Borghi, Olaf; Shankar, Sahana

**Working Paper**

# A Comment on "Bayesianism and Wishful Thinking are Compatible"

I4R Discussion Paper Series, No. 198

**Provided in Cooperation with:**
The Institute for Replication (I4R)

*Suggested Citation:* Borghi, Olaf; Shankar, Sahana (2025) : A Comment on "Bayesianism and Wishful Thinking are Compatible", I4R Discussion Paper Series, No. 198, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/309441

INSTITUTE for REPLICATION

No. 198
I4R DISCUSSION PAPER SERIES

# A Comment on "Bayesianism and Wishful Thinking are Compatible"

Olaf Borghi

Sahana Shankar

January 2025

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 198

# A Comment on "Bayesianism and Wishful Thinking are Compatible"

**Olaf Borghi[1], Sahana Shankar[1]**

*[1]Royal Holloway, University of London, Egham/Great Britain*

JANUARY 2025



E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

# A comment on “Bayesianism and wishful thinking are compatible”

*This comment was written as part of the UKRN Replication Games on December 5th 2024.*

**Reproduced article:**

Melnikoff, D.E., Strohminger, N. Bayesianism and wishful thinking are compatible. *Nat Hum Behav* 8, 692–701 (2024). https://doi.org/10.1038/s41562-024-01819-6

**Authors of comment:**

Olaf Borghi[1], MSc

Sahana Shankar[1], PhD

[1]*Department of Psychology, Royal Holloway, University of London*

**Code and data provided by the authors:**

https://osf.io/59dmr/?view_only=b8ea1a66b5e84d1e8d67391662b60d82

**Code used to reproduce results:**

https://osf.io/eyw5u/?view_only=a7eafa500ff1460cb5d590adf26fb36d

**Abstract**

Melnikoff and Strohminger (2024) report that affective prediction errors drive wishful belief updating, i.e., the tendency to adjust beliefs in the direction of one's desires. Historically, this phenomenon has posed a challenge to Bayesian accounts of reasoning, which assume that beliefs are updated relative to the available evidence and prior beliefs. However, the authors propose that wishful belief updating can align with Bayesian principles when affective prediction errors as “hidden information signals” are taken into account. Across five experiments, the authors provide compelling evidence for this account and show that affective prediction errors systematically influence belief updates in the direction of desires, and they formalise this in a Bayesian model.

We were partially able to computationally reproduce the findings from Experiments 1, 2, and 3 using the provided code and data. Reported results from experiments 1 and 2 could be fully reproduced by inferring the statistical models based on reported results and rewriting part of the analysis code, as the provided code appeared incomplete. To assess robustness, we reanalyzed Experiments 1, 2, and 3 by including participants previously excluded for failing attention checks. The findings remained robust under these alternative specifications. For Experiment 4, data and code were not available on OSF. The data in the folder labeled “Experiment 4” instead appeared to provide data and code for Experiment 5. We used this to partially reproduce the findings from Experiment 5, but for more complex analyses, information on models and code was incomplete, making it impossible for us to fully reproduce the results. We provide the additional code we used to reproduce the findings.

# 1. Introduction

Melnikoff and Strohminger (2024) report that affective prediction errors (APEs) drive wishful belief updating, providing empirical evidence for this hypothesis across five experiments. In the present comment, prepared as a collaboration between the Institute for Replication and Nature Human Behaviour (Brodeur et al., 2024), we assess the computational reproducibility of the findings. We further evaluate the replicability and robustness of the reported results by re-analysing data without the exclusion of participants that failed attention checks. For all reproduction analyses, we display the p-values as reported in tables or figures from the main text labelled with a (e.g., Table 1a, Figure 1a) and as reproduced for this commentary labelled with the letter b (e.g., Table 1b). Figures and tables from robustness analyses are labelled with the letter c (e.g., Table 1c).

