Replicating the “Seductive Allure of Neuroscience Explanations” effect in a classroom experiment and an online study

Pearl Amber Väth, Jakob von Petersdorff, Christof Neumann, Roger Mundry & Julia Fischer

Corresponding Author: Julia Fischer jfischer@dpz.eu

Short title: Replicating the SANE Effect

Keywords: classroom experiment, neuroscience, replication, seductive allure, science communication

Abstract

The “Seductive Allure of Neuroscience Explanations” (SANE) effect refers to the observation that superfluous neuroscience information (SNI) added to an explanation can bias judgments of information quality. Here, we report the results of a classroom experiment to sensitize undergraduate students to that issue. In contrast to previous studies, students rated good explanations without SNI as the highest. There was also considerable variation between student cohorts. Inspired by these observations, we set out to conceptually replicate the original study using an online experiment that allowed us to directly assess the statistical interactions between explanation quality, the presence of SNI, and expertise levels. In this preregistered study, participants (n=409) with varying levels of expertise rated the quality of good and bad explanations, with or without SNI. Irrespective of the presence of SNI, participants across all expertise levels rated good explanations more favourably than bad ones, but the differences were surprisingly small. We also found a significant interaction between the impact of SNI and expertise, with SNI boosting ratings mostly in participants with less expertise ($p=<0.001$), corroborating previous findings. Developing a curriculum that trains students to distinguish between actual explanations and “bullshit” would ultimately also sensitize teachers and experts that produce and review scientific information.

1. Introduction

Neuroscientific information carries a unique persuasive appeal. Even when not directly relevant, it can present an argument with an air of authority and scientific credibility [1–3]. As neuroscience continues to gain prominence in public discourse, understanding its influence on beliefs, attitudes, and behavior becomes increasingly important [4]. The “Seductive Allure of Neuroscience Explanations” (SANE) effect highlights how neuroscience explanations and brain imagery can influence information quality and persuasiveness judgments. In a seminal study on the SANE effect, Weisberg and colleagues explored the impact of adding irrelevant neuroscience information to explanations in three experiments [3]. The authors first presented participants with different psychological phenomena. They then offered different explanations and asked participants to rate their satisfaction. The explanations varied with regard to quality and the presence of superfluous neuroscience information (SNI) that did not add to the understanding of the phenomenon. Good explanations provided a reason (mechanism) for a given psychological phenomenon, while bad explanations merely repeated the results in other words. These good and bad explanations were presented with or without superfluous neuroscience information in a 2×2 design.

The three experiments by Weisberg and colleagues involved different groups of participants: novices (n=81), students in a cognitive neuroscience class (n=22), and neuroscience experts (n=48), and both between-subject and within-subject designs were used [3]. Across all participant groups, good explanations received higher ratings than bad ones, and superfluous neuroscience information boosted the rating of bad explanations in novices and students but not in experts [3]. However, due to differences in the design between the three experiments, a direct assessment of interactions with expertise level was not possible.

62

4

Several studies have shown that it is not mere explanation length (associated with adding SNI) that affects the positive rating of explanations combined with SNI [5–7]. A follow-up study showed that participants rated explanations containing reductive information more favorably across several scientific disciplines [1]. These authors also showed that participants with higher scientific literacy or those who had taken science courses demonstrated higher proficiency in distinguishing between good and bad explanations. In a large meta-analysis, Bennett and McLaughlin confirmed the presence of the SANE effect across 60 experiments from 28 publications with a sample size of 13,800 participants. Considering only the ratings of laypeople, they found a mild though highly significant impact of neuroscience information on evaluating information [8].

Here, we report the outcome of two studies. Study 1 comprises a classroom experiment that we had developed to alert students to ‘crap’ [9] and how to distinguish valid explanations from mere repetitions of facts. The classroom experiment was built on the study by Weisberg and colleagues [3], and used one item (see Table 1) of their stimulus material. Over the years, we observed major variations in the aggregated ratings of the students and, overall, a pattern that deviated from the one described in the original study. We thus decided to embark on a conceptual replication study of the original investigation. The aim of Study 2 was to extend systematic investigations of the SANE effect into the German-speaking community and to place our findings from the classroom experiment into a broader context. Following the hypothesis that the impact of SNI on the assessment of explanation quality depends on both the quality of the explanation per se and the degree of science literacy, we predicted that subjects would rate good explanations higher than bad ones and that SNI would boost the ratings of bad explanations specifically for participants with less expertise in the

neurosciences. We fitted a three-way interaction between explanation quality, SNI, and expertise to test this prediction.

6

2. Methods

2.1 Classroom experiment (Study 1)

The classroom experiment was part of the lecture “Good Scientific Practice” by the last author. This lecture is mandatory for 2nd or 3rd year Bachelor students of Biology and Biochemistry at the University of Göttingen. The experiment was first run in 2008 and every year since then. The cover story was that students would learn about the “Curse of Knowledge.” They were presented with the text (translated into German, see Supplementary Material) of item 1 of the phenomena used by Weisberg and colleagues in their original study (Table 1). Mirroring the 2x2 design of the original study, four different types of explanations (good and bad explanation with or without SNI) were printed separately on paper slips, mixed, and distributed among the students. Students were then asked to rate their satisfaction with the explanation (without consulting each other) on a scale from -2 to 2 (simplifying the original design that used a rating from -3 to 3 to facilitate data collection in the classroom). Each student rated only one explanation. We then collected the ratings, typed the numbers into a prepared spreadsheet, and visualized the results in front of the students. Subsequently, we debriefed the students and compared their ratings to those in the original study. During the COVID-19 pandemic (2020 and 2021), data collection took place online, and students participated in multiple ratings; we did not use these data in the present analysis. The data presented here span 2008-2022 and involved ratings by N=887 students. We collected no personal information (age, gender) from the students, and participation was voluntary. We did not fit a statistical model to these data but simply illustrated the raw data with the means and the bootstrapped confidence intervals.

