Precise Answers to Vague Questions: Issues With Interactions

Julia M. Rohrer[1*] & Ruben C. Arslan[2*]

Psychological theories often invoke interactions but remain vague regarding the details. As a consequence, researchers may not know how to properly test them, and potentially run analyses that reliably return the wrong answer to their research question. We discuss three major issues regarding the prediction and interpretation of interactions. First, interactions can be removable in the sense that they appear or disappear depending on scaling decisions, with consequences for a variety of situations (e.g., binary or categorical outcomes, bounded scales with floor- and ceiling-effects). Second, interactions may be conceptualized as changes in slope or changes in correlations, and since these two phenomena do not necessarily coincide, researchers might draw wrong conclusions. Third, interactions may or may not be causally identified, and this determines which interpretations are valid. We illustrate each of these issues with examples from psychology and issue recommendations for how to best address them in a productive manner.

"Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years' work?"

"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

"But it was the Great Question! The Ultimate Question of Life, the Universe and Everything," howled Loonquawl."

From: Douglas Adams. "The Hitchhiker's Guide to the Galaxy"

[1] Department of Psychology, Leipzig University. julia.rohrer@posteo.de
[2] Max Planck Institute for Human Development, Berlin. ruben.arslan@gmail.com
* Contributed equally
Code and simulated data on OSF (osf.io/apxtv) and Github (https://rubenarslan.github.io/interactions).
Acknowledgements: We thank Felix Elwert for his valuable input provided over the course of multiple discussions, as well as Paul-Christian Bürkner, Daniël Lakens and Stefan C. Schmukle for their helpful comments on an earlier version to this draft.

Interactions are ubiquitous in psychological science. Person-situation interactions, stress-vulnerability models, gene-environment interactions: a large number of theoretical perspectives postulate that the effect of one variable depends on another variable. Interactions often seem highly plausible from a substantive perspective, and many empirical investigations consider them the central target of inquiry. Thus, it is no surprise that statistical procedures to test for interactions—running an ANOVA or multiplying the supposedly interacting variables and entering the product as a predictor in a regression analysis—are standard part of the training of psychological researchers.

Establishing an interaction might seem like a straightforward task; however, there is a range of potential complications. For one, *reliably* detecting interactions can be a challenge. There is some empirical evidence suggesting that interaction claims are less replicable (Beck & Jackson, 2020; Open Science Collaboration et al., 2015), and there are a-priori reasons to expect so. To find interactions where an effect is attenuated in one group, but not reversed compared to the other group, surprisingly large sample sizes are needed (e.g., Gelman, 2018; Giner-Sorolla, 2018; Lakens, 2020). This means that existing interaction effects are harder to confirm in empirical studies—but also that reportedly significant interactions are more likely to be false-positives, or even point in the wrong direction. Furthermore, studies with experimental interventions might only afford one plausible main effect but many different interactions to be considered (e.g., intervention times *any* demographic variable assessed), leading to a large number of researcher degrees of freedom which may further increase the risk of false-positives.

These issues affect the replicability of interactions. However, there are other issues with interactions that can result in *reliably* mistaken conclusions. Re-running the same study will not fix or reveal those issues—instead, one may get the wrong answer, every time. Some of these issues can be identified by close investigation of the data alone. For example, Hainmueller et al. (2016) discuss two problems and how to address them: the standard approach to interaction assumes a linear interaction effect which can be a misspecification; and estimates can be misleading if there is little data underlying certain regions of values (e.g., if, at some values of one of the interacting variables, there is little variability in the other interacting variable). But there is another set of issues that cannot be resolved by careful consideration of the data alone.

It is three such replicable issues with interactions on which we want to focus in the present article: The scale dependence of interactions, the distinction between moderation of slopes versus moderation of correlation, and the causal identification of interactions. We will

work through each of these issues with a motivating example based on data that have been simulated (which ensure that we know the true model) followed by relevant examples from the literature.

