October 27, 2024

Dear Replication Team,

We thank you for your careful replication attempt of our work. We also thank you for the creation and support of the Transparent Replications project, which represents an important contribution to the field.

We are obviously disappointed by the failure to replicate our initial findings. Generally-speaking, we agree with the Team's interpretation that our initial experimental design could be modified to increase sensitivity to the prevalence-induced concept change effect. Namely, future work should endeavour to more equally divide the range of bodies presented to participants to maximize the number of critical "ambiguous" judgements—those most likely to show the prevalence-induced effect.

As pointed out in the blog post, our initial assumption when computing a-priori power was that the neutral stimulus (BMI = 19.79) would be viewed as maximally ambiguous and thus judgements of this stimulus would be most sensitive to our hypothesized effect. After collecting the data however, we realized that this was not the case, and that the prevalence effect was most pronounced at slightly higher model BMIs. This finding, in retrospect, suggests that our initial power estimate was probably underestimated. Because the sample size in this replication attempt was based on that prior power analysis, we agree with the Replication Team's assessment that their replication may be underpowered and that their findings "may not constitute substantial evidence against the hypothesis itself".

With this in mind, it is worth noting that there are many ways to define a replication. On the one hand, the team's findings do not constitute a statistical replication of our prior results: the key three-way interaction effect was not statistically significant. This is not too surprising, given the issues with sensitivity and potentially with statistical power pointed out above and by the Replication Team. On the other hand, Figure 5a—showing the prevalence-induced concept change effect in Devine et al. (2022)—and Figure 7 in the Replication Report—showing the same result in the replication—are remarkably similar. Indeed, the shape of the functions in both Figures are nearly identical, barring increased noisy responses for ambiguous stimuli from participants assigned to the Stable condition. Figure 10 reinforces this similarity for the most important observations, showing that the pattern of effects in the replication study and Devine et al. (2022) are very similar in both magnitude and direction, even though they were undersampled. From our perspective, it seems unlikely that these response patterns would match so closely if this effect were truly null (i.e., no difference in responses between conditions).

How to reconcile this apparent qualitative reproduction of our key results with a statistically null finding? When power is a concern, p-values are not very informative. Instead, a quantification of the strength of evidence for (or against) the null hypothesis—in this case that the coefficient for the three-way interaction between condition, trial, and model size is zero—is desirable. To accomplish this, we can turn to Bayesian statistics.

Reproducing the multilevel logistic model reported in Devine et al. (2022) and in this replication in a Bayesian framework (see Code Snippet R1 below), and using uninformative priors (we will return to this shortly), we find strong support that the three-way interaction is not zero (BF = 11.72) and moderate support that it is greater than zero (i.e., in the hypothesized direction; BF = 6.59) (Lee & Wagenmakers, 2014). Put another way, there is an 87% chance that the key three-way interaction encoding the prevalence-induced concept change effect is positive (as hypothesized) and only a 13% chance that it is zero or below (contrary to our hypothesis). If informative priors are used, taking the results of Devine et al. (2022) as prior expectations for parameter values (see Code Snippet R2)[1], the likelihood for the key interaction being greater than zero increases to 100%, which constitutes strong evidence for the hypothesis (BF > 1000).

So, while the current replication attempt failed to find a statistically significant three-way interaction effect in a frequentist framework, reanalysis in a Bayesian framework is suggestive of a prevalence effect in young women's body judgements. This Bayesian reanalysis is appropriate given concerns about statistical power. Indeed, even if power was not a concern, such analyses would be needed to quantify the strength of evidence in favour of the null hypothesis and provide more clarity on the nature of the effect. In this case, the Bayesian reanalysis provides moderate-to-strong evidence in favour of the hypothesis, depending on how priors are specified. Accordingly, we believe the primary issue in this replication is one of statistical power for the frequentist approach to significance. Despite our initial beliefs—and the Replication Team's careful preparation—the current replication study may be underpowered to detect the key interaction of interest.