7

Table 1. Item 1 of 18, as established by Weisberg and colleagues [3], with good and bad explanations in the absence and presence of superfluous neuroscience information.

Phenomenon: Researchers created a list of facts that about 50% of people knew. Subjects in this experiment read the list of facts and had to say which ones they knew. They then had to judge what percentage of other people would know those facts. Researchers found that the subjects responded differently about other people's knowledge of a fact when the subjects themselves knew that fact. If the subjects did know a fact, they said that an inaccurately large percentage of others would know it too. For example, if a subject already knew that Hartford was the capital of Connecticut, that subject might say that 80% of people would know this, even though the correct answer is 50%. The researchers call this finding "the curse of knowledge."		
	Explanation	
	Good	Bad
Without SNI	The "curse of knowledge" happens because subjects have trouble switching their point of view to consider what someone else might know. People mistakenly project their own knowledge onto others.	The "curse of knowledge" happens because subjects make more mistakes when they have to judge the knowledge of others. People are much better at judging what they themselves know.
With SNI	**Brain scans indicate that this "curse" happens because of the frontal lobe brain circuitry, known to be involved in self-knowledge**. Subjects have trouble switching their point of view to consider what someone else might know, mistakenly projecting their own knowledge onto	**Brain scans indicate that this "curse" happens because of the frontal lobe brain circuitry, known to be involved in self-knowledge.** Subjects make more mistakes when they have to judge the knowledge of others. People are much better at judging what they themselves know.

8

	others.	

The neuroscience information is highlighted here, but subjects did not see such markings. The full table with all items used in Study 2 can be found at https://osf.io/4mwje/?view_only=f576e2a957f448718b431d59db4a27dd.

9

2.2 Online experiment (Study 2)

The study design conceptually followed the design of the original study (Weisberg et al., 2008), using a within-subject design throughout, however. Study 2 was preregistered (https://doi.org/10.17605/OSF.IO/7BZ84). Each participant was presented with four out of the 18 scientific phenomena previously established in [3]. After preregistration, we slightly edited the stimulus material. Specifically, we presented the neuroscience information in a separate sentence to keep the actual explanation constant in the conditions with and without SNI.

Moreover, we edited content that referred to gender and race-related issues to avoid repeating gender stereotypes and issues with using the term “race” in German, where it has a pejorative connotation. The original, adapted, and translated versions of the survey items are deposited at https://osf.io/4mwje/?view_only=f576e2a957f448718b431d59db4a27dd. The study was hosted on the survey platform “shinyapps.io” by Posit Software (https://www.shinyapps.io). Initially, the planned sample size was n = 1000, but due to time constraints, we had to terminate the study when we had reached 584 participants.

Participants accessed the survey through a shared link. This link was distributed across various platforms, including social media, email distributions, and websites. Participants started the survey by selecting their preferred language (German or English) and then gave consent for the use of their data. Next, they received a concise overview of the study procedure, providing a brief description without revealing too much detail (Supplementary Material).

Participants were asked to provide information on their expertise by choosing one of the six levels depicted in Table 3. Each participant was then presented with four different phenomena (items), which previously had been randomly paired with one of the factor combinations of interest (good/bad and with

or without SNI), such that each participant was confronted once with each of the four factor combinations of interest. They were required to rate the quality of the explanation on a 7-point scale, ranging from -3 to +3. Between rating the items, participants were asked to enter their age in years and gender (options: female, male, non-binary, and prefer not to say) and solve a simple math question. This information was not used in the analysis. The survey took approximately 10 minutes to complete. In the end, participants could contact the study's first author via email to receive updates on the study's results.

We randomized the order in which we presented the factor combinations and items to each participant. However, we restricted the randomization such the sample was as close as possible to being counterbalanced with regard to (i) the frequency with which items were presented, (ii) the frequency with which items were combined with each of the four combinations of explanation quality and presence of SNI, (iii) the frequency with which items were presented at the first, second, third, and fourth position to participants, and (iv) the frequency with which the four combinations of explanation quality and presence of SNI were presented to the participants.

After excluding data entries with missing values (e.g., lacking quality rating), the original sample size of 584 participants was reduced to 430. The final data set comprised 1624 ratings from 430 participants, rating 18 different phenomena. Three hundred eighty-six participants completed all four questions, eleven participants rated only three of the four explanations, 14 rated only two, and 19 rated only one explanation. Unlike the preregistered plan, we also included participants with less than four ratings in our analysis. The reason for doing so was that we had reached much less than the originally planned 1000 participants and thus wanted to maximize the use of the ratings available to us.

177

12

Table 2: Level of expertise queried from the participants who gave at least one rating.

Level	Expertise	N
0	No High School Diploma	37
1	High School Diploma	124
2	University Degree	112
3	Science Degree	73
4	PhD in Neuroscience and related fields	33
5	Postdoc, Lecturer, or Professor in Neuroscience and related fields	51

2.3 Data analysis

To estimate the extent to which satisfaction ratings were influenced by explanation quality, the presence of superfluous neuroscience information (SNI), and level of expertise, we fitted a Generalized Linear Mixed Model [10]. It included fixed effects of explanation quality and SNI, level of expertise, and all their interactions up to the third order. We included a fixed effect of first language (German or English) and trial number to control for their possible effects on the ratings. Since the response was a rating, we used a cumulative logit link, also known as ordinal model [11].

