How statistical challenges and misreadings of the literature combine to produce unreplicable science: An example from psychology*

Andrew Gelman† and Nicholas J. L. Brown‡

19 Jul 2024

Abstract

Given the well-known problems of replicability, how is it that researchers at respected institutions continue to publish and publicize studies that are fatally flawed in the sense of not providing evidence to support their strong claims? We argue that two general problems are: (a) difficulties of analyzing data with multilevel structure and (b) misinterpretation of the literature. We demonstrate with the example of a recently published claim that altering patients' subjective perception of time can have a notable effect on physical healing. We discuss ways of avoiding or at least reducing such problems, including comparing final results to simpler analyses, moving away from shot-in-the-dark phenomenological studies, and more carefully examining previous published claims. Making incorrect choices in multilevel modeling is just one way that things can go wrong, but this example also provides a window into more general problems with complicated designs, cutting-edge statistical methods, and the connections between substantive theory, experimental design, data collection, and replication.

1. Introduction

A dozen years ago, Bem (2011) published a paper claiming to find extra-sensory perception, and this kicked off awareness of a replication crisis in psychology. The experiments in question indeed failed to replicate (Ritchie et al., 2012), but the more general issue remained that this unreplicable and scientifically implausible result had appeared to be supported by rigorous experimentation and analysis (Carey, 2011).

A few years later, the episode was summarized as follows in the news media (Engber, 2017):

> Even with all that extra care, Bem would not have dared to send in such a controversial finding had he not been able to replicate the results in his lab, and replicate them again, and then replicate them five more times. His finished paper lists nine separate ministudies of ESP. Eight of those returned the same effect. . . . But for most observers, at least the mainstream ones, the paper posed a very difficult dilemma. It was both methodologically sound and logically insane.

In fact, Bem's paper contains zero actual replications. What it has could be called conceptual replications, open-ended studies that could be freely interpreted as successes through the "garden of forking paths" of data-dependent choices of data coding and analysis (Gelman & Loken, 2014). And the paper is not "methodologically sound." Its conclusions are based on p-values, which are statements regarding what the data summaries would look like, had the data come out differently, but that article offers no evidence that, had the data come out differently, the analyses would have been the same. Indeed, the nine studies in that paper feature all sorts of different data analyses.

[*] To appear in *Advances in Methods and Practices in Psychological Science*. Data and code are at `https://osf.io/rh6g5/`. We thank Pamela Davis-Kean, Susan Gelman, and two reviewers for helpful comments and Alia Crum for helpful discussions and for sharing data from Crum and Langer (2007).

[†] Department of Statistics and Department of Political Science, Columbia University, New York.

[‡] Department of Psychology, Linnaeus University, Växjö, Sweden.

What's stunning in retrospect is how (a) at the time, the Bem (2011) paper looked like standard practice, maybe nothing special but nothing horrible either; but (b) in retrospect, its problems are obvious and just jump out, once you know what to look for. It's like one of those color-vision tests the eye doctor gives you, where when you wear the 3-D glasses the images just leap off the page. We are reminded of the notorious photographic images of fairies from the early twentieth century which fooled Arthur Conan Doyle and others but to modern eyes are obvious fakes (Smith, 1997).

The problems of Bem (2011) are now clear, but publication and promotion of unreplicable research remains a problem despite the progress made in the science reform movement during the past decade.

In the present paper, we explore some general issues by examining in detail a recent psychology paper and investigating problems that might not be apparent in a casual reading but nonetheless lead to unreplicability.

1.1. Two factors leading to unreplicable research

We consider two factors that lead to overconfidence in empirical claims from noisy data.

First, psychology experiments often include multilevel structure, for example from repeated measurements, manipulations applied at the group level, or different raters. A certain amount of complexity in experimental data is typically unavoidable in psychology, given that modern research often focuses on interactions: the hypothesis of interest is how a manipulation affects a change rather than an absolute level, or how effects differ among groups, or how the effect of one variable depends on the level of another. In addition, high variation between people makes it advisable to perform within-person comparisons where possible, for reasons of substantive theory as well as statistical efficiency. But analysis of multilevel data is difficult: it is easy to get apparently strong statistical results from correlated errors, it is not always clear how to perform simple sanity checks of complex analyses, and multilevel modeling introduces its own challenges.

The second problem is misreading of the empirical literature. Results from any single experiment will be open to multiple interpretations, as no intervention occurs in a vacuum. New empirical findings are understood in the context of previous work on the topic. It is well known, although perhaps not so well understood, that published results tend to be overly optimistic about effect sizes as a result of low power and selection on statistical significance (Ioannidis, 2008). Beyond this, there are qualitative challenges in interpretation of the literature, with a mismatch between claims being made and the evidence that was used in support of those claims.

Although these problems have been discussed in general terms, they can only really be understood in the context of particular examples: each statistical analysis presents its own challenges, and each literature review has its own concerns. As a result, we fear that these quantitative and qualitative problems of interpretation of evidence are insufficiently conveyed in the statistics and methods literatures.

Our paper focuses on the data analysis and cited papers of a single work in psychology, a recently published article reporting that altering patients' subjective perception of time can have a notable effect on physical healing. Beyond revealing specific problems with that work, our investigation demonstrates the effort that can be required to track down what went wrong in a published paper. Our purpose here is not specifically to critique that one article; indeed, we could have chosen many others that exhibit very similar problems. This particular paper was chosen because it was by researchers from a respected institution, published in a journal that is generally considered legitimate, and demonstrates three key factors:

1. The published paper reports large estimated effects from a small study with no clear theoret-

ical justification, the sort of finding that is characteristic of the wave of unreplicable results in psychology, as discussed, for example, by Bishop (2020b).

2. On the other hand, the results appear at first glance to be unambiguously statistically significant and based on a solid experimental design.

3. In addition, the published article refers to a substantial literature reporting similar findings, thus potentially reducing the concern about the lack of clear mechanism of action of the treatment.

All three of these attributes are relevant. Without the first attribute—implausibly large effects—this would just be unexceptional science. Without the second attribute—apparent statistical significance—the results could be dismissed as an artifact of noisy data. And without the third attribute—connection to an existing literature—the results would not lead to any clear scientific interpretation. In the present paper, we investigate these factors in the context of a single published research article and explain why its apparent statistical significance is a mirage and why its cited literature does not say what is claimed.