# 2. Computational Reproducibility

The authors provided raw data, code for data cleaning, and (incomplete) analysis code for four of the five experiments on OSF. Code used to generate figures and calculate confidence intervals was not provided. For experiment 5, the folder containing data and code was mislabelled as experiment 4. Data and code from experiment 4 were missing.

| | **Experiment** | | | | |
|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** |
| Raw data provided | Fully | Fully | Fully | No | Fully |
| Cleaning code provided | Fully | Fully | Fully | No | Fully |
| Analysis data provided | Fully | Fully | Fully | No | Fully |
| Analysis code provided | Fully | Partial | Partial | No | No/Partial? |
| Reproducible from raw data | Fully | Fully | Partial | No | No/Partial? |
| Reproducible from analysis data | Fully | Fully | Partial | No | No/Partial? |

# Experiment 1

We successfully reproduced the main results of the experiment from the raw data using the analysis code provided. The results reported in the article and from our reproducibility analyses are described in Table 1a and 1b and visualised in Figures 1a and 1b. All p-values could be exactly reproduced based on the provided data and code. Code for data visualisation was not provided, but a similar visualisation could be reproduced from the data.

# Experiment 2

We could computationally reproduce the results of experiment 2. However, for experiment 2, the provided analysis code was incomplete and did not match the reported coefficients in the article. In specific, to obtain the coefficients (p-values, and t-values) reported by the authors for the error based updating hypotheses and the follow up test, we had to specify linear models that include only main effects. The provided analysis code included interactions that result in coefficients that differ from those reported in the article.

To give an example, the following analysis code was used for the linear models in R for experiment 2:

*Model 1: lm(update ~ 1, data = dfClean)*

*Model 2: lm(delta ~ 1, data = dfClean)*

*Model 3: lm(update ~ delta*noise, data = dfClean)*

*Model 4: lm(update ~ obs*noiseLow + pred*noiseLow, data = dfClean)*

Model 1 tests wishful belief updating, i.e., whether the change in P(Guilty) after being assigned to the prosecutor role is greater than before any role assignment ($Guilty_{Post}$ - $Guilty_{Pre}$). Model 2 tests the underestimation hypothesis, stating that participants would underestimate how positive they would feel about being assigned the prosecutor role ($Affect_{post}$ - $Affect_{pre}$). These two models exactly reproduce the reported results from the data.

The next reported result in the main text tests the error-based updating hypothesis, namely that the wishful belief updating is predicted by affective prediction errors. However, the output of none of the provided linear models align with the coefficients reported in the main text "(*t*(567) = 6.29, two-tailed *P* < 0.001, *b* = 0.02, 95% CI = 0.016 to 0.03)." The most intuitive test of the hypothesis is a linear model that just predicts the belief update based on the error. Running this model allowed us to reproduce the reported results from the provided data (see Tables 2a and 2b).

The same was the case for the next test on the effects of predicted and observed affect. The provided code only includes Model 4, however, the interaction of observed and predicted affect with noise was not reported in the article. Instead, to reproduce the results it was again required to run an additional model that only includes main effects for the two affect measures, i.e., lm(update ~ obs + pred, data = dfClean).

These inconsistencies between code and reported results were not the only ones we noticed. While the code for the model for the interaction effect between the affective prediction error and observation noise appeared to be the intended one, for this model, the coding of effect directions seemed inconsistent with other models and cannot be directly reproduced from the provided code.

More specifically, the authors report “The main effect of APE was qualified by an interaction with our continuous measure of subjective observation noise (t(565) = 2.56, two-tailed P = 0.01, b = 0.01, 95% CI = 0.002 to 0.017; Fig. 2c ).” However, running the provided codes gives the following output: t(565) = -2.56, two-tailed *P* = 0.011, b = -0.009, 95% CI = -0.017, -0.002. To obtain the effect coded in a positive direction, in contrast to other analyses, either the affective prediction error needs to be coded as ($Affect_{pre}$ - $Affect_{post}$) or observation noise needs to be reverse coded.

This is only a minor inconsistency in reporting and coding, and we want to highlight that the authors still draw conclusions in the correct direction, i.e., “greater levels of subjective observation noise were associated with weaker effects of APEs on wishful belief updating”. However, it is unclear at what stage of the analysis the recoding of effect directions took place, and why this is inconsistent across reported results.