Equation 1: Simplified model equation. The full model equation can be found in the supplemental material.

rating ~ quality*presence of SNI*education level + language + trial.nr +
(1|item ID) + (1|participant ID)

Most participants rated up to four items, and items were used multiple times (between 83 and 106 times), so we included random intercept effects for these two factors. These control for possible variation among participants and also among the items with regard to the average rating. We included all theoretically identifiable random slopes to keep the type 1 error rate at the nominal level of 0.05 and avoid an “overconfident” model [12,13]. These were those of explanation quality and presence of SNI within participant ID, as well as those of explanation quality, SNI, first language, level of expertise, their interactions up to order three, and trial number within item ID. Initially, we also included a random slope of trial number within-participant and item and parameters for the correlations among random intercepts and slopes. However, to avoid an overly complex model and convergence issues, we removed them from the model again.

As an overall test of the effects of explanation quality, SNI, level of expertise, and their interactions and to avoid cryptic multiple testing [14], we compared the full model as described above with a null model lacking explanation quality, SNI, level of expertise, and their interactions in the fixed effects part. For the full-null model comparison, we used a likelihood ratio test [15]. We tested the significance of individual fixed effects by dropping them from the model, one at a time, and comparing the likelihood of the resulting reduced models with that of the full model (R function drop1). As the full-null model comparison indicated a significant effect of the predictor variables, but the three-way interaction between the presence of SNI, explanation quality, and explanation quality was not significant, we removed the three-way interaction from the model. The removal

allowed us to assess the contributions of the three two-way interactions between the presence of SNI, explanation quality, and level of expertise. Two of these were not significant (quality*SNI and quality*expertise), so we removed them, too, which allowed us to assess the effect of the remaining interaction between SNI and expertise and the main effect of quality in the final model.

We fitted the model in R (version 4.4.0 [16]) using the function clmm of the package ordinal (version 2023.12-4 [17]). We included the level of expertise as a numeric predictor, although it is an ordered factor. We did so for three reasons: first, doing so reduced model complexity considerably, particularly as we would have had to add random slopes for that factor and the interactions with it within item. Second, it turned out that level of expertise only had a moderate influence on satisfaction ratings, in the sense that, per combination of level of expertise, question quality, and presence of SNI, the actual satisfaction ratings varied a lot and generally covered the entire space of the available ratings (see Fig. 2). Finally, plotting the data and the model did not reveal hints for level of expertise being included as a quantitative predictor having led to an obvious mismatch between the model and the response.

Before fitting the model, we z-transformed level of expertise and trial number to a mean of zero and a standard deviation of one to ease model convergence. We manually dummy-coded and centred all factors before including them as random slopes. We determined model stability by dropping individual participants and individual items, one at a time, fitting the full model to each of the subsets, and finally comparing the range of estimates obtained with those that we had gathered for the full data set. This procedure revealed the model to be of good stability. We estimated confidence limits of model estimates and fitted values (see below) using a parametric bootstrap (n=1000 bootstraps). Collinearity among the

predictors, assessed using Variance Inflation Factors [VIF; 18] was not an issue (maximum VIF: 1.2; determined for a standard LMM lacking the interactions using the function vif of the package car version 3.1-2 [19]).

In Fig. 2 and 3, we included ‘fitted values’ with confidence limits. These were derived as follows: we first determined fitted values (also for each bootstrap) in terms of the probability of observing a given satisfaction rating for a given constellation of values of the predictors. We then determined the weighted mean of the response, treating it as a quantitative variable, with the weights being the probabilities of the individual ratings. Given the ordinal nature of the satisfaction ratings, we know that treating the response as a quantity is not fully appropriate. However, adding such ‘fitted values’ and their confidence limits gives an intuitive illustration of the model’s findings.

Of the 72 possible item-explanation combinations (18 items x 4 explanation types), each combination occurred 16 to 32 times in the survey. The ages of the participants ranged from 16 to 83 years. The genders in the sample consisted of N=286 persons who identified themselves as females, N=129 as males, N=4 persons who identified themselves as non-binary, and N=9 who preferred not to provide information on their gender. Over half of the participants selected German (n=246) as their language, while 184 participants chose English. The sample consisted of participants across all six levels of expertise (Table 3). Data and code are deposited at https://osf.io/4mwje/?view_only=f576e2a957f448718b431d59db4a27dd.

2.4 Ethical Note

No personal data were gathered, and participants maintained complete anonymity throughout the study. Before participation, participants were asked to provide consent. They were explicitly informed that they could withdraw anytime without a reason. Minimal information about individual participants, such as age, gender, and level of expertise, was collected. There was no possibility of identifying participants through this information. The Ethics Committee of the Georg Elias Müller Institute of Psychology approved the study at the Georg-August-University Göttingen under application number 359 on November 29, 2023.

17

3. Results

3.1 Study 1

For the item used with undergraduate students in the classroom experiment, good explanations without SNI yielded the highest ratings; good explanations with SNI yielded the second-highest ratings. The compound rating was 0.83 (on a scale from -2 to 2) for good explanations without SNI, 0.28 for good explanations with SNI, -0.04 for bad explanations without SNI, and 0.07 for bad explanations with SNI (Figure 1a). However, there was considerable variation between the student cohorts (Figure 1b).