Much has been written on the replication crisis in psychology, including methodological studies, recommendations for changes in research and publication practices, organized replication studies, and surveys of the literature. Here we present a detailed examination of a particular claim, a case study which we see as complementary to the broader takes on replication concerns. We have many times seen illusory statistical significance and inaccurate literature reviews, and it typically takes a bit of digging to track down these problems in each case. By carefully going through these steps in the context of a high-profile example, we can get a sense of how a result can misleadingly appear to be well-founded both empirically and theoretically.

1.2. A questionable finding in psychology

A recent article (Aungle & Langer, 2023) reported an experiment that "tested whether cupping marks produced by identical cupping treatments healed faster or slower as a function of perceived time." Cupping "involves creating a localized suction on the skin . . . [leading to] bruising." Each of 33 participants were given this treatment three times; in each case, the participant was given a 28-minute-long task and in each instance a photograph of the skin was taken before and after the 28-minute interval. The three instances differed in that the experimenters manipulated "perceived time" of the recovery interval, telling the participant it was actually 14 minutes, 28 minutes, or 56 minutes. The following results were reported:

> Healing in the 14-min condition had a mean rating of 6.17 (SD = 2.59, 32 Subjects, 800 ratings); healing in the 28-min condition had a mean rating of 6.43 (SD = 2.54, 33 Subjects, 825 ratings); and healing in the 56-min condition had a mean rating of 7.30 (SD = 2.25, 32 Subjects, 800 ratings).

Healing was measured by 25 external raters who were given the before and after photographs and asked to rate on a scale: "0.0 = not at all healed, 5.0 = somewhat healed, 10.0 = completely healed." For each of three comparisons (56 min vs. 28 min condition, 56 min vs. 14 min, 28 min vs. 14 min), the t score (estimate divided by standard error) was calculated: the resulting values were reported as 7.2, 10.7, and 2.5.

Based on our experiences with small-sample studies, these results did not seem plausible. An indirect intervention given to 32 or 33 people, yielding a t statistic of 7.2? A t statistic as high as

7 in such a setting would typically only occur for a manipulation check, not for the main finding of a study of a speculative effect; it is a red flag leading us immediately to question the data analysis.

We were also concerned about the larger claims made in the article's summary: "we show that the effect of time on physical healing is significantly influenced by the psychological experience of time . . . Our results demonstrate that the effect of time on physical healing is inseparable from the psychological experience of time." After a careful look at the statistical analysis, we find that the data do not provide strong evidence of the claimed effect, and even if the statistical results in the paper had been correct, this would not demonstrate the inseparability claimed in that statement.

We report our efforts to understand where the claimed results came from, followed by our reanalysis of the data and our consequent understanding of the study and the literature to which it refers. In doing this, we came across a challenge in using multilevel modeling to account for experimental design. We hope this work is useful for future researchers who are analyzing multilevel data structures, as well as for people who are trying to interpret the existing literature.

2. Reanalysis and reassessment

Even when the substantive theory underlying a flawed research project is speculative or implausible, it can be helpful to reanalyze the data to better understand how the results of a noisy experiment could have been arranged in a way that convinced authors and reviewers alike that they were seeing strong evidence. In the case of Aungle and Langer (2023), challenges arose when analyzing a multilevel data structure.

2.1. Published analysis: multilevel model with varying intercepts

The data and code for the healing experiment are linked from the journal article's webpage, so we can check the authors' analyses and conduct our own as well. It also makes sense to check things with a simpler model, which we do in Section 2.3 by collapsing the data and comparing averages. We start by examining the analysis that appeared in the published paper.

One recommended practice when analyzing data with clustering is to fit a multilevel model (Snijders & Bosker, 1999). In this case, the data are clustered by participant (coded as `Subject` in the dataset) and rater (coded as `ResponseId`). For their Table 1, the authors choose the best fit among several models. Here we show the simplest, as fit in R:

```
lmer(formula = Healing ~ Condition + (1 | Subject) + (1 | ResponseId),
    data = DFmodel, REML = FALSE)
            coef.est coef.se
(Intercept) 6.20     0.31
Condition28 0.23     0.09
Condition56 1.05     0.09

Error terms:
 Groups     Name        Std.Dev.
 Subject    (Intercept) 1.07
 ResponseId (Intercept) 1.22
 Residual               1.87
---
number of obs: 2425, groups: Subject, 33; ResponseId, 25
```

In this model, the 14-min condition is the baseline, and the estimates of 0.23 ± 0.09 and 1.05 ± 0.09 correspond to the effects of the 56-min condition compared to this baseline. It is also possible to extract the comparison to the 28-min condition from this analysis, but to understand what is going on here that is not really necessary, so we will just focus on the two comparisons here. The t scores of $0.23/0.09 = 2.4$ and $1.05/0.09 = 11.1$ are close to those reported in the article, with the only difference being that the published analysis included additional predictors for participant and session characteristics. Including these additional predictors induced only very small changes in the estimates and standard errors for the treatment effects, and so for simplicity we do not consider them further.

Before going on, we shall interpret the error terms in the above fitted model. Based on the fitted model, the measurements vary with a standard deviation of 1.07 across participants and 1.22 across raters, and the unexplained or residual error has standard error 1.87, all relative to the ten-point scale of measurement. This all makes sense: some participants' bruises will look more serious than others', different raters use different subjective scales, and this will vary across measurements.