## Experiment 3

We were partially able to reproduce the code for this experiment. Overall, the logic of the provided code paralleled the one for experiment 2. We again identified a few minor errors in the analysis. For this analysis, all participants were assigned the defending role. We assume to keep the sign of APEs positive, APEs for this study were calculated as $Affect_{pre}$ - $Affect_{post}$ in the provided code. Wishful belief updating, however, as in experiment 2 was coded as $Guilty_{Post}$ - $Guilty_{Pre}$.

In the reported results reported for the interaction of affective prediction errors and noise on wishful belief updating, the values for degrees of freedom, p, and b seem to have been copied from the results of experiment 2 and are thus incorrect in the context of experiment 3. To quote the corresponding paragraph in the article “The effect of APE on wishful belief updating was

qualified by a significant interaction with our continuous measure of subjective observation noise ($t$(565) = 2.56, two-tailed $P$ = 0.01, $b$ = 0.01, 95% CI = 0.002 to 0.019; Fig. 2f).”

The reproduced results are $t$(307) = -2.328, $P$ = 0.021, 95% CI = -0.019 to -0.002. Given the sample size of $N$ = 311 of experiment 3, the degrees of freedom, t- and p-values are likely those from experiment 2, and only the confidence interval appears to match the data and code for experiment 3. Note again that on several occasions reported signs in the article have been flipped, e.g., the authors report the correct confidence intervals but with positive signs.

In addition, experiment 3 includes a variable on “free choice”. In the methods section it is reported that this variable is based on a Likert scale ranging from 0 to 6. However, in the provided data, the range of the variable is 1 to 7. To obtain the same mean, median, mode and standard deviation of this variable as reported here “the modal response on our free choice scale was zero (median = 1, mean = 1.74, s.d. = 1.95)”, we had to recode the scale to 0-6.

Finally, we had several issues in our attempts to reproduce the following part of the results: “When predicted and observed affect were used simultaneously to predict wishful belief updating, we found a positive effect of observed affect ($t$(307) = 4.93, two-tailed $P$ < 0.001, $b$ = 0.03, 95% CI = 0.015 to 0.04) and a negative effect of predicted affect ($t$(307) = 2.01, two-tailed $P$ = 0.046, $b$ = −0.02, 95% CI = −0.03 to −0.0003).” Code for linear models for these analyses was not reported. We assumed that as in experiment 2, a model only including the main effects of observed and predicted affect, i.e., lm(update ~ obs + pred, data = dfClean) could allow us to reproduce the findings. We ran this model, and noticed that contrary to experiment 2, here it does not reproduce the reported results. Notable differences were the following: First of all, the model had 308 degrees of freedom, while the reported df = 307. It may be that a third variable was included as a fixed effect in the model, but from the article and code we could not find out which variable that was. Importantly, this also leads to a second difference. The p-value for the main effect of predicted affect in our model was $P$ = 0.052 and thus not significant, whereas the reported $P$ = 0.046. Providing the code used for this part of the analysis and a clearer indication of included variables can clarify these discrepancies.

Similarly, there was no code provided for the models testing the effect of free choice on whishful belief updating. We ran a model with a simple main effect (Update ~ freeChoice), but due to discrepancies in the degrees of freedom and results, again it appears additional variables were included in the analyses. However, as we were unable to infer what predictors the model the authors ran included, we could not reproduce these analyses.

## Experiment 4

The data and code for experiment 4 do not seem to be available on the OSF repository linked in the article (https://osf.io/59dmr/?view_only=b8ea1a66b5e84d1e8d67391662b60d82). The code and data for experiment 5 appears to have been mislabelled as experiment 4.

## Experiment 5

We tried, but ultimately did not reproduce the results of experiment 5. Given the increased complexity of analyses for Study 5, and as it was hard to match the provided code to the reported results, we were unable to computationally reproduce the findings of Study 5.

We noticed some minor inconsistencies in the provided code that had to be changed in order to attempt to reproduce results. In particular, dfClean and df_clean as names of dataframes were used, but only dfClean was defined in the code. E.g., for demographic statistics, the following code was used, but only the line with dfClean runs without error.

```
length(dfClean$subject) # N
mean(df_clean$sex == 2) # % female
median(df_clean$age) # median age
```

## 3. Robustness Reproduction

We carried out robustness reproduction by rerunning the analysis for experiments 1, 2 and 3 without excluding the data from participants who failed attention checks. We did not run robustness checks on experiment 4 (due to the unavailability of data) and experiment 5 (as we could not computationally reproduce the reported findings). We used the same code as provided by the authors (updated where necessary as highlighted above). For experiment 1 and 2 all tests remained significant and in the direction of the results reported in the manuscript (see Tables 1c and 2c).