18

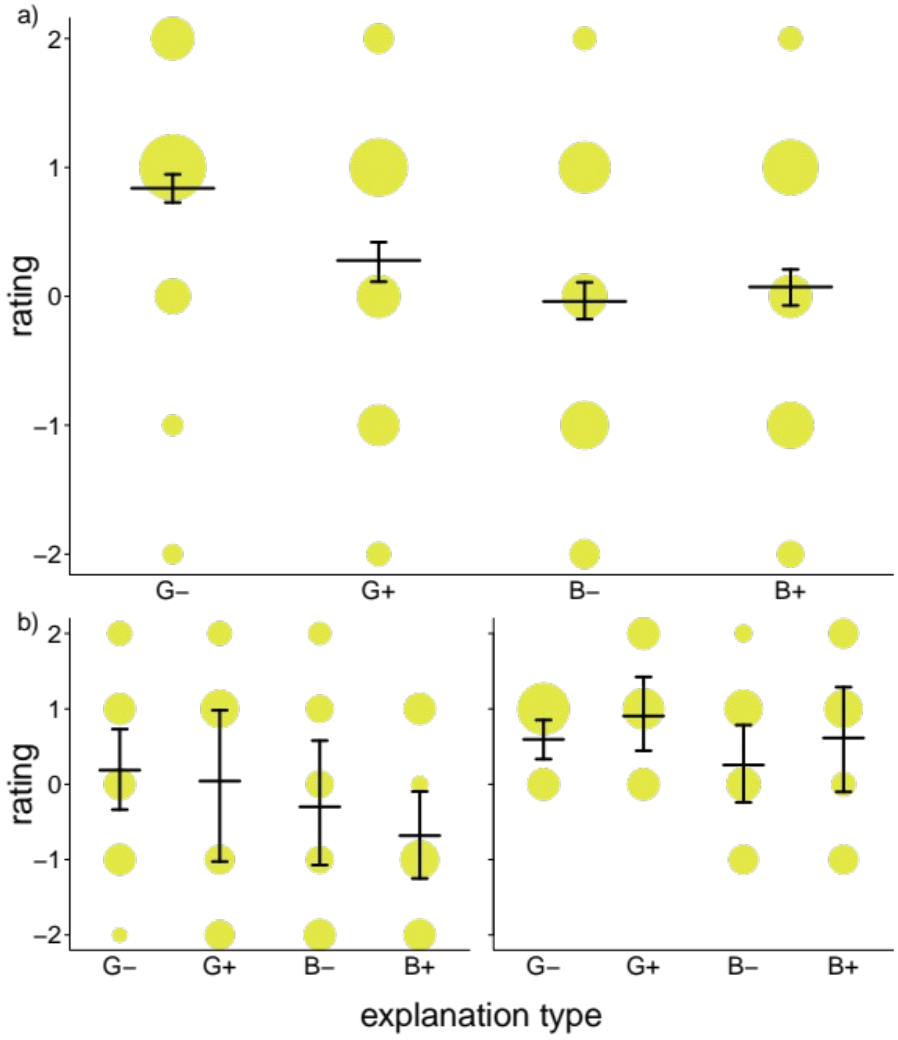

Figure 1. Satisfaction rating for good ('G') and bad ('B') quality explanations in the absence ('-') and presence of ('+') superfluous neuroscience information (SNI) in the classroom experiment. The area of the dots corresponds to the proportion of ratings per explanation type category. a) Pooled ratings across all study years (N=887 participants). b) Ratings for the years 2014 (N=62) and 2019 (N=48). Horizontal lines with error bars are fitted means with 95% confidence limits.

3.2 Study 2

In the online survey, we found a clear effect of explanation quality, presence of superfluous neuroscience information (SNI), level of expertise, and one or several

19

interactions on participants' rating behavior (full-null model comparison, likelihood ratio test: LRT, $\chi^2 = 66.3$, df = 7, $P < 0.001$). Supplementary Table S1 provides the full model output. As the predicted three-way interaction between explanation quality, SNI, and level of expertise was not significant, we removed it to assess the effects of the remaining two-way interactions. Of these, the one between superfluous neuroscience information and level of expertise showed a significant effect ($\chi^2 = 17.0$, df = 1, $P < 0.001$). The other two-way interactions between explanation quality and SNI, as well as explanation quality and level of expertise, were not statistically significant (Supplementary Table S2). We, therefore, removed these non-significant two-way interactions to assess the main effect of explanation quality.

The final model revealed that good explanations were rated higher than bad ones (effect of quality: $\chi^2 = 14.6$, df = 1, $P < 0.001$; see Table 2 for the full model output; Fig. 2). Furthermore, at lower levels of expertise, the presence of SNI was associated with higher ratings than when SNI was absent, and the effect of the presence of SNI decreased with increasing levels of expertise (Fig. 3).

20

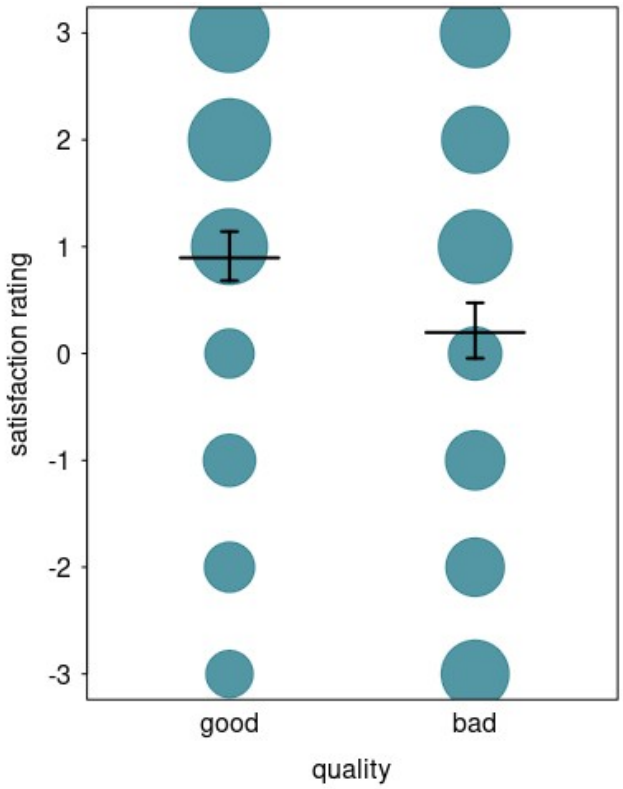

Figure 2. Satisfaction rating as a function of explanation quality. The area of the dots corresponds to the number of data points with identical values in satisfaction rating and story quality (range = 63 to 190, n = 1624 ratings). Horizontal lines with error bars represent fitted average ratings and 95% confidence limits (obtained from the reduced model lacking all non-significant interactions).