2.2. Multilevel model with varying intercepts and slopes

When estimating a treatment effect under a cluster design, it is not enough to fit a multilevel model with varying intercepts. The slopes—that is, the treatment effects—must also be allowed to vary, in accordance with the general principle of the design and analysis of experiments that the error term for any comparison should be at the level of analysis of the comparison; see, for example, Cochran and Cox (1957) and Barr et al. (2013). In the terminology of the analysis of variance, the treatment is applied between groups (subjects and raters), and so the estimated effect must be compared to a between-group variance. This can be done by slightly extending the fitted multilevel model to allow the treatment effects to vary by subjects and raters:

```
lmer(formula = Healing ~ Condition + (1 + Condition | Subject) +
    (1 + Condition | ResponseId), data = DFmodel, REML = FALSE)
            coef.est coef.se
(Intercept) 6.18     0.39
Condition28 0.25     0.36
Condition56 1.09     0.37

Error terms:
 Groups     Name        Std.Dev. Corr
 Subject    (Intercept) 1.71
            Condition28 1.99     -0.71
            Condition56 2.03     -0.72  0.65
 ResponseId (Intercept) 1.24
            Condition28 0.07      1.00
            Condition56 0.13     -1.00 -1.00
 Residual               1.51
---
number of obs: 2425, groups: Subject, 33; ResponseId, 25
```

The estimated average treatment effects are similar to before, but the standard errors are much bigger. The fitted model estimates a high variation of effects across participants: the effect of 28 min (compared to the baseline of 14 min) is estimated to have a standard deviation of 1.99, and the effect of 56 min is estimated to have a standard deviation of 2.03. Those standard deviations are much higher than the estimated mean effects of 0.25 and 1.09, respectively; thus, according to the fitted model, the estimated effects are not consistent in their sign or their magnitude. In this case, the most important step was to allow treatment effects to vary by subjects. The variation of effects by raters is small, but it did not hurt to include that in the model. It would also be possible for the variation as well as the mean to vary by group, and various alterations of the model will slightly change the estimated effects and uncertainties. Our point here is not to offer a definitive analysis of these data but rather to understand how the published results had been so inappropriately strong.

An additional concern arises from the high estimated correlations of the varying slopes; indeed, the estimated correlation matrix for the varying intercepts and slopes for raters is not positive definite, which results in a warning message when the model is fit in R. Given the small number of groups, this sort of degeneracy in the maximum likelihood estimate of the covariance is not unexpected; see Chung et al. (2014). We checked our result by running a fully Bayesian analysis, which accounts for uncertainty in the estimation of these variance components. In this case the result was essentially the same and so we stick with the analysis shown above. Including other predictors into the model also left the estimates and standard errors of the average treatment effects essentially unchanged.

To summarize: the fitted model shows evidence for an average effect of the 56 min compared to the 14 min condition, but not of the 28 min compared to the 14 min condition. Both effects are estimated to vary by a large amount across participants, implying that both the sign and the magnitude of the effects are highly variable. There is essentially zero variation in treatment effects across raters, which makes sense, given that the raters are reacting to the images given to them and are not otherwise affected by the treatments.

Referring to the estimated intraclass correlation in their fitted model, the authors write, "A lower ICC value suggests that there was less variability in healing outcomes between subjects, indicating that the condition effect was relatively consistent across subjects." This claim is in error: the model that they fit assumes constant treatment effects and thus offers no information at all regarding consistency of effects across subjects.

2.3. Simple paired-comparisons analysis

One frustrating aspect of this problem is that the statistical issue is clear—we want to obtain estimates and uncertainties of treatment effects in the presence of clustering—but textbook recommendations for analysis can be hard to follow. The multilevel analysis performed by the authors looks superficially reasonable but is missing the all-important variation in treatment effects. We solved this problem by including varying slopes, but this leaves a lingering suspicion that some additional analysis step might still be missing.

One way to get a handle on the problem is to perform a simpler analysis. To start with, we will consider each of the comparisons (56 min vs. 28 min, 56 min vs. 14 min, and 28 min vs. 14 min) as its own problem, thus avoiding the difficulties arising from analyzing an experiment with three treatment levels. Next we simplify further by working with the mean of the 25 measurements for each person and each condition. This leaves us with a simple matched-pair design for each of the three comparisons, which we can then estimate in the usual way by computing the difference in outcome between the two conditions for each person and then summarizing by the mean of these differences and their standard deviation divided by $\sqrt{n}$, where n is the number of people who participated in both conditions. The resulting estimates ± standard errors for the three comparisons are 0.79 ± 0.31, 1.10 ± 0.38, and 0.20 ± 0.36, respectively. These estimates and uncertainties are close to those from the multilevel model but not identical because they do not account for rater effects.

This simple analysis is not intended to be an alternative to the multilevel model; it is just a way to compare it to something more easily understandable. In this case we see no clear alternative to fitting a multilevel model with varying intercepts and slopes or using a procedure such as clustered standard errors, which would have its own theoretical and practical complications (Abadie et al., 2023).

2.4. Re-evaluating the published claims

How do we think about the claims of Aungle and Langer (2023), now that its t scores have been downgraded from 10.7 to 3.0 (for the 56 min vs. 14 min comparison) and from 2.5 to 0.7 (for the 28 min vs. 14 min comparison)? As noted above, in addition to these possible average effects, any effect of the manipulation on healing is estimated to be highly variable, sometimes positive and sometimes negative.

We are skeptical that this study reveals anything about the effect of perceived time on physical healing, for four reasons.

First, the statistically-significant result that appeared is one of many comparisons that could have been made. Data were also gathered on participants' anxiety, stress, depression, mindfulness, mood, and personality traits, implying many possible analyses that could have been performed. In the absence of preregistration, there is just no way of knowing what might have been done had the data turned out differently, and the result is that the appearance of a comparison that is 3 standard errors away from zero does not necessarily represent strong evidence of an effect (Simmons et al., 2011). To the extent that perceived time could affect healing, it would be easy to come up with hypotheses why such effects would only occur for patients with high or low levels of anxiety or stress, different levels of mindfulness, or under only some conditions of mood or for only some sorts of personality profiles, as any of these could be related to "specific networks of expectations, physiological responses, and beliefs associated with participants' concepts of time," which is one of several speculative explanations offered by the authors for their findings.

Second, the large estimated variation in effect size across people implies that any estimated average effect will be highly contingent on who happens to be in the study, and there is no reason to believe that the particular 33 people in the experiment are representative of any larger population of interest. This issue would always arise when attempting to use data from a lab experiment to generalize to the outside world, but it would have been less of a concern if the effect estimate had really had been 10 standard errors away from zero, as that would imply a consistency in the effect that would make generalization easier to swallow.