For experiment 3, again, all main hypothesis tests of interest remain significant and in the direction of the results reported in the manuscript (see Table 3c). However, there are some notes to this. Above, we reported that we could not reproduce the significant effect of predicted affect on the belief update ($p = 0.052$). When including all participants in the model this effect is significant ($p < 0.001$), but the p-value still differs from the one reported in the paper ($p = 0.046$). This finding does not appear robust, but this was not one of the main hypothesis tests.

The authors also report that “perceptions of free choice had no effect on wishful belief updating ($t(304) = 0.57$, two-tailed $P = 0.569$, $b = -0.003$, 95% CI = −0.015 to 0.008)”. When including the full sample of respondents, free choice had a significant effect on wishful belief updating ($p = 0.027$). It is thus unclear, if with a larger sample size, the authors would have found an effect of free choice.

## 4. Conclusion

Our general assessment is that the reproducibility of the article by Melnikoff and Strohminger (2024) can be improved. The article is of overall high quality, but we had to reverse-engineer linear models from results to reproduce several findings (and failed to do so for experiment 5). In addition, the coding of effect directions is inconsistent in some cases even within experiments. Most importantly, data of experiment 5 was mislabelled as experiment 4 in the linked OSF repository, and data and code from experiment 4 was not provided. No code for the generation of figures and the calculation of confidence intervals was provided. The manuscript, supplementary materials, and provided data and code thus did not allow us to computationally reproduce all results. The results that we could reproduce, however, in most cases were robust against alternative data analytic choices. In particular we assessed whether the inclusion of dropped cases would alter the significance of tests, which was only the case for two secondary models.

## References

Brodeur, A., Dreber, A., Hoces de la Guardia, F. et al. Reproduction and replication at scale. Nat Hum Behav 8, 2–3 (2024). https://doi.org/10.1038/s41562-023-01807-2

Melnikoff, D.E., Strohminger, N. Bayesianism and wishful thinking are compatible. *Nat Hum Behav* 8, 692–701 (2024). https://doi.org/10.1038/s41562-024-01819-6

