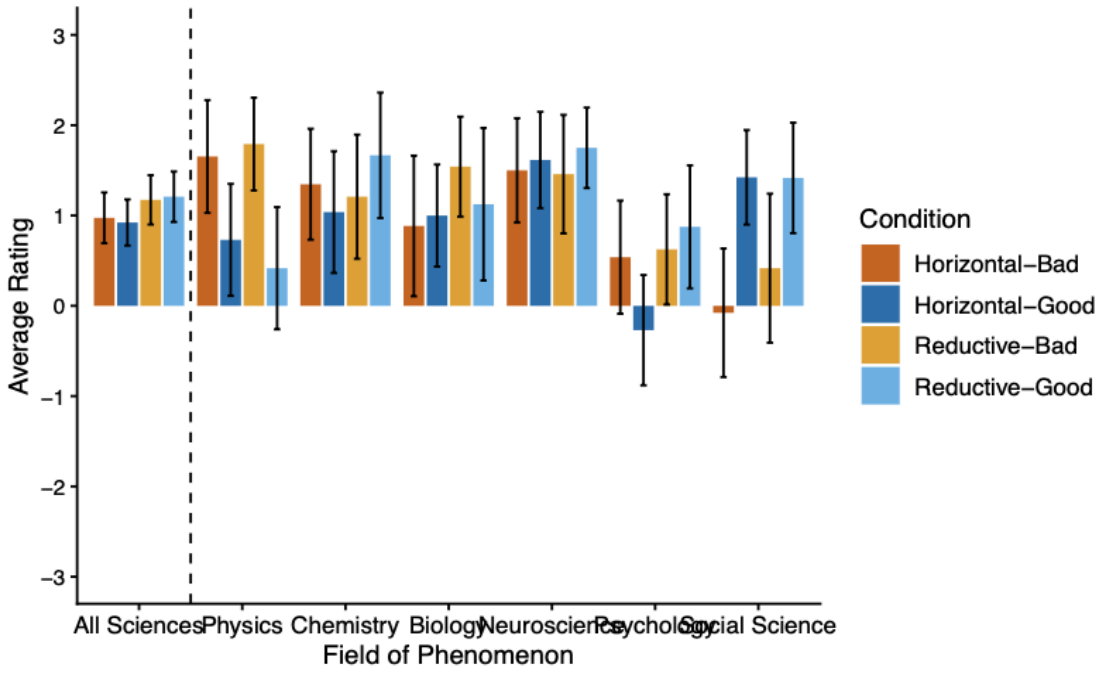
* (yellow = original study; blue = replication study; gray = not collected in replication)

Predictor	Estimate		[95% CI]		*t* value	
Intercept	1.13	1.070092	[0.96, 1.29]	[0.85, 1.27]	13.32	9.938
Sample	-0.25	-	[-0.44, -0.07]	-	-2.66	-
Quality	1.27	0.004387	[1.07, 1.47]	[-0.14, 0.16]	12.20	0.059
Explanation Level	0.21	-0.121374	[0.02, 0.42]	[-0.32, 0.08]	2.23	-1.210
Science						
Chemistry	0.08	0.246075	[-0.24, 0.42]	[-0.07, 0.57]	0.49	1.489
Biology	0.05	0.066559	[-0.27, 0.42]	[-0.23, 0.40]	0.33	0.403
Neuroscience	0.28	0.509638	[-0.07, 0.63]	[0.19, 0.84]	1.67	3.083
Psychology	-0.27	-0.625458	[-0.60, 0.08]	[-0.97, -0.32]	-1.59	-3.784
Social Science	-0.14	-0.278989	[-0.45, 0.15]	[-0.60, 0.01]	-0.86	-1.688
Sample x Quality	0.54	-	[0.30, 0.76]	-	4.92	-
Sample x Explanation Level	-0.13	-	[-0.52, 0.27]	-	-0.67	-
Sample x Science						
Chemistry	0.05	-	[-0.17, 0.29]	-	0.43	-
Biology	-0.14	-	[-0.37, 0.11]	-	-1.13	-
Neuroscience	0.20	-	[-0.05, 0.42]	-	1.66	-