A third reason for skepticism is that any effect would be expected to vary not just across people but across situations, leading to the same concern about interactions and stability. The article refers to "healing," but the experiment did not involve the participants experiencing any sickness or injury beyond very mild bruising. Indeed, the technique of cupping is often promoted by proponents of alternative medical treatments as having healing properties of its own; on this account, what the patients experienced could have been explained as an interaction between psychological factors and the purported mechanism of action of cupping itself (which, to the best of our knowledge, remains unidentified; Singh & Ernst, 2008, p. 307).

A fourth concern is with the mechanism of action, as it is not clear how the perception of time would affect changes in the skin in this setting. The authors refer to "mind–body unity" and "the importance of psychological factors in all aspects of health and wellbeing," and we would not want to rule out the possibility of such an effect, but no mechanisms are examined in this study, so the result seems at best speculative, even taking the data summaries at face value. During the half hour of the experimental conditions, the participants were performing various activities on the computer that could affect blood flow, and these activities were different in each condition (watching videos under one condition, playing Tetris in another, playing a different video game in the third). In addition there is no mention that the experimenter, who took the photos, was blind to the condition. There were many things going on in the experiment, and it seems to us to be a strong claim to attribute the observed differences to "perceived time" rather than to any of the other factors that were varying across the three conditions, or even just to the general ability of researchers conducting uncontrolled studies to find patterns from noise.

In raising these concerns, we are not saying that the substantive conclusions of the study are necessarily wrong, just that there are many alternative explanations for the results which we find just as scientifically plausible as the published claim that "the effect of time on physical healing is significantly influenced by the psychological experience of time."

3. Problems with citation of the previous literature

Unreplicable claims based on weak theory can gain apparent support by connections to related published work. Three problems can arise.

First, the connections between the cited literature and the new study can be tenuous, and this can particularly be an issue when the underlying theory is vague. Ideas such as embodied cognition, evolutionary psychology, nudging, mindfulness, or mind-body unity are general enough to encompass a wide range of potential phenomena, to the extent that there is almost no limit to the past studies that could be thought to have some possible relevance to any new experiment.

Second, informal literature reviews are subject to selection bias. An article promoting a controversial idea can easily cite studies claiming to have found evidence for related ideas, while avoiding citations of failed replications or papers suggesting alternative theories. This can even be a problem with systematic meta-analyses, if the entire subfield being meta-analyzed is full of studies with uncontrolled researcher degrees of freedom (Gelman, 2022).

Third, the interpretation of individual studies being cited can be seriously flawed. This is a problem of citing past literature as support for a general claim without looking at exactly what was done in the cited research and without following up on that work. Here we discuss three different examples of this sort of misinterpretation of the literature cited in the paper under discussion.

3.1. Doctors' assurance and allergic reactions

Aungle and Langer (2023) provide context for their study by citing related findings on "surprising mind-body unity phenomena." One of these came from Leibowitz et al. (2018), which reports that patients who received a histamine skin prick reported less itchiness if they were assured by their healthcare provider, "From this point forward your allergic reaction will start to diminish, and your rash and irritation will go away." However, the statistically-significant ($z = 1.96$, $p = 0.05$) result in this paper represents only the tip of the iceberg of possible analyses that could have been performed for this experiment. The participants also appear to have completed a fairly extensive survey, whose results are included in the data file accompanying the article. Several participants took many hours to complete this, which suggests that it was not done on a computer in the lab or wherever the skin prick was performed. The survey seems to have included a lot of questions about people's feelings about the procedure. Most of these responses seem not to have been used in the article, and indeed the entire survey is not mentioned, but six of the responses made it into a posted shorter version of the dataset. These appear to be the actual itching and the participant's expectation of something (how much it would itch, perhaps?) at baseline, 3 minutes, and 9 minutes. But if this survey was indeed completed retrospectively, this would suggest that the measurement of itchiness perceived by the participant was also done retrospectively, and that would not seem to be very reliable, especially for six time points. There are also some inconsistencies between the online data files and various other numbers that do not quite line up. We conclude from all this that the study had many researcher degrees of freedom, and we have no strong reason to believe that it would hold up under replication, especially given that, in citing this result, Aungle and Langer (2023) refer to it as "surprising." When an experiment is connected with many possible data selection and analysis paths, it is easy to obtain statistical significance even in the absence of any underlying effect, a concern that is magnified given the uncertainties of how the data were collected.

Leibowitz et al. (2018) conclude: "We suspect the present study is a conservative test of this effect since participants were healthy volunteers whose allergic reactions were unlikely to be highly stressful or concerning, and allergic reactions were expected to decline over time even without intervention." This is highly speculative, and indeed it would be easy to make a case for the exact opposite. If you want to make this claim—or its opposite—, it would make sense to go and test some more difficult patients.

If a one-sentence reassurance could reliably reduce short-term pain, this could have immediate implications in health care practice, and so it would seem advisable for someone who believes in this result to conduct a careful replication study. A Google Scholar search conducted five years after the study's publication showed 28 citations, none of which replicated the original experiment. The closest empirical study we found in these references was Leibowitz et al. (2019), which was also cited by Aungle and Langer as one of the "surprising mind-body unity phenomena," and which studied "various mechanisms of open-label placebo treatments: a supportive patient-provider relationship, a medical ritual, positive expectations, and a rationale about the power of placebos," looking at outcomes on allergic responses but not on itchiness. That study reported "no main effects of condition on allergic responses" but statistical significance on some particular interactions. Again, given the many possible interactions that could be studied, this does not represent strong evidence for the replicability of whatever happened to show up in this particular sample.

3.2. "Countless studies, many of which stand up to replication and rigorous scrutiny"

Aungle and Langer (2023) write, "previous research has found the skin is quite responsive to expectations. For example, patients who received physician assurances after skin pricks healed significantly faster, and the suggestion that one had touched poison ivy resulted in stronger symptoms than actually touching poison ivy." The first of these claims refers to the aforementioned Leibowitz et al. (2018) study; the second points to Ikemi and Nagakawa (1962), which includes a qualitative study of 13 boys who were exposed on one arm to the poisonous leaves of the lacquer or wax tree (not actually poison ivy) and on the other arm to inert leaves, but were told that the exposures were the reverse. Under various conditions, skin reactions occurred on the arm where the boys were told the poisonous leaves had been applied, rather than at the actual location of the contact. A later review article (W. A. Brown, 2015) describes that study as "as baffling today as it was when it first appeared" and continues, "this contact dermatitis study has not been replicated so it's hard to know just how solid its remarkable findings may be. Nevertheless this study does not stand alone. . . . Countless studies, many of which stand up to replication and rigorous scrutiny, show that the power of expectation is as dramatic and perplexing as it was in the poison leaf study."