21

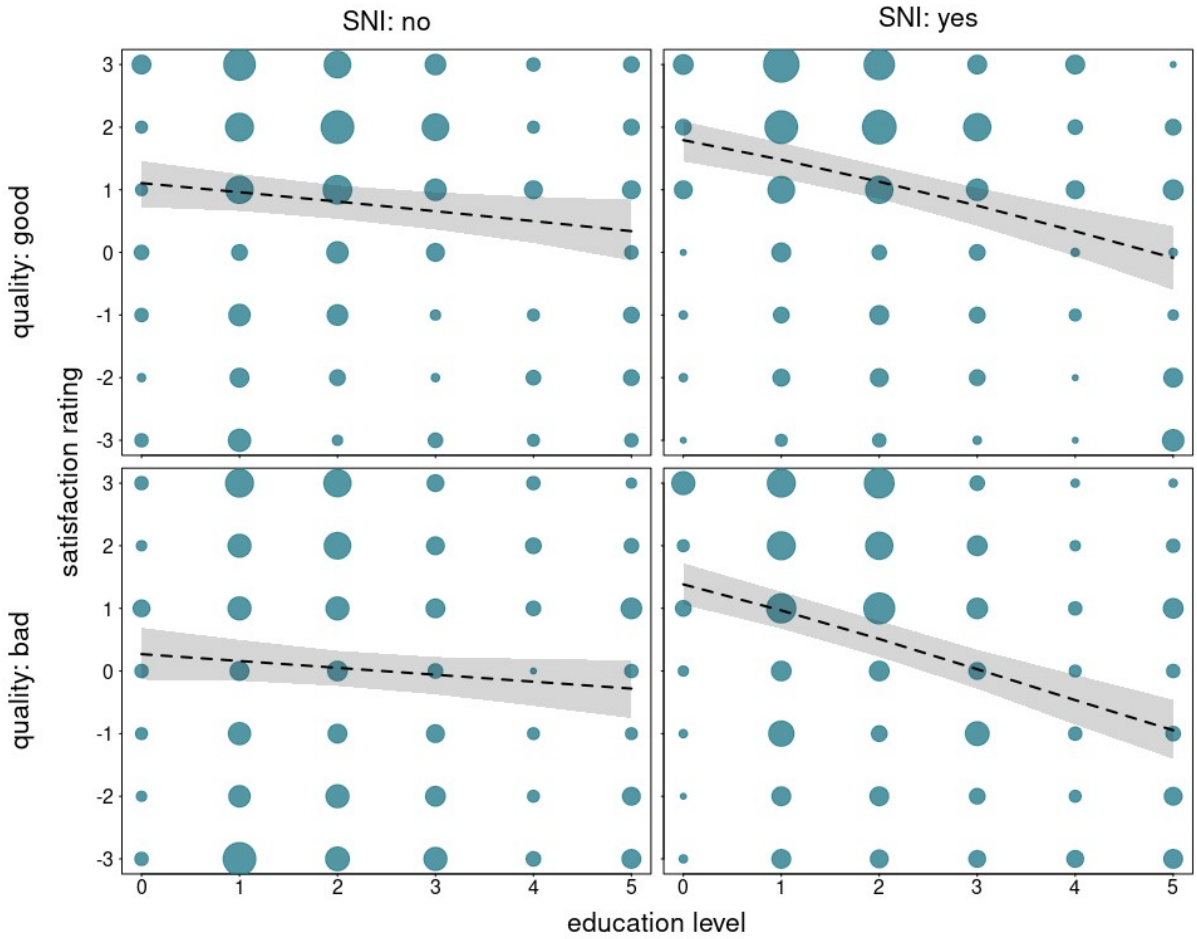

Figure 3. Satisfaction rating for good (upper) and bad (lower) quality explanations as a function of level of expertise in the absence (left) and presence (right) of superfluous neuroscience information (SNI). The areas of the dots correspond to the number of data points for level of expertise, satisfaction level, and SNI presence within each of the two explanation qualities (range = 1 to 35, n = 1624 ratings). The dashed lines with grey polygons represent the fitted averaged rating and their 95% confidence limits (obtained from the full model).

22

Table 2. Results for fixed effects part of the reduced model lacking all non-significant interactions.