The important question is: Do changes in the prevalence of thin bodies **actually** bias young women's judgements about body size? Another way to address this question (in addition to the Bayesian analysis reported above) is to consider all the available data on the question so far. As the current replication project is a direct replication of our initial dataset, doing so is straightforward: data can be pooled across the original sample (Devine et al., 2022) and the current replication. It is appropriate to do so, because the samples are comparable (the same exclusion criteria were applied). By pooling data across studies and accounting for the study of origin using a nested-random effects multilevel model (see Code Snippet R3), we find that the key three-way interaction emerges as statistically significant (b = 2.49, CI = [1.93, 3.06], p < .0001; Table R1). Figure R1 visualizes this effect using the pooled dataset. Unsurprisingly, the pattern is similar to that found in both Devine et al. (2022) and the replication (though the error bars are narrowed owing to a larger, pooled, sample size). In line with Figures 7, 10, and our Bayesian reanalysis, this pooled analysis suggests to us again that the current replication may have been underpowered to detect this three-way

[1] Incorporating prior information into Bayesian analyses is a well-known technique for mitigating concerns around sample sizes. We think this is appropriate here given that 1) more data was available in Devine et al. (2022), making it the best available estimate for the population effect size, and 2) the Team conducted a direct replication using the same procedure and recruited a comparable sample, so parameter estimates should be somewhat similar.

interaction and when more data is available, the effect emerges as statistically significant in a frequentist framework.

Overall then, we thank the Replication Team for their hard work, diligence, and transparency. We also thank them for their good suggestions for modifying the experimental design so as the lower the burden on data collection by increasing the task's sensitivity. Overall, we believe the current replication results provide weak-to-moderate evidence for a prevalence-induced concept change effect in young women's judgements about body image. We welcome future replication work which either 1) modify the Bodies Task in accordance with the Team's suggestions; or 2) collect a sufficient number of participants to detect the effect under the original experimental design.

Best,

Sean Devine
Nathalie Germain
Ben Eppinger

References

Devine, S., Germain, N., Ehrlich, S., & Eppinger, B. (2022). Changes in the Prevalence of Thin Bodies Bias Young Women's Judgments About Body Size. *Psychological Science*, *33*(8), 1212–1225. https://doi.org/10.1177/09567976221082941

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Code Snippet R1. Bayesian logistic multilevel model R code using uninformative priors.

```
library(brms)
library(bayestestR)

## Start with uninformative priors for fixed effects (needed to compute BF for parameters)
## In log-odds space, this is quite diffuse
Prior = c(
  set_prior("normal(0, 10)", class = "b")
)

m6_bayes_noprior = brm(key_press ~ conditionc*trial0*size0c + (trial0|subject),
                family=bernoulli(link = "logit"),
                chains = 3,
                iter = 5000,
                warmup = 1000,
                cores=3,
                data=d,
                prior = Prior,
                sample_prior = T,
                seed=2024,
                save_pars = save_pars(all = TRUE)
                )

## Hypothesis testing
h0   = hypothesis(m6_bayes_noprior, hypothesis = 'conditionc:trial0:size0c = 0')
hgt0 = hypothesis(m6_bayes_noprior, hypothesis = 'conditionc:trial0:size0c > 0')
```

Code Snippet R2. Bayesian logistic multilevel model R code using informative priors taken from Devine et al., 2022.

```
## Priors from Devine et al., 2022
## SDs correspond to SEs from frequentist MLM in that paper
Prior = c(
  set_prior("normal(-1.90, 0.06)", class = "Intercept"),
  set_prior("normal(0.08,  0.06)", coef = "conditionc"),
  set_prior("normal(-0.62, 0.08)", coef = "trial0"),
  set_prior("normal(21.21, 0.20)", coef = "size0c"),
  set_prior("normal(-0.65, 0.08)", coef = "conditionc:trial0"),
  set_prior("normal(-0.48, 0.19)", coef = "conditionc:size0c"),
  set_prior("normal(2.05,  0.41)", coef = "trial0:size0c"),
  set_prior("normal(3.85,  0.38)", coef = "conditionc:trial0:size0c")
)

m6_bayes_prior = brm(key_press ~ conditionc*trial0*size0c + (trial0|subject),
                family=bernoulli(link = "logit"),
                chains = 3,
                iter = 5000,
                warmup = 1000,
                cores=3,
                data=d,
                prior = Prior,
                sample_prior = T,
                seed=2024,
                save_pars = save_pars(all = TRUE)
                )

## Hypothesis testing
h0   = hypothesis(m6_bayes_prior, hypothesis = 'conditionc:trial0:size0c = 0')
hgt0 = hypothesis(m6_bayes_prior, hypothesis = 'conditionc:trial0:size0c > 0')
```