The phrase, "countless studies, many of which stand up to replication and rigorous scrutiny," is interesting in that it asks the reader to consider as evidence some large number of studies (the difference between "countless" and "many") that do *not* stand up to replication and rigorous scrutiny. In this case, Aungle and Langer (2023) cite small uncontrolled studies that have not been replicated, which suggests to us that the many studies that stand up to replication and rigorous scrutiny are either hard to find or else not directly relevant to the claims being made in their paper.

3.3. Exercise beliefs and biometric outcomes

Aungle and Langer (2023) also report the following claim, which would be stunning if it held up under replication:

> "If a person who does not exercise weighed themselves, checked their blood pressure, took careful body measurements, wrote everything down, maintained their same diet and level of physical activity, and then repeated the same measures a month later, few would expect exercise-like improvements. But in a study involving hotel housekeepers, that is effectively what the researchers found."

After a careful study of the reference (Crum & Langer, 2007), we are again skeptical. The treatment in this experiment was to inform the hotel housekeepers (in this study, 84 women working at 7 hotels) "that the work they do (cleaning hotel rooms) is good exercise and satisfies the Surgeon General's recommendations for an active lifestyle."

We have two reasons to doubt the above-quoted summary.

First, the reported changes seem implausibly large for population effects, with the women receiving the brief intervention seeing an average drop of two pounds of weight, half a percentage point of body fat, and five or ten points of blood pressure a month later, compared to the women in the control group. We would be inclined to attribute such large apparent effects to chance variation in the data. However, for many of the outcomes being studied these differences are two or three standard errors from zero, which would seem to be unlikely to occur by chance alone.

Some of the apparent strength of the statistical patterns arose from clustering in the design that was not accounted for in the analysis, similar to the problem with the multilevel analysis discussed in Section 2. In the 2007 paper, the intervention was applied at the hotel level, with workers at four hotels receiving the treatment and at the other three receiving the control, but the published analysis does not appear to have accounted for this clustering. A more appropriate analysis would use a multilevel model with intercept and treatment effect varying by group, as in Section 2.2 but with groups being hotel rather than participant in this case. This correction for clustering reduces the t statistics for the changes in biometric outcomes, but some of them still remain above the conventional level of statistical significance.

There is also a concern about possibility of systematic error in body measurements if the experimenter was not blinded to the treatment. In addition, there were many missing observations and some people in the dataset whose BMI data were not consistent with their recorded heights and weights. Flexibility in data coding, measurement, and analysis could suffice to explain the observed patterns in the data.

Stepping back, it is a stretch to expect that a presentation on how work is good exercise would result in major changes, especially for a group of people who have "maintained their same diet and level of physical activity." We would expect that a one-shot study of 84 people would be too noisy to discover any plausible effect after four weeks—a period that is not only short, but also arbitrary in the absence of any theoretical account of how the intervention might work without inducing change in diet and exercise. Even if such an effect exists, we would not expect it to work or to go in the same direction for everyone, and any average effect should be small. As explained by Button et al. (2013), when a noisy experiment is performed to study a small effect, any statistically significant result will tend to greatly overestimate the true underlying effect; this is called the winner's curse or statistical significance filter or type M (magnitude) error (Gelman & Carlin, 2014).

Our second concern with the above quote is the claim that the women in the study "maintained their same diet and level of physical activity." Crum and Langer (2007) indeed state that "actual behavior did not change," but that article does not report any direct measures of diet and physical activity at either the start or end of the study, just information from a retrospective questionnaire. It is problematic to take survey responses as measures of actual behavior, especially in the context of a study of an intervention specifically designed to alter perceptions of exercise. The interpretation given by Aungle and Langer (2023) requires that the words given to the participants at the beginning of the study could affect measures that are the result of physiological processes such as weight, body-fat percentage, and blood pressure, without having any effect on survey responses on exercise and diet.

Beyond this, the data in the study actually *do* show a large increase in perceived amount of exercise (the average going from 3.8 to 5.7 on a 0–10 scale), so if the survey responses are to be believed, this directly contradicts the claim that the participants "maintained their same diet and level of physical activity." The paper also reports that "there were no significant changes in subjects' substance abuse and diet." However, with such a small sample, a lack of statistical significance does not imply that real changes were zero or even that they were small. All of this is in addition to the differences between actual diet and retrospective self-reports.

In short, to the extent that the intervention caused the physiological changes measured in the study, the data are consistent with the reasonable hypothesis that these were associated with behavioral changes, and so we do not think it makes sense for Aungle and Langer (2023) to cite this study as evidence for the claim that "the benefits of exercise" do not "require the act of exercising, or at the very least an increase in physical activity or change in diet." To make such a conclusion requires, first, the statistical error of treating a non-statistically-significant result as zero ("no change in workload, exercise habits, overall physical activity, or diet"), and, second, discounting the actual survey responses in which participants reported exercising more in that study.

4. Discussion

A naive reader of many discussions of the replication crisis in science might gain the impression that all would be well if scientists were merely to follow open-science protocols and avoid certain questionable research practices. The problems go deeper, however: difficulties of statistical analysis of real data, along with misinterpretations of the scientific literature, two issues that arise more generally in unreplicable subfields of research. We hope that our careful exploration of these issues in a particular example gives insight into a problem that goes far beyond the literature in mind-body unity.

Also, to the extent that there is general interest in the claim that physical healing can be affected by manipulating the psychological experience of time, or that a one-sentence reassurance could reliably reduce short-term pain, or that it is possible to gain the benefits of exercise while maintaining the same diet and level of physical activity—and, given publication, citation, media attention to such claims, they do seem to be of general interest—it should also be of interest to learn that the evidence for such claims is not nearly as strong as has been presented in the literature. And it is worth understanding the specifics of how a paper published by researchers at a respected university could go so wrong.