Term	Estimate	SE	Cllower	Clupper	□2	df	P
-3|-2	-2.174	0.207	-2.528	-1.851			
-2|-1	-1.272	0.196	-1.589	-0.956			
-1|0	-0.579	0.192	-0.886	-0.267			
0|1	-0.078	0.191	-0.388	0.225			
1|2	0.932	0.193	0.635	1.244			
2|3	2.164	0.204	1.850	2.512			
					14.604		<0.001
Quality	0.699	0.146	0.475	0.914		1	
SNI	0.326	0.093	0.149	0.504			
education level	-0.192	0.083	-0.357	-0.040			
Language	-0.291	0.150	-0.543	-0.032	3.630	1	0.057
					11.628		
trial nr.	0.169	0.047	0.081	0.264		1	0.001
SNI:education					17.014		<0.001
level	-0.446	0.095	-0.616	-0.253		1	

Indicated are fixed effects estimates, standard errors, 95% confidence limits, significance tests, and the range of estimates obtained when dropping individual participants and items, one at a time. All factors were dummy coded with the reference levels being ‘bad’ (Quality), ‘present’ (SNI, ‘superfluous neuro information’), and ‘English’ (language). All covariates were z-transformed to a mean of zero and a standard deviation of one; the mean and standard deviation of the original variables were 2.223 and 1.454 (education level) and 2.455 and 1.120 (trial number).

23

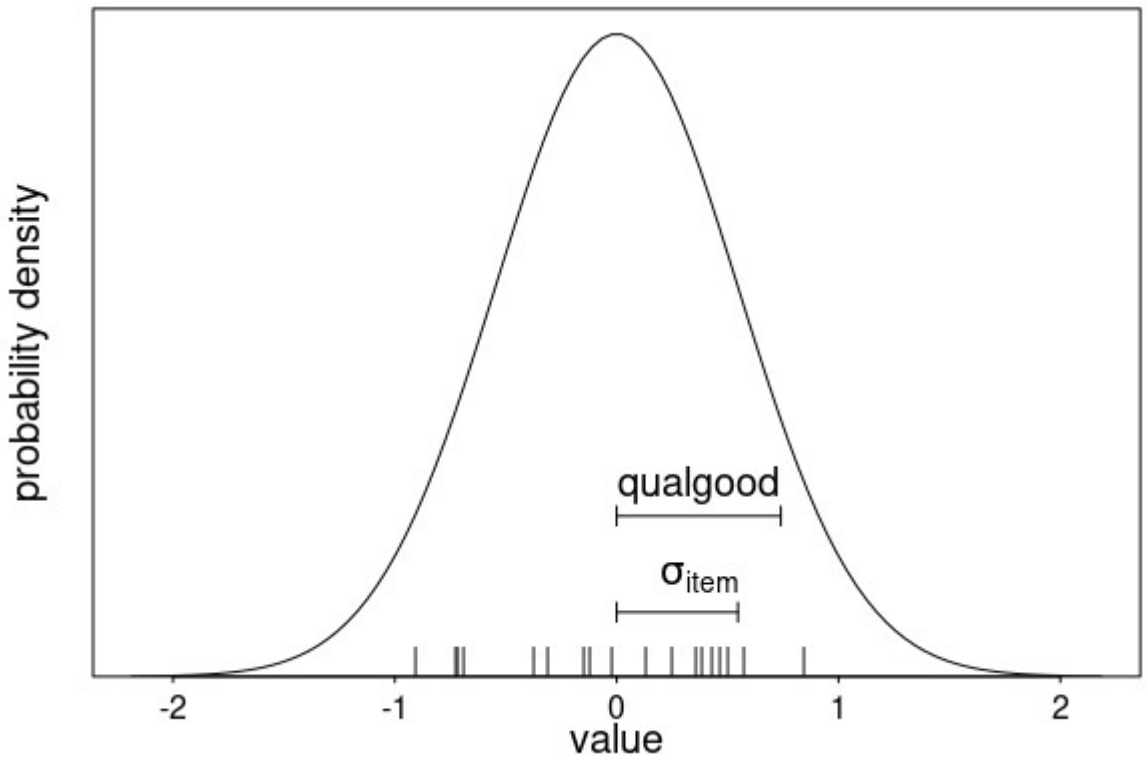

Figure 4. Illustration of the relative strength of the fixed main effect of quality and the random intercept effect of item. Vertical lines above the x-axis depict the estimated Best Linear Unbiased Predictors (BLUPs) for the random intercepts of the 18 different items. The depicted distribution is the probability density of BLUPs, which are assumed to be random samples from a normal distribution with a standard deviation as estimated for the random intercepts effect of item (σ_{item}, mean = 0, estimated SD = 0.55; the x-axis is depicted in link space and the results shown are taken from the full model). In addition, we show the estimated fixed effect of quality (good vs. bad, 0.68). Note that the variation in ratings due to different questions was of about the same magnitude as the difference in ratings of good and bad explanations.

There was considerable variation in ratings in relation to the items used in the study. Figure 3 shows the distribution of the Best Linear Unbiased Predictors (BLUPs) for the 18 items in the study. BLUPs ranged from -0.91 to 0.84. For comparison, the effect of explanation quality in the full model was 0.74. Interestingly, the pattern was very similar for item 1, which was used in the classroom and the online study. For this comparison, we extracted the ratings for

24

level of expertise 2-4 (N = 209), which broadly corresponded to the levels expected of the students in the classroom experiment. Good explanations without SNI received the highest ratings (1.56 in the classroom experiment (maximum value: 2) and 0.89 in the online study (max. value: 3), good explanations with SNI the second highest ratings (1.00 and 0.28), while bad explanations received similarly poor ratings (without SNI -0.33 and -0.04; with SNI -0.40 and 0.07, respectively.

4. Discussion

Both in the classroom experiment and the online study, participants rated proper explanations higher than mere repetitions of the phenomenon. In the online study, the addition of superfluous neuroscience information differentially affected satisfaction ratings for persons with different levels of expertise: participants with less expertise rated bad explanations with SNI higher than those without SNI. Broadly, we corroborated the findings from previous studies on the SANE effect. There were, however, also some differences compared to the original study. While the original study [3] identified a significant interaction between explanation quality and the presence of superfluous neuroscience information in novices and students, we did not find evidence for such an interaction (Supplementary Material Table S1).