# Experiment 1

**Figure 1a.** *Visualisation of results of experiment 1 as reported in Melnikoff & Strohminger (2024)*

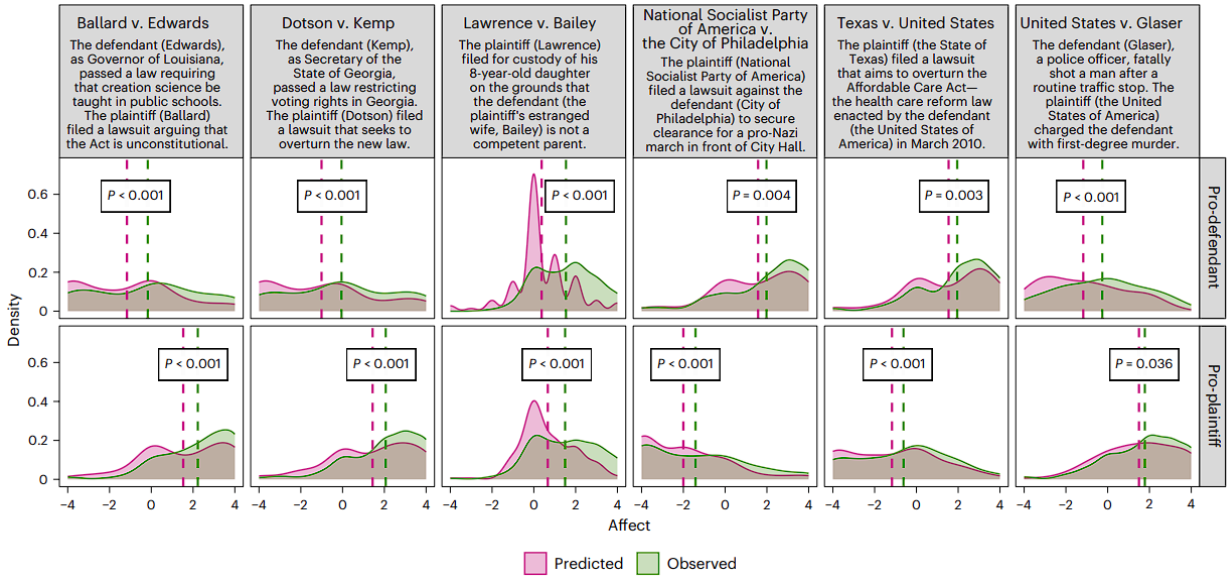


**Table 1a.** *Reported P-values of experiment 1 in Melnikoff & Strohminger (2024)*

| Case | Role | Coefficient | p-value |
|---|---|---|---|
| United States v. Glaser | Plaintiff | N/A | 0.036 |
| Texas v. United States | Plaintiff | N/A | < 0.001 |
| Lawrence v. Bailey | Plaintiff | N/A | < 0.001 |
| Ballard v. Edwards | Plaintiff | N/A | < 0.001 |
| National Socialist Part of America v. The City of Philadelphia | Plaintiff | N/A | < 0.001 |
| Dotson v. Kemp | Plaintiff | N/A | < 0.001 |
| United States v. Glaser | Defendant | N/A | < 0.001 |
| Texas v. United States | Defendant | N/A | 0.003 |
| Lawrence v. Bailey | Defendant | N/A | < 0.001 |
| Ballard v. Edwards | Defendant | N/A | < 0.001 |
| National Socialist Part of America v. The City of Philadelphia | Defendant | N/A | 0.004 |
| Dotson v. Kemp | Defendant | N/A | < 0.001 |