All of these issues have been discussed before, sometimes in great detail and clarity, often within specific substantive contexts, and we will highlight some of these works throughout the manuscript. But a look at contemporary publications in psychological journals suggests that researchers with a substantive focus often do little to address them. This may have multiple reasons: (1) researchers may not be aware of them; or (2) they may consider them esoteric details without consequences for their own substantive conclusions; or (3) they simply may not know how to address them best. Thus, in the present manuscript, we aim to (1) clearly explain these problems using simulated data; (2) illustrate how they can indeed affect substantive conclusions in published research; and (3) provide constructive recommendations on how to address them.

Now You See It, Now You Don't: Interactions are Scale Dependent

Motivating Example

A large company is testing a mentoring program with the hope of increasing employee retention. For this purpose, they have randomly assigned 10 percent of their employees to participate in the program, and they plan to evaluate its effects on quitting one year later. But then a global pandemic forces a change of plan and due to distancing measures, only half of the employees can work on-site, with the other half working remotely. The decision who still works on site has been randomized. All the while, the mentoring program is continued through video conferencing.

Once the time has come for the evaluation of the mentoring program, there are now two questions that can be evaluated: Did the mentoring program reduce quitting? And did it do so equally for on-site and remote workers—in other words, was there an interaction between the mentoring program and work location?

All models reported in this manuscript have been estimated in *brms* (Bürkner, 2017, 2018) using default, weakly informative priors. Note that all the problems discussed in this manuscript occur regardless of whether a frequentist or a Bayesian approach is chosen,

whether you run an ANOVA or a regression; however, *brms* as a highly flexible statistical package allows us to adapt the models and procedures to easily mitigate the problems discussed here. Table 1 shows results from a logistic regression on simulated data in which the binary outcome quitting (0: no, 1: yes) was regressed onto three predictors: mentoring program participation (0: no, 1: yes), working on-site (0: no [remote work], 1: yes), and the product term of the two (mentoring*working on-site).

Table 1

Results from a Logistic Regression Analysis Predicting Quitting from Participation in the Mentoring Program, Working On-Site, and Their Interaction

Variable	Coefficient	95% Credible Interval
Intercept	-3.06	[-3.20; -2.92]
Mentoring	-1.46	[-2.38; -0.70]
Working on-site	2.90	[2.74; 3.05]
Mentoring*working on-site	0.87	[0.09; 1.81]

Note. N = 10,000, data have been simulated.

As hoped for, the mentoring program had a negative effect on subsequent quitting. However, working on-site increased subsequent quitting quite dramatically, possibly because employees had health concerns or child care duties that could not have been fulfilled otherwise. Intriguingly, the coefficients from the logistic regression indicate an interaction: Among those who worked on-site, mentoring led to a smaller reduction in quitting—in other words, it seems like the program was less effective on-site (and more effective remotely).

A logistic regression models the effect of the predictors on an underlying continuous unbounded scale (here, a latent propensity towards quitting, which might be understood as job dissatisfaction) which is linked to the observed binary outcome (quitting yes/no) with a logistic function, see Figure 1.