Psychology	0.12	-	[-0.10, 0.35]	-	0.98	-
Social Science	-0.12	-	[-0.36, 0.13]	-	-1.00	-
Quality x Explanation Level	-0.14	-0.19	[-0.37, 0.10]	[-0.12, 0.14]	-1.33	0.399
Quality x Science						
Chemistry	-0.19	-0.040784	[-0.62, 0.20]	[-0.36, 0.28]	-0.83	-0.247
Biology	0.00	0.069764	[-0.49, 0.43]	[-0.23, 0.40]	0.00	0.422
Neuroscience	-0.30	-0.107349	[-0.75, 0.14]	[-0.42, 0.24]	-1.31	-0.649
Psychology	0.44	0.137363	[0.00, 0.87]	[-0.20, 0.45]	1.90	0.831
Social Science	-0.02	-0.633155	[-0.42, 0.38]	[-0.96, -0.32]	-0.09	-3.830
Explanation Level x Science						
Chemistry	0.04	-0.002485	[-0.19, 0.27]	[-0.28, 0.27]	0.32	-0.018
Biology	0.03	-0.072970	[-0.20, 0.27]	[-0.34, 0.21]	0.21	-0.520
Neuroscience	0.08	0.099336	[-0.16, 0.33]	[-0.19, 0.39]	0.69	0.708
Psychology	0.22	-0.188645	[-0.02, 0.44]	[-0.49, 0.10]	1.79	-1.344
Social Science	-0.39	0.003347	[-0.65, -0.15]	[-0.25, 0.33]	-3.19	0.024
Sample x Quality x Explanation Level	0.49	-	[0.01, 0.94]	-	2.23	-

Figure 1. Mean Explanation Ratings by Scientific Discipline

Error bars represent 95% confidence intervals



Discussion

Summary of the Replication Attempt

We **failed** to replicate the main findings of Hopkins et al. (2016). We did not find a significant main effect of explanation level (β = -0.12, 95% CI[-0.32, 0.08]). In comparison, the original study found a significant main effect of explanation level (β = 0.21, 95% CI[0.02, 0.42]). Altogether, the estimated direction of the effect was reversed in our replication (positive in the original, negative in the replication), although the magnitudes were of somewhat similar absolute values. We also failed to replicate any of the interactions found in the original study. Our replication study's regression analysis (Table 2) did not reveal significant interaction between any of these three interactions: (1) Explanation Level x Quality x Science ($\chi^2(5) = 6.60$, $p = .252$), (2) Explanation Level x Science ($\chi^2(10) = 9.92$, $p = .447$), (3) Explanation Level x Quality ($\chi^2(6) = 6.72$, $p = .348$).

Commentary

Our failure to replicate the results from Hopkins et al. (2016) has a host of potential explanations. The first considers the possibility of fundamentally theoretical failure; namely, the "seductive allure" of reductive information attempts to describe a psychological phenomenon does not actually exist. Important to consider is the precedent study to Hopkins et al. (2016) that focused on the seductive allure of irrelevant neuroscientific fMRI images in psychological explanations (Weisberg et al., 2008). Numerous critiques and failed replications of this study have followed, including arguments for the true epistemic value of brain images rather than a seductive allure (Farah & Hook, 2013; Michael et al., 2013; Gruber & Dickerson, 2012). However, Hopkins et al. (2016) and our replication — although based on this foundation of the idea of a seductive allure of fMRI images — critically did *not* use images as a part of our given stimuli. Nonetheless, Wilson et al. (2025) also fails to replicate Hopkins et al. (2016), meaning even with solely written explanations, bringing the seductive allure of reductive information into question.