Moreover, in contrast to the original results, where experts rated good explanations with SNI as less satisfying than those without SNI and bad explanations similarly, irrespective of the presence or absence of SNI [3], we found no evidence for similar effects in the current study. Our study also revealed substantial differences in satisfaction ratings across the different items. Although

Weisberg et al. [3] reported that subjects tended to respond similarly to all 18 items (Cronbach's = .79), the variation in ratings for different items in our study partly exceeded that of the differences between good and bad explanations. Note that we could not calculate Cronbach's for our data set due to differences in study design. Future studies need to consider such variation; it may also be necessary to check the stimulus material in greater depth before embarking on future studies.

The minor discrepancies between our study and the original one may be due to several not mutually exclusive reasons, including stochastic effects, the inclusion of a second language (German), the slight editing of the items and explanations, differences in the design of the study (within- and between-subject designs), and the total number of ratings per participant. The absence of significant interactions between explanation quality and the presence of SNI and between explanation quality and level of expertise may also be attributed to differences in statistical analyses. This study used a cumulative logit link model (CLLM), whereas the original study employed repeated measure ANOVA to assess significance [3]. The advantage of CLLM models is that they do not assume normally distributed and homogeneous residuals, which are unlikely to be met with a bound and discrete response [20]. Moreover, they correctly limit fitted values and their confidence limits to the bound space of the response [20]. Yet, despite this range of possible sources of divergence, our findings are broadly comparable to those of previous studies, supporting the reliability and generalizability of the SANE effect [3,7,8].

The SANE effect has been attributed to neuroscience's ability to offer reductive explanations [1,3,21,22]. Reductionist explanations provide insights into fundamental principles, rendering neuroscience explanations fitting for

understanding psychological phenomena [5,23]. Considering the esteemed status of neuroscience as a scientific discipline and the tendency for scientific jargon to create a false sense of comprehension, it is reasonable to suggest that the allure of neuroscience may be driven by its prestige, aligning with the “prestige of science” hypothesis [5,23]. There is ample evidence that expertise moderates the SANE effect [8]. Individuals who exhibited lower levels of reflection and reduced proficiency in verbal fluidity and numeracy were found to be more susceptible to accepting what could be termed “bullshit” [24], leading individuals to mistake vagueness for profundity [24,25]. Such individuals were also likelier to judge pompous explanations as accurate and meaningful despite being hollow [24]. However, if individuals receive specific training on the SANE effect in their education, they may be more likely to detect it [1,3,7,24].

What we found striking (not only in our results) was the small differences in the ratings of good and bad explanations. The mean ratings for the good explanations differed very little: they were 0.89 in our study and 0.88 in the original one; greater differences were observed in the ratings of bad explanations, which were 0.20 in our study and -0.28 in the original one [3]. In other studies, good explanations received ratings of 0.70 [5] and 1.23 [1] higher than bad ones. Hopkins et al. [21] explored the effect of expertise on rating behavior in more detail. They found that expert participants rated good explanations (mean = 1.53, s.d. = 1.56) significantly higher than bad explanations (mean = −0.66, s.d. = 1.97), but there was considerable variation. Notably, the ‘bad’ explanations were not explanations at all but simple descriptions of the results in other words. Yet, in our study, even some of the experts with several years in academia rated such ‘explanations’ highly with 2 or 3 points (Fig. 2). Indeed, in our study, experts did not rate bad explanations significantly worse compared to non-experts, as evidenced by the lack of an interaction between quality and expertise.

27

Why were experts unable to distinguish between good and bad explanations more accurately than non-experts, even though expertise should improve the ability to evaluate explanations critically? In a previous study by Goldberg and Thompson-Schill [26], biology professors demonstrated less accuracy in their responses to statements about biology that contradicted basic principles than statements aligning with them. Known as the “curse of expertise,” experts tend to overestimate their knowledge in their specialized field [27]. Similar to the “illusion of explanatory depth” in laypeople, expertise can create the illusion of competence within experts, leading individuals to believe they have a deeper understanding of a particular topic than they do [27–29]. Developing a curriculum that trains students to distinguish between actual explanations and “bullshit” would ultimately also fire up the “crap detectors” [9] of teachers and experts.

For science communicators, insights into the SANE effect and its variants underscore the responsibility of communicators to present information accurately and transparently without relying on jargon to influence perception [1,21,30,31]. As emphasized by Silas and colleagues [31], there is a pressing need for public health communication to prioritize clarity, simplicity, and accessibility. They conducted a study that explored how explanations with scientifically irrelevant neuroscience affect intentions to vaccinate against COVID-19, demonstrating a significant impact of poor scientific communication on vaccination intentions [31]. Additionally, it was suggested that public health efforts are at risk of being sabotaged by misinformation that successfully uses technical language to persuade people to believe nonsensical explanations [31]. Contrary to the notion that technical language lends credibility, the data suggest that clear, simple, and straightforward information is ultimately more effective in public health communication [30–32].