**Figure 1b.** *Visualisation of results of experiment 1 as reproduced for this commentary*

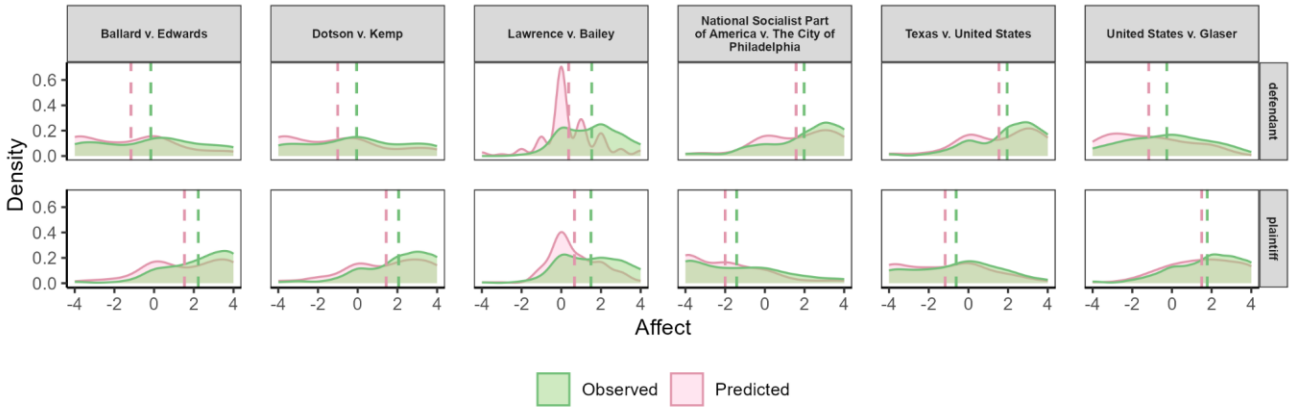


**Table 1b.** *Reproduced coefficients and p-values of experiment 1*

| Case | Role | Coefficient | p-value |
|---|---|---|---|
| United States v. Glaser | Plaintiff | 0.280 | 0.036 |
| Texas v. United States | Plaintiff | 0.556 | < 0.001 |
| Lawrence v. Bailey | Plaintiff | 0.830 | < 0.001 |
| Ballard v. Edwards | Plaintiff | 0.697 | < 0.001 |
| National Socialist Part of America v. The City of Philadelphia | Plaintiff | 0.582 | < 0.001 |
| Dotson v. Kemp | Plaintiff | 0.627 | < 0.001 |
| United States v. Glaser | Defendant | 0.910 | < 0.001 |
| Texas v. United States | Defendant | 0.411 | 0.003 |
| Lawrence v. Bailey | Defendant | 1.164 | < 0.001 |
| Ballard v. Edwards | Defendant | 1.000 | < 0.001 |
| National Socialist Part of America v. The City of Philadelphia | Defendant | 0.404 | 0.004 |
| Dotson v. Kemp | Defendant | 0.955 | < 0.001 |

**Table 1c.** *Robust coefficients and p-values of experiment 1 (no data exclusions)*

| Case | Role | Mean Delta | p-value |
|---|---|---|---|
| United States v. Glaser | Plaintiff | 0.352 | 0.007 |
| Texas v. United States | Plaintiff | 0.582 | < 0.001 |
| Lawrence v. Bailey | Plaintiff | 0.821 | < 0.001 |
| Ballard v. Edwards | Plaintiff | 0.694 | < 0.001 |
| National Socialist Part of America v. The City of Philadelphia | Plaintiff | 0.557 | < 0.001 |
| Dotson v. Kemp | Plaintiff | 0.628 | < 0.001 |
| United States v. Glaser | Defendant | 0.856 | < 0.001 |
| Texas v. United States | Defendant | 0.405 | 0.004 |
| Lawrence v. Bailey | Defendant | 1.065 | < 0.001 |
| Ballard v. Edwards | Defendant | 0.945 | < 0.001 |
| National Socialist Part of America v. The City of Philadelphia | Defendant | 0.385 | 0.005 |
| Dotson v. Kemp | Defendant | 0.894 | < 0.001 |

# Experiment 2

**Table 2a.** *Reported results of different linear models for experiment 2 in Melnikoff & Strohminger (2024)*

| | Update | APE | Update ~ APE | Update ~ obs + pred | Update ~ APE * Noise |
|---|---|---|---|---|---|
| (Intercept) | 0.04 [0.03, 0.05] | 0.53 [0.4, 0.65] | | | |
| | p < 0.001 | p < 0.001 | | | |
| APE | | | 0.02 [0.016, 0.030] | | |
| | | | p < 0.001 | | |
| obs | | | | 0.03 [0.02, 0.04] | |
| | | | | p < 0.001 | |
| pred | | | | -0.01 [-0.02, -0.005] | |
| | | | | p = 0.002) | |
| APE x Noise | | | | | 0.01 [ 0.002, 0.017] |
| | | | | | p = 0.01 |

**Table 2b.** *Reproduced results of different linear models for experiment 2*

| | Update | APE | Update ~ APE | Update ~ obs + pred | Update ~ APE * Noise |
|---|---|---|---|---|---|
| (Intercept) | 0.041 [0.030, 0.053] | 0.525 [0.397, 0.654] | 0.029 [0.017, 0.041] | 0.017 [0.004, 0.029] | 0.009 [-0.017, 0.035] |
| | p < 0.001 | p < 0.001 | p < 0.001 | p = 0.009 | p = 0.492 |
| APE | | | 0.023 [0.016, 0.030] | | 0.041 [0.026, 0.057] |
| | | | p < 0.001 | | p < 0.001 |
| obs | | | | 0.032 [0.024, 0.040] | |
| | | | | p < 0.001 | |
| pred | | | | -0.013 [-0.021, -0.005] | |
| | | | | p = 0.002) | |
| APE x Noise | | | | | -0.009 [-0.017, -0.002] |
| | | | | | p = 0.011 |