Another possible factor contributing to our failed replication considers characteristics of the sample. Most obviously, our replication deviated by only recruiting participants through Prolific, while the original study had two sample sets (participants recruited through MTurk and undergraduate students). First, Prolific and MTurk have been found to have differential quality of participants (Douglas et al., 2023). MTurk is an online marketplace for businesses and individuals to outsource whatever tasks they need completed, from transcribing short videos to data entry to testing new tools, and participating in academic studies. Since the MTurk platform hosts a wide variety of tasks, it is possible that the individuals participating in the experiment weren't approaching the task with the same educational background, expectations, and mindset as the Prolific participants. Second, we did not recruit a secondary sample of undergraduate students; however, the original study did not find significant differences in their two samples (MTurk and undergraduates). More speculatively, our replication comes almost a decade after the original study, and meaningful characteristics of our sample in 2025 may be significantly

different than that of 2016. Since this time, the presence of medical and scientific terminology has been more prominent in the media with COVID-19. When people are more exposed to any stimulus more often, like scientific jargon, it becomes less novel. Perhaps the participants have become more desensitized to language that would have previously impressed them.

The implementation of our replication also differed from the original study in a few other ways. We would argue that although our participants were not exposed to tasks including perceptions of science ratings and other cognitive and science-understanding tests, this would not impact our results as these tasks were presented to participants after the main rating components of the survey in the original study. Perhaps a more meaningful difference was that our study presented the original phenomenon description alongside the evaluated explanation, while the original study had them presented consecutively but separately. It is possible that when participants were able to refer to the exact description and question, they were able to recognize the irrelevance of the reductive information.

There are a few theoretical implications of our failure to replicate. Taken most extremely, our results may undermine the fundamental existence of the 'seductive allure of reductive information' phenomenon. More likely and more moderately, however, our failure to replicate indicates that this phenomenon has particular qualifying contexts, including sample characteristics and the set-up of the description and explanation. In other words, irrelevant reductive information may become more appealing to lay people given their personal exposure to scientific communication or when they are unable to refer to the exact explanandum of the explanation they are evaluating. Future research could investigate more precisely the specific qualifying or moderating factors that induce the seductive allure, including demographics, presentation of the stimuli, perceived expertise of the explanation source, or cognitive load on the participants. Taken altogether, while we failed to replicate the key results of Hopkins et al. (2016), the differences in experiment implementation and speculation of changes in population-level science understanding opens new questions to what exactly leads to lay perceptions of explanations and their quality.

References

Douglas, B.D, Ewell, P. J., Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE*, 18(3): e0279720. https://doi.org/10.1371/journal.pone.0279720

Farah, M. J., & Hook, C. J. (2013). The Seductive Allure of "Seductive Allure". *Perspectives on psychological science : a journal of the Association for Psychological Science*, *8*(1), 88–90. https://doi.org/10.1177/1745691612469035

Fernandez-Duque, D., Evans, J., Christian, C., & Hodges, S. D. (2015). Superfluous neuroscience information makes explanations of psychological phenomena more appealing. *Journal of Cognitive Neuroscience*, 27(5), 926–944. http://dx.doi.org/10.1162/jocn_a_00750.

Gruber, D., & Dickerson, J. A. (2012). Persuasive images in popular science: Testing judgments of scientific reasoning and credibility. *Public understanding of science (Bristol, England)*, *21*(8), 938–948. https://doi.org/10.1177/0963662512454072

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. V. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, *155*, 67–76. https://doi.org/10.1016/j.cognition.2016.06.011

Michael, R. B., Newman, E. J., Vuorre, M., Cumming, G., & Garry, M. (2013). On the (non) persuasive power of a brain image. *Psychonomic bulletin & review*, *20*(4), 720–725. https://doi.org/10.3758/s13423-013-0391-6

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872. http://dx.doi.org/10.1016/j.jesp.2009.03.009

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of cognitive neuroscience*, *20*(3), 470–477. https://doi.org/10.1162/jocn.2008.20040

Wilson, K. D., Lonergan, M., Nagel, C., & Meier, B. P. (2025). Does reductive information increase satisfaction with scientific explanations? Three preregistered tests of the reductive allure effect. *Cognition*, *254*, 105941. https://doi.org/10.1016/j.cognition.2024.105941