28

References

1. Hopkins EJ, Weisberg DS, Taylor JCV. 2016 The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition* **155**, 67–76. (doi:10.1016/j.cognition.2016.06.011)

2. McCabe DP, Castel AD. 2008 Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition* **107**, 343–352. (doi:10.1016/j.cognition.2007.07.017)

3. Weisberg DS, Keil FC, Goodstein J, Rawson E, Gray JR. 2008 The Seductive Allure of Neuroscience Explanations. *J. Cogn. Neurosci.* **20**, 470–477. (doi:10.1162/jocn.2008.20040)

4. O'Connor C, Rees G, Joffe H. 2012 Neuroscience in the Public Sphere. *Neuron* **74**, 220–226. (doi:10.1016/j.neuron.2012.04.004)

5. Fernandez-Duque D, Evans J, Christian C, Hodges SD. 2015 Superfluous Neuroscience Information Makes Explanations of Psychological Phenomena More Appealing. *J. Cogn. Neurosci.* **27**, 926–944. (doi:10.1162/jocn_a_00750)

6. Rhodes RE, Rodriguez F, Shah P. 2014 Explaining the alluring influence of neuroscience information on scientific reasoning. *J. Exp. Psychol. Learn. Mem. Cogn.* **40**, 1432–1440. (doi:10.1037/a0036844)

7. Weisberg DS, Taylor JCV, Hopkins EJ. 2015 Deconstructing the seductive allure of neuroscience explanations. *Judgm. Decis. Mak.* **10**, 429–441. (doi:10.1017/S193029750000557X)

8. Bennett EM, McLaughlin PJ. 2024 Neuroscience explanations really do satisfy: A systematic review and meta-analysis of the seductive allure of neuroscience. *Public Underst. Sci.* **33**, 290–307. (doi:10.1177/09636625231205005)

9. Postman N. 1969 Bullshit and the art of crap-detection. See https://aquadoc.typepad.com/files/bs_speech_postman-1.pdf.

10. Baayen RH. 2008 Analyzing linguistic data. Cambridge, UK.

11. Agresti A. 2012 *Categorical Data Analysis*. John Wiley & Sons.

12. Barr DJ, Levy R, Scheepers C, Tily HJ. 2013 Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255–278. (doi:10.1016/j.jml.2012.11.001)

13. Schielzeth H, Forstmeier W. 2009 Conclusions beyond support: overconfident estimates in mixed models. *Behav. Ecol.* **20**, 416–420. (doi:10.1093/beheco/arn145)

14. Forstmeier W, Schielzeth H. 2011 Cryptic multiple hypotheses testing in linear

models: overestimated effect sizes and the winner's curse. *Behav. Ecol. Sociobiol.* **65**, 47–55. (doi:10.1007/s00265-010-1038-5)

15. Dobson AJ. 2001 *An Introduction to Generalized Linear Models*. 2nd Edition. New York, NY, USA: Chapman & Hall/CRC.

16. R Core Team. 2024 R: A Language and Environment for Statistical Computing.

17. Christensen RHB. 2018 Cumulative Link Models for Ordinal Regression with the R Package ordinal.

18. Field A. 2013 *Discovering Statistics Using IBM SPSS Statistics*. SAGE.

19. Fox J, Weisberg S. 2018 *An R Companion to Applied Regression*. SAGE Publications.

20. Bürkner P-C, Vuorre M. 2019 Ordinal Regression Models in Psychology: A Tutorial. *Adv. Methods Pract. Psychol. Sci.* **2**, 77–101. (doi:10.1177/2515245918823199)

21. Hopkins EJ, Weisberg DS, Taylor JCV. 2019 Does expertise moderate the seductive allure of reductive explanations? *Acta Psychol. (Amst.)* **198**, 102890. (doi:10.1016/j.actpsy.2019.102890)

22. McCabe DP, Castel AD. 2008 Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition* **107**, 343–352. (doi:10.1016/j.cognition.2007.07.017)

23. Fernandez-Duque D. 2017 Lay Theories of the Mind/Brain Relationship and the Allure of Neuroscience. In *The Science of Lay Theories* (eds CM Zedelius, BCN Müller, JW Schooler), pp. 207–227. Cham: Springer International Publishing. (doi:10.1007/978-3-319-57306-9_9)

24. Pennycook G, Cheyne JA, Barr N, Koehler DJ, Fugelsang JA. 2015 On the reception and detection of pseudo-profound bullshit. *Judgm. Decis. Mak.* **10**, 549–563. (doi:10.1017/S1930297500006999)

25. Sperber D. 2010 The Guru Effect. *Rev. Philos. Psychol.* **1**, 583–592. (doi:10.1007/s13164-010-0025-0)

26. Goldberg RF, Thompson-Schill SL. 2009 Developmental "Roots" in Mature Biological Knowledge. *Psychol. Sci.* **20**, 480–487. (doi:10.1111/j.1467-9280.2009.02320.x)

27. Fisher M, Keil FC. 2016 The Curse of Expertise: When More Knowledge Leads to Miscalibrated Explanatory Insight. *Cogn. Sci.* **40**, 1251–1269. (doi:10.1111/cogs.12280)

28. Keil FC. 2003 Folkscience: coarse interpretations of a complex reality. *Trends Cogn. Sci.* **7**, 368–373. (doi:10.1016/S1364-6613(03)00158-X)

29. Rozenblit L, Keil F. 2002 The misunderstood limits of folk science: an illusion of explanatory depth. *Cogn. Sci.* **26**, 521–562. (doi:10.1207/s15516709cog2605_1)

30. Baur C, Prue C. 2014 The CDC Clear Communication Index Is a New Evidence-Based Tool to Prepare and Review Health Information. *Health Promot. Pract.* **15**, 629–637. (doi:10.1177/1524839914538969)

31. Silas J, Jones A, Weiss-Cohen L, Ayton P. 2021 The seductive allure of technical language and its effect on covid-19 vaccine beliefs and intentions. *Vaccine* **39**, 7590–7597. (doi:10.1016/j.vaccine.2021.11.027)

32. Bullock OM, Colón Amill D, Shulman HC, Dixon GN. 2019 Jargon as a barrier to effective science communication: Evidence from metacognition. *Public Underst. Sci.* **28**, 845–853. (doi:10.1177/0963662519865687)