4.1. Challenges of accounting for complex designs

As illustrated in Section 2, data analysis in the real world can be difficult. Aungle and Langer (2023) followed general recommendations to use multilevel modeling when analyzing clustered data, but even so they got tripped up and didn't realize they needed to allow the treatment effects, not just the intercept, to vary by participants. And it seems that none of the reviewers of the paper caught this error either. Indeed, we only noticed it because we were tipped off by the unrealistically-high t scores and then were able to download and interpret the authors' code.

The most natural advice at this point would be to say that, when estimating a causal effect (or, more generally, a regression coefficient) from data with a multilevel structure, it is necessary to allow both the intercepts and effects to vary along each grouping structure. In practice, though, this is a challenge, first because this particular issue is not always covered in textbooks; second because including additional variance components to a model can make it less stable to fit, especially when the number of measurements and people in the study is small; and third because modeling choices will still arise, for example which other predictors to interact with the treatment indicator and how to code a treatment with multiple levels.

Similar problems arise with the general recommendation in observational studies to adjust for all relevant pre-treatment variables. The number of possible adjustment variables can be large, in which case researchers will need to choose what predictors to include and how to parameterize them, and it can be necessary to go beyond simple least-squares adjustment.

A different direction would be to obtain standard errors by bootstrapping, with the resampling respecting the design of the study. Such an approach can work well but again will require care in getting it to work in problems with non-nested structures such as in the Aungle and Langer (2023) study. Our message here is not that complex designs cannot be analyzed or should not be conducted—indeed, one of us is on record as recommending within-person comparisons in psychology experiments (Gelman, 2018). Rather, this is just a reminder that statistical analysis of such data is far from routine, even for randomized experiments.

Applied researchers remain in the awkward position of needing to use statistical methods without clear guidance and whose results can be difficult to understand. Ultimately, this reflects a fundamental issue that, in all but the simplest designs, uncertainties in estimation depend on aspects of treatment-effect variation that themselves must be estimated from the data. This represents a challenge not just for practitioners but also for software developers and authors of textbooks and expository articles: it is impossible to give advice that is general, easy to follow and understand, and correct. As the saying goes, choose at most two.

4.2. Misinterpretation of the literature and relevance to the replication crisis

The article under discussion, Aungle and Langer (2023), follows the pattern of much of the unreplicable research we have seen in psychology: a study of a highly speculative claim, with results whose apparent statistical significance fades upon careful analysis. The problems were not apparent to casual reviewers, and the paper was published in a legitimate journal and received some uncritical publicity (Carroll; 2024; DeSmith, 2024; Langer and Aungle, 2024; Levitt, 2024; Peterson, 2023; Plain, 2024; Rand, 2023). Beyond the specific problems with the statistical analysis, the paper featured a claim ("the effect of time on physical healing is inseparable from the psychological experience of time") that was not directly addressed by the experiment and thus could not have been supported, even if one were to take the reported quantitative summaries at face value. Other work in this literature has been similarly criticized on both methodological and theoretical grounds, and these criticisms are not new (Liberman, 2009; Coyne, 2014).

One recurring feature in the replication crisis is a style of writing and presentation which can give a misleading appearance of coherence to a diverse literature. Conceptual replications are valuable, but they represent yet another source of researcher degrees of freedom (Simmons et al., 2011): if a conceptual replication goes in the direction that is consistent with the story that you want to tell, you can label it as a replication and use it as evidence in favor of your theory; if it yields a null result or goes in the opposite direction, you can emphasize the differences between the different studies. This relates to the points made by Bishop (2020a) regarding researchers' cognitive processes when designing experiments and interpreting their results.

Another issue we have seen before is reliance on earlier studies with flawed design and analysis. This arose in a recent meta-analysis of nudge interventions that was based on a large number of published papers that were subject to selection bias in what summaries they included, along with several papers that had been retracted or discredited because of potential fraud. Even without the inclusion of the fraudulent research, we judged that the potential for selection bias and effect-size variation in the mass of studies in the meta-analysis made its conclusions close to worthless (Szászi et al., 2022).

Aungle and Langer (2023) had similar problems, uncritically citing work that, when studied carefully, did not offer strong evidence in favor of their claims. One of these cited papers, Crum and Langer (2007), supported its argument for the plausibility of large effects by pointing to various studies of the placebo effect, including a newspaper report (Blakeslee, 1998) that referred to the unreplicated study of Ikemi and Nagakawa (1962) discussed in Section 3.2 above. This is a sort of game of telephone where various problematic or unreplicated studies get referenced in a way that can make them appear to represent a consistent literature or a web of evidence, and it arises from a disconnect between scientific procedures and scientific theories (Devezer et al., 2021). Loose theory plus loose criteria for evidence combine to allow a literature to be built upon sand, an issue discussed by Oberauer and Lewandowsky (2019).

4.3. Recommendations

In this paper we have considered several challenges:

- At the technical level, there is not always a clear pathway for researchers to analyze data from complex designs to obtain efficient inferences while avoiding overconfidence. This is an issue that will not be going away, given the increasing interest in varying treatment effects, within-person studies, and the need for more elaborate analyses to generalize from observational or experimental data to larger populations under more realistic scenarios.

- The evidence provided in any particular study can depend strongly on the average size and variation of the underlying effect. Claims—explicit or implicit—about effect sizes are commonly based on a literature which is full of estimates that are biased wildly upward.

- Readers of published articles often need to resort to a sort of forensics, especially with non-preregistered studies where there can be many researcher degrees of freedom in data exclusion, coding, and analysis.

- All these problems arise even when authors are acting in good faith. It is important to be able to criticize published research without impugning the integrity of the researchers; conversely, researchers should not fool themselves into thinking, just because they are morally upright, that they cannot publish work with serious and avoidable errors.