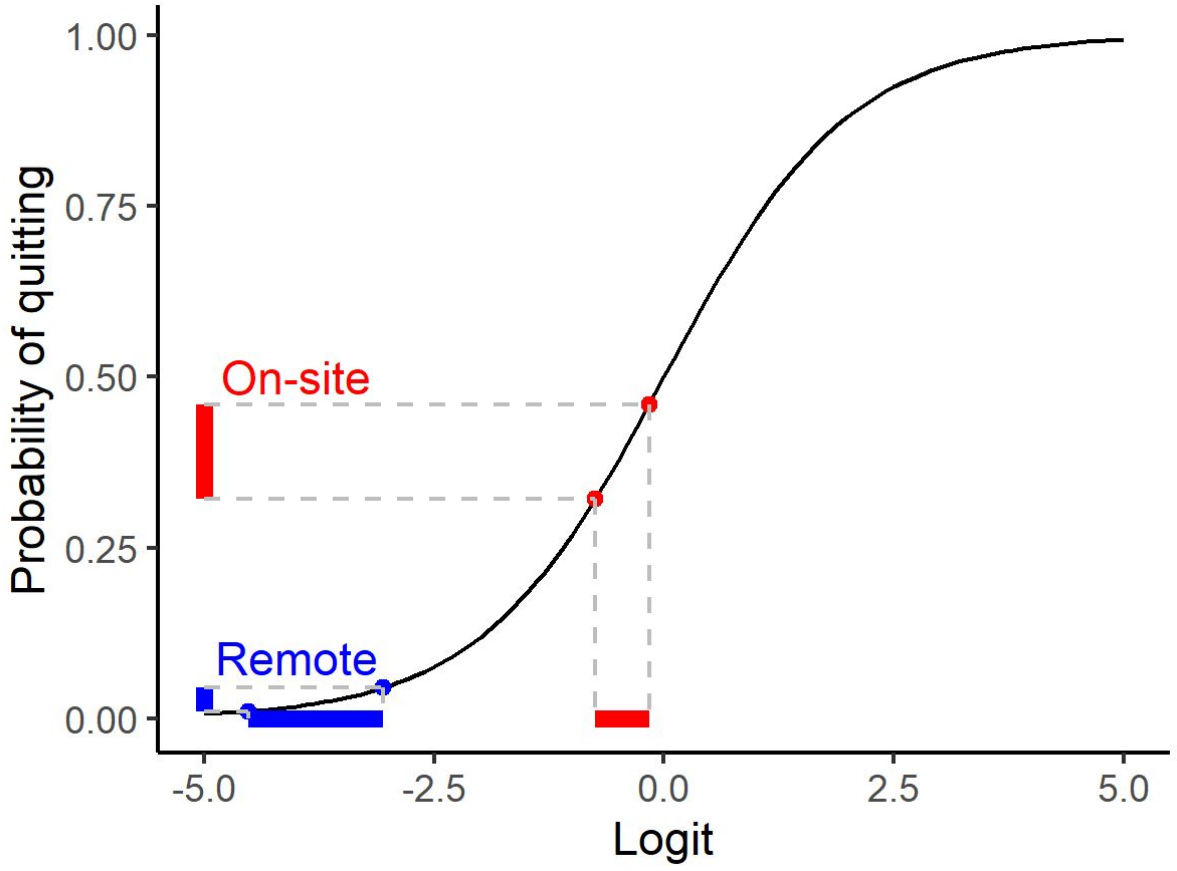

Figure 1. Effect of the mentoring program on quitting on logit scale (x axis) and on probability scale (y axis), for employees who worked remotely (blue) or on-site (red).

This link function ensures that the model only predicts possible values, probabilities that lie between 0 and 1. The horizontal bars on the x-axis visualizes the effect of the intervention for on-site employees (red bar) and remote employees (blue bar) on the logit scale. The intervention led to a larger reduction for remote employees (i.e., the horizontal blue bar is longer than the horizontal red bar). While the model coefficients need to be interpreted on the the assumed underlying Logit scale, our model also allows us to make statements about the effect of the program on a different scale, the *probability* of quitting. This can be done by comparing the *vertical* bars in Figure 1, or by simply plugging the model coefficients into the logistic function (see Table 2).

Table 2

Latent Quitting Propensities as Predicted by the Logistic Regression Model and the Corresponding Predicted Probabilities of Quitting

	Latent quitting propensity *q*	$P(q) = \frac{e^q}{1+e^q}$
not on-site, no mentoring	-3.06	4.5%
not on-site, mentoring	-3.06 - 1.46 = -4.52	1.1%
on-site, no mentoring	-3.06 + 2.90 = -0.16	46.0%
on-site, mentoring	-3.06 + 2.90 - 1.46 + 0.87 = -0.75	32.1%

Now, the picture changes. Mentoring changed the probability of quitting for remote employees from 4.5% to 1.1%, a decrease of 3.4 percentage points. But for on-site employees, mentoring reduced the probability of quitting from 46.0% to 32.1%, by as much as 13.9 percentage points. Because we have estimated the models in brms, we can easily estimate credible intervals for any particular metric of interest (see osf.io/apxtv) for more details). The estimated differences between the mentoring effects in the two groups is an astounding 10.5 percentage points, 95% CI: [-15.2; -5.8]. Hence, it very much looks like the intervention is *much* more effective on-site.