**Table 2c.** *Robust results of different linear models for experiment 2 as reproduced for this commentary (no data exclusions)*

| | Update | APE | Update ~ APE | Update ~ obs + pred | Update ~ APE * Noise |
|---|---|---|---|---|---|
| (Intercept) | 0.041 [0.029, 0.052] | 0.537 [0.411, 0.664] | 0.027 [0.016, 0.039] | 0.017 [0.005, 0.029] | 0.012 [-0.013, 0.037] |
| | p < 0.001 | p < 0.001 | p < 0.001 | p = 0.007 | p = 0.343 |
| APE | | | 0.025 [0.018, 0.032] | | 0.041 [0.026, 0.057] |
| | | | p < 0.001 | | p < 0.001 |
| obs | | | | 0.034 [0.026, 0.041] | |
| | | | | p < 0.001 | |
| pred | | | | -0.016 [-0.024, -0.008] | |
| | | | | p < 0.001 | |
| APE x Noise | | | | | -0.008 [-0.015, -0.001] |
| | | | | | p = 0.023 |

# Experiment 3

**Table 3a.** *Reported results of different linear models for experiment 3 in Melnikoff & Strohminger (2024)*

| | Update | APE | Update ~ APE | Update ~ obs + pred | Update ~ APE * Noise | Update ~ freeChoice |
|---|---|---|---|---|---|---|
| (Intercept) | -0.08 [-0.100, -0.06] | -0.82 [0.64, 1.01] | | | | |
| | p < 0.001 | p < 0.001 | | | | |
| APE | | | 0.03 [0.01, 0.04] | | | |
| | | | p < 0.001 | | | |
| Observed Affect | | | | 0.03 [0.015, 0.04] | | |
| | | | | p < 0.001 | | |
| Predicted Affect | | | | -0.02 [-0.03, 0.0003] | | |
| | | | | p = 0.046 | | |
| APE x Noise | | | | | 0.01 [0.002, 0.019] | |
| | | | | | p = 0.01 | |
| Freedom of Choice | | | | | | -0.003 [-0.015, 0.008] |
| | | | | | | p = 0.569 |

**Table 3b.** *Reproduced results of different linear models for experiment 3*

| | Update | APE | Update ~ APE | Update ~ obs + pred | Update ~ APE * Noise | Update ~ freeChoice |
|---|---|---|---|---|---|---|
| (Intercept) | -0.078 [-0.100, -0.055] | -0.823 [-1.008, -0.638] | -0.057 [-0.081, -0.033] | -0.046 [-0.072, -0.020] | -0.072 [-0.120, -0.024] | -0.071 [-0.101, -0.041] |
| | p < 0.001 | p < 0.001 | p < 0.001 | p < 0.001 | p = 0.004 | p < 0.001 |
| APE | | | 0.025 [0.012, 0.038] | | 0.049 [0.024, 0.074] | |
| | | | p < 0.001 | | p < 0.001) | |
| Observed Affect | | | | -0.030 [-0.044, -0.017] p < 0.001) | | |
| Predicted Affect | | | | 0.016 [0.000, 0.031] | | |
| | | | | p = 0.052 | | |
| APE x Noise | | | | | -0.011 [-0.019, -0.002] | |
| | | | | | p = 0.021 | |
| Freedom of Choice | | | | | | -0.004 [-0.015, 0.008] |
| | | | | | | p = 0.516 |

**Table 3c.** *Robust results of different linear models for experiment 3*

| | Update | APE | Update ~ APE | Update ~ obs + pred | Update ~ APE * Noise | Update ~ freeChoice |
|---|---|---|---|---|---|---|
| (Intercept) | -0.030 [-0.047, -0.013] | -0.632 [-0.771, -0.493] | -0.018 [-0.036, 0.000] | -0.019 [-0.041, 0.003] | -0.024 [-0.061, 0.014] | -0.057 [-0.086, -0.028] |
| | p < 0.001 | p < 0.001 | p = 0.049 | p = 0.092 | p = 0.216 | p < 0.001 |
| APE | | | 0.019 [0.009, 0.029] | | 0.036 [0.017, 0.055] | |
| | | | p < 0.001 | | p < 0.001 | |
| Observed Affect | | | | -0.019 [-0.030, -0.008] | | |
| | | | | p < 0.001 | | |
| Predicted Affect | | | | 0.019 [0.008, 0.031] | | |
| | | | | p = 0.001 | | |
| APE x Noise | | | | | -0.007 [-0.013, 0.000] | |
| | | | | | p = 0.037 | |
| Freedom of Choice | | | | | | 0.012 [0.001, 0.022] |
| | | | | | | p = 0.027 |