All these issues arose with the study we have considered here on physical healing as a function of perceived time. The underlying claim—that manipulating a clock to alter patients' subjective recovery time can affect actual physical recovery—does not fit in well with the standard paradigm of medicine, and the authors of the paper under question do not offer any mechanistic theory of action. As a result, they are under some burden to argue for the plausibility of this entire line of research, which unfortunately is itself based on studies with similar methodological flaws (small samples, measurements too noisy to study effects which are realistically small and highly variable, misapplied statistical methods, and workflows with enough researcher degrees of freedom to make it possible to find apparent statistical significance even in the absence of any underlying consistent effects). Finding these errors took effort on our part, although this was facilitated by the openly-posted data of Aungle and Langer (2023) and the openness of Crum and Langer to share the data from their 2007 paper.

Beyond our immediate recommendation not to trust these particular published claims of mind-body effects on healing, weight loss, and blood pressure, we can offer some general advice.

In the short term, interpret studies in light of realistic possible effect sizes, and accept that many experiments are just too small and noisy to provide useful scientific information: even when some statistically-significant comparisons can be found, they can be explained by chance variation. In the analysis stage, be aware of the challenges of design and adjustment in the presence of multilevel structure. When it is time to summarize and write up the study, present all comparisons of interest, ideally in graphical form, rather than focusing only on the largest or those that reach some statistical significance threshold. When an interesting result arises, nail down the finding by designing and carrying out an exact replication. Contrary to all your expectations, the replication might fail; indeed that is the reason for performing the replication in the first place (Nosek et al., 2012).

In the medium term, design new studies on the basis of plausible hypothetical models and then preregister before collecting data. Preregistration is a floor, not a ceiling: use it to specify initial analyses with the understanding that you can and should go further when you see unexpected patterns in the data. The design and preregistration stage is a good time to think hard about effect sizes and their variation and to understand an experimental design using simulated data (Gelman, 2024). If a small study appears to reveal a potentially interesting result; the next scientific step is to probe it carefully with future experiments, not to treat it as an established fact in the later literature.

Recognizing the importance of preregistration, journals, including *AMPPS*, are steadily increasing their support for it. The strongest current form of preregistration, Registered Reports, requires authors' prespecified statistical analysis plans to be reviewed before data collection starts, which brings external, impartial eyes to the process. Registered Reports show considerable promise in reducing publication bias (Scheel et al., 2021).

In the longer term, we hope that default analyses and workflows will keep up with advances in data collection and modeling and that the presence of stronger studies in the literature, along with formal replication studies, will allow researchers to avoid being trapped in a loop of pseudo-replication. When it comes to the study of mind-body interaction, we recommend moving away from shot-in-the-dark phenomenological studies—black-box experiments designed to demonstrate that intervention X has a large and statistically significant effect on outcome Y—toward studies designed to probe more fully specified theories by controlling and measuring intermediate outcomes.

The above recommendations should seem reasonable, but none of them are easy, even if—*especially if*—we are working from a position of honesty and transparency, which we believe characterizes ourselves and also the authors of the papers under discussion. Analyzing multilevel, panel, time-series, spatial, and network data is hard. There are statistical and computational literatures on all these topics, but said literature often gives conflicting recommendations. One of us has written a book on multilevel modeling and still remains confused about general recommendations for the inclusion of interactions when analyzing data with non-nested multilevel structure. So it's not like we can even just recommend practitioners to solve their analysis problems by asking a nearby statistician for advice. Hypothesizing effect sizes for simulated-data experimentation is another difficult task, requiring the hard work of making strong assumptions and committing to them, at least for the moment. We argue that this effort would be well spent, but it adds to the cognitive cost of conducting a study. It's a lot more work to design, hypothesize, preregister, and conduct a study than to just gather some data and run with it. Reading the literature with a critical eye can also be hard work—as we demonstrated to ourselves in the preparation of the present article—which often can seem wasted if the ultimate conclusion is not to trust that literature. Far easier to just take titles and abstracts at face value and perhaps to pipe previously published results into a meta-analysis.

In short, we are recommending a steady dose of blood, toil, tears, and sweat, with the argument being that this is the only way to make progress when working in a field where effects are small and highly variable.

4.4. Statistical and conceptual problems go together

We have focused our inquiry on the Aungle and Langer (2023) paper, which, despite the evident care that went into it, has many problems that we have often seen elsewhere in the human sciences: weak theory, noisy data, a data structure necessitating a complicated statistical analysis that was done wrong, uncontrolled researcher degrees of freedom, lack of preregistration or replication, and an uncritical reliance on a literature that also has all these problems.

Any one or two of these problems would raise a concern, but we argue that it is no coincidence that they all have happened together in one paper, and, as we noted earlier, this was by no means the only example we could have chosen to illustrate these issues. Weak theory often goes with noisy data: it is hard to know to collect relevant data to test a theory that is not well specified. Such studies often have a scattershot flavor with many different predictors and outcomes being measured in the hope that something will come up, thus yielding difficult data structures requiring complicated analyses with many researcher degrees of freedom. When underlying effects are small and highly variable, direct replications are often unsuccessful, leading to literatures that are full of unreplicated studies that continue to get cited without qualification. This seems to be a particular problem with claims about the potentially beneficial effects of emotional states on physical health outcomes; indeed, one of us found enough material for an entire Ph.D. dissertation on this topic (N. J. L. Brown, 2019).

Finally, all of this occurs in the context of what we believe is a sincere and highly motivated research program. The work being done in this literature can feel like science: a continual refinement of hypotheses in light of data, theory, and previous knowledge. It is through a combination of statistics (recognizing the biases and uncertainty in estimates in the context of variation and selection effects) and reality checks (including direct replications) that we have learned that this work, which looks and feels so much like science, can be missing some crucial components. This is why we believe there is general value in the effort taken in the present article to look carefully at the details of what went wrong in this one study and in the literature on which it is based.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *Quarterly Journal of Economics* **138**, 1–35. https://doi.org/10.1093/qje/qjac038

Aungle, P., and Langer, E. J. (2023). Physical healing as a function of perceived time. *Scientific Reports* **13**, 22432. https://doi.org/10.1038/s41598-023-50009-3

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* **100**, 407–425. https://doi.org/10.1037/a0021524

Bishop, D. V. M. (2020a). The psychology of experimental psychologists: Overcoming cognitive constraints to improve research. *Quarterly Journal of Experimental Psychology* **73**, 1–19. https://doi.org/10.1177/1747021819886519

Bishop, D. V. M. (2020b). Low-level lasers. Part 2. Erchonia and the universal panacea. *Bishop Blog*, 5 Dec. https://deevybee.blogspot.com/2023/

Blakeslee, S. (1998). Placebos prove so powerful even experts are surprised. *New York Times*, 13 Oct, F1.