We have already arrived at conclusions that are diametrically opposed (mentoring was more effective remotely vs. mentoring was more effective on-site), but let's consider yet another outcome metric. A decrease from 4.5% to 1.1% corresponds to a relative risk of 1.1/4.5 = 0.24, 95% CI: [0.09; 0.52], which means that the risk of quitting was reduced by 1 - 0.24 = 0.76 = 76% for remote employees. In the group working on-site, a decrease from 46.0% to 32.1% corresponds to a relative risk of 0.70, 95% CI: [0.60, 0.80], which means that the risk of quitting was reduced by 1 - 0.70 = 0.30 = 30%. Thus, the relative risk reduction was much larger in the group working remotely, and we might once again conclude that the intervention is more effective among remote employees.

Which interpretation is the correct one?

We have now arrived at conclusions that seemingly contradict each other: a positive interaction, a negative interaction, and once again a positive interaction. However, these findings are perfectly compatible, they simply re-state the same pattern on different scales.

On the assumed latent continuous quitting propensity, which might be understood as dissatisfaction with the job, the mentoring effect is larger for remote employees. However, the reduction in the probability of quitting is much larger for on-site employees. And the relative reduction in the probability of quitting is once again larger for remote employees.

Readers from psychology might favor to evaluate the interaction by interpreting the interaction term from the logistic regression model; hence, their conclusions would apply to the latent continuous quitting propensity. For example, Simonsohn (2017) distinguishes between “conceptual interactions” that arise from “variables actually influencing each other”, captured by model coefficients (here: the logit coefficients); and “mechanical interactions” that arise from the non-linearity of the model (and are implied to be less interesting because they will supposedly arise *in any case*). The substantive interpretation of the coefficients from the nonlinear model assumes a continuous underlying latent variable that is linked to the observed outcome (quitting) following a certain functional shape (here: a logistic function). Psychologists may or may not be willing to endorse these assumptions, but they are hardly ever made explicit—interpreting interaction coefficients from nonlinear models directly seems a default solution rather than a principled decision.

Researchers from other fields have arrived at diametrically opposed preferences, generally favoring probabilities as the relevant outcome scale. For example, an editorial comment in *American Sociological Review* states that “[t]he case is closed: don’t use the coefficient of the interaction term to draw conclusions about statistical interaction in categorical models such as logit, probit, Poisson, and so on” (Mustillo et al., 2018). Likewise, the seminal paper on interaction terms in nonlinear models in economics (Ai & Norton, 2003) does not even consider the possibility that the coefficient of the interaction term might correspond to anything of particular interest.[3] However, we should not let disciplinary norms dictate our scaling assumptions, but instead motivate them for the question at hand (Hand, 1994).

In the example presented above, the “right” scale depends on the decision that has to be made. If the company wants to keep as much of their workforce as possible while saving mentoring costs during the pandemic, it might be most effective to restrict mentoring to on-site workers, because that would prevent more resignations. In other words, the probability scale would be of central interest. But after the pandemic, things may look differently. The company decides to remain partly remote, and they have good reason to

[3] In general, economists are fond of linear probability models which model the probability of the outcome as a linear function of the predictors, rather than assuming a non-linear link function.

believe that the high quitting propensity among on-site workers was restricted to pandemic conditions. Thus, it might be most effective to focus mentoring efforts on remote workers, because their job dissatisfaction (i.e., the latent quitting propensity) is reduced more strongly. In other words, the logit scale would be of central interest.