Code Snippet R3. R code for pooled frequentist analysis, combining data from Devine et al., 2022 and this replication attempt.

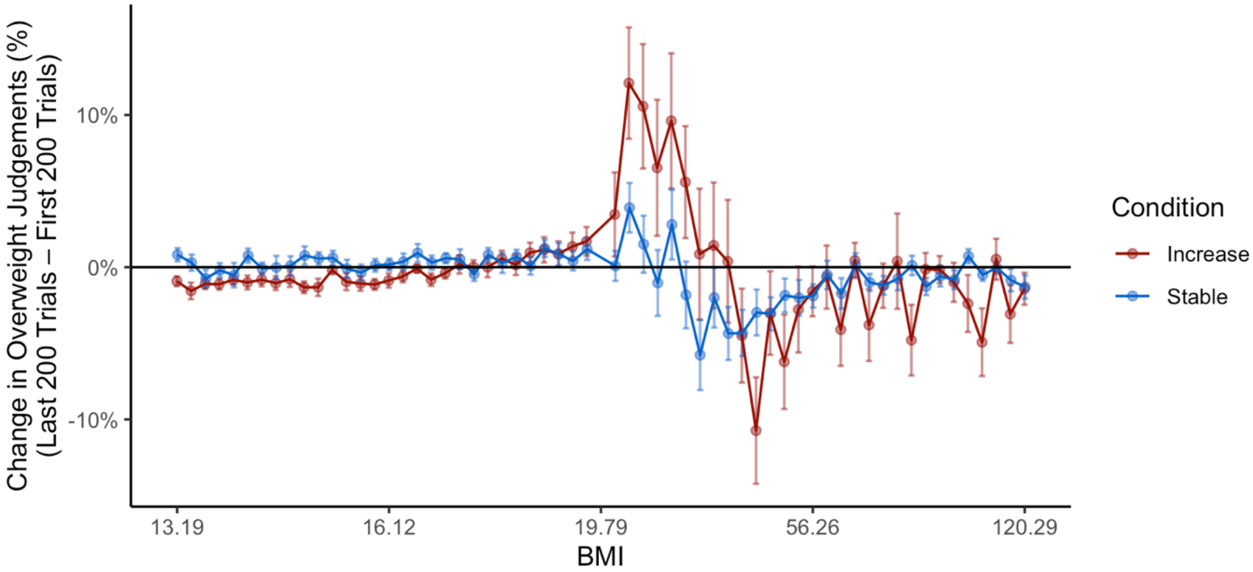
```
library(lme4)

pooled_mod =
 glmer(key_press ~ conditionc*trial0*size0c + (trial0|study/subject),
      family='binomial', glmerControl(optimizer = 'bobyqa'),
      data=d)
```

Table R1. Fixed-effects output from the pooled multilevel model analysis.

Predictor	Log-Odds	Std. Error	p
Intercept	-1.77464	0.06935	< 2e-16
Condition	0.10701	0.04583	0.019548
Trial	-0.68697	0.08455	4.46E-16
Size	20.82213	0.15536	< 2e-16
Condition x Trial	-0.64923	0.0603	< 2e-16
Condition x Size	-0.45827	0.14856	0.002037
Trial x Size	1.01132	0.30582	0.000943
Condition x Trial x Size	**2.49386**	**0.28897**	**< 2e-16**

Figure R1. Visual representation of changes in responses across the Bodies Task using the Pooled Dataset (*N* = 620).


Change in Overweight Judgements (%)
(Last 200 Trials – First 200 Trials)
BMI
Condition
Increase
Stable