Brown, N. J. L. (2019). *Can Positive Emotions Improve Physical Health? An Examination of Some Claims from Positive Psychology*. Ph.D. thesis, University of Groningen. https://doi.org/10.33612/diss.99196913

Brown, W. A. (2015). Expectation, the placebo effect and the response to treatment. *Rhode Island Medical Journal* **98** (5), 19–21.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 1–12. https://doi.org/10.1038/nrn3475

Carey, B. (2011). Journal's paper on ESP expected to prompt outrage. *New York Times*, 6 Jan, A1.

Carroll, S. (2024). Ellen Langer on mindfulness and the body. *Mindscape Podcast*, 17 Jun. https://www.preposterousuniverse.com/podcast/2024/06/17/279-ellen-langer-on-mindfulness-and-the-body/

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., and Dorie, V. (2014). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics* **40**, 136–157. https://doi.org/10.3102/1076998615570945

Cochran, W. G., and Cox, G. M. (1957). *Experimental Designs*, second edition. Wiley.

Coyne, J. (2014). Re-examining Ellen Langer's classic study of giving plants to nursing home residents. *Plos Blogs*, 5 Nov. https://web.archive.org/web/20141109090707/http://blogs.plos.org/mindthebrain/2014/11/05/re-examining-ellen-langers-classic-study-giving-plants-nursing-home-residents

Crum, A. J., and Langer, E. J. (2007). Mind-set matters: Exercise and the placebo effect. *Psychological Science* **18**, 165–171. https://doi.org/10.1111/j.1467-9280.2007.01867.x

DeSmith, C. (2024). Glimpse into how mind may affect healing. *Harvard Gazette*, 1

Mar. https://news.harvard.edu/gazette/story/2024/03/glimpse-into-how-mind-may-affect-healing/

Devezer, B., Navarro, D. J., Vandekerckhove, J., and Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science* **8**, 200805. https://doi.org/10.1098/rsos.200805

Engber, D. (2017). Daryl Bem proved ESP is real, which means science is broken. *Slate*, 7 Jun. https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin* **44**, 16–23. https://doi.org/10.1177/0146167217729162

Gelman, A. (2022). The real problem of that nudge meta-analysis is not that it includes 12 papers by noted fraudsters; it's the GIGO of it all. *Statistical Modeling, Causal Inference, and Social Science*, 10 Jan. https://statmodeling.stat.columbia.edu/2022/01/10/the-real-problem-of-that-nudge-meta-analysis-is-not-that-it-include-12-papers-by-noted-fraudsters-its-the-gigo-of-it-all

Gelman, A. (2024). Before data analysis: Additional recommendations for designing experiments to learn about the world. *Journal of Consumer Psychology* **34**, 190–191. https://doi.org/10.1002/jcpy.1378

Gelman, A., and Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., and Loken, E. (2014). The statistical crisis in science. *American Scientist* **102**, 460–465. https://doi.org/10.1511/2014.111.460

Ikemi, Y., and Nakagawa, S. (1962). A psychosomatic study of contagious dermatitis. *Kyushu Journal of Medical Science* **13**, 335–350.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Langer, E., and Aungle, P. (2024). How much do our thoughts shape our health? *Scientific American*, 3 May. https://www.scientificamerican.com/article/how-much-do-our-thoughts-shape-our-health/

Leibowitz, K. A., Hardebeck, E. J., Goyer, J. P., and Crum, A. J. (2018). Physician assurance reduces patient symptoms in US adults: An experimental study. *Journal of General Internal Medicine* **33**, 2051–2052. https://doi.org/10.1007/s11606-018-4627-z

Leibowitz, K. A., Hardebeck, E. J., Goyer, J. P., and Crum, A. J. (2019). The role of patient beliefs in open-label placebo effects. *Health Psychology* **38**, 613–622. https://doi.org/10.1037/hea0000751

Levitt, S. (2024). Pay attention! (Your body will thank you). *Freakonomics Podcast*, 7 Jun. https://freakonomics.com/podcast/pay-attention-your-body-will-thank-you/

Liberman, M. (2009). Generalization and truth. *Language Log*, 3 May. https://languagelog.ldc.upenn.edu/nll/?p=1396

Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* **7**, 615–631. https://doi.org/10.1177/1745691612459058

Oberauer, K., and Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review* **26**, 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Peterson, J. B. (2023). Mindset, health, and life. *Jordan Peterson Podcast*, 31 Aug. https://www.youtube.com/watch?v=MPVMcBPlKeY

Plain, C. (2024). Perception of time can dramatically alter this crucial biological process, new study reveals. *The Debrief*, 3 Jan. https://thedebrief.org/perception-of-time-can-dramatically-alter-this-crucial-biological-process-new-study-reveals

Rand, P. (2023). Why the secret to health lies in the mind-body connection, with Ellen Langer. *Big Brains Podcast*, University of Chicago, 7 Sep. https://news.uchicago.edu/why-secret-health-lies-mind-body-connection

Ritchie, S. J., Wiseman, R., and French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLos One* **7**, e33423. https://doi.org/10.1371/journal.pone.0033423

Scheel, A. M, Schijen, M. R. M. J., and Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science*, **4**, 1–12. https://doi.org/10.1177/25152459211007467

Simmons, J., Nelson, L., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* **22**, 1359–1366. https://doi.org/10.1177/0956797611417632

Singh, S., and Ernst, E. (2008). *Trick or Treatment: Alternative Medicine on Trial.* London: Bantam Press.

Smith, P. (1997). The Cottingley fairies: The end of a legend. Inn Narváez, Peter, *The Good People: New Fairylore Essays*, ed. P. , Narváez, 371–405. University Press of Kentucky.

Snijders, T. A. B., and Bosker, R. J. (1999). *Multilevel Analysis.* London: Sage.

Szászi, B., Higney, A. C., Charlton, A. B., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., and Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences* **119**, e2200732119. https://doi.org/10.1073/pnas.2200732119