The Scale Dependence of Interactions

In cognitive psychology, the scale dependence of interactions has been pointed out more than 40 years ago by (Loftus, 1978).[4] He noted that when response probabilities are used as the dependent variable, some interactions can be "removed" by assuming a different mapping between response probability and the assumed underlying component of memory, and concluded that such interactions may be uninterpretable. Three decades later, Wagenmakers et al. (2012) followed up on the phenomenon. Studying citation histories, questionnaires, statistical textbooks and published articles, they illustrated how experimental psychologists had largely remained unaware of the problem. Judging from the published literature, psychologists from many non-experimental fields have remained at least as unaware, although there are of course exceptions. For example, Johnson (2007) provides a clear discussion of the problem in the context of research on gene-environment interactions, and Murray et al. (2016) develop scaling recommendations for this field.

Thus, it may be important to note that the scale dependence of interactions is a broad phenomenon that affects all substantive fields, and that applies to a wide variety of situations. Conclusions about the magnitude of interactions, and, in some cases about their presence and sign (see Wagenmakers et al., 2012, for different scenarios) depend on scaling decisions. Scaling decisions might be trivial when there is a single natural mapping between the observed measure and the underlying process or construct of interest. However, this is rarely the case. The following examples illustrate the relevance of this phenomenon.

Substantive Examples

Flattening the Curve. Our World in Data (Roser et al., 2020) presents the incidence of people testing positive for SARS-CoV-2 per capita, across countries, in an interactive chart. Viewers can toggle whether they want to see the comparison on a linear or a log-linear scale. The log-linear scale makes it easiest to judge which countries are doing

[4] There is also a large body of literature on the subject in other fields such as epidemiology (see e.g., Rothman et al., 1980, for a recap of an earlier debate), which has largely gone unacknowledged within psychology.

better at flattening the curve, that is, reducing new infections below the numbers expected based on the current number of infected people. We might want to make this comparison if we want to find out whether public health interventions in one country are more effective than in another. However, the linear scale makes it easier to judge which countries are currently worst affected, and makes it easier to see, for instance, which countries will exhaust the number of available intensive care units sooner.

Summer break gaps. It has been reported that test score gaps between advantaged and disadvantaged students grow fastest during summer vacation, which has been interpreted as evidence that the major sources of inequality lie outside of the school context. Note that the pattern describes an interaction between time and socio-economic status on students' academic abilities. There is no single defensible mapping between students' abilities and their responses on the test items—thus, scale dependence may be an issue. von Hippel and Hamrock (2019) demonstrated empirically that conclusions about the growth of gaps are sensitive to whether one analyzes (an estimate of) the number of correct answers, or an ability estimate from an item response theory model.

The Paradox of Declining Female Happiness. Happiness research routinely employs single-item measures that seem at best ordinal in nature. Hence, the question arises how actual happiness—which we can plausibly assume to be continuous—relates to the observed categorical answers. Bond and Lang (2019) explore how different assumptions about the underlying distribution (such as skewness) of happiness affect substantive conclusions. One of their examples highlights how the supposed relative decline in US-American women's happiness since the 1970s can be "removed" by assuming that the happiness distribution is left-skewed.

Recommendations

There is no way to circumvent the fact that conclusions regarding interactions often rest on scaling assumptions. But as with all assumptions, there are ways to address them that are more productive than simply glossing over them.

First of all, not all interactions can be removed or reversed. If an interaction is non-removable, the qualitative conclusion that there is a certain interaction pattern is robust under various scaling assumptions, namely strictly[5] monotonic transformations of the link

[5] Wagenmakers et al. (2012) state that "a nonremovable interaction can never be undone by a monotonic transformation of the measurement scale" (pp. 145). However, consider the (admittedly pathologic) case of the constant function: if we transform the outcome scale so that every single

function. Cross-over interactions can be considered non-removable; Wagenmakers et al. (2012) provide a more detailed investigation into the conditions under which interactions are removable (or not).

Second, as with all assumptions, scaling assumptions should be spelled out and reflected upon. For example, many psychologists might not be aware of the precise assumptions that they implicitly endorse by analyzing an ordinal outcome scale as if it were continuous (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018); or by analyzing sum scores as opposed to estimates from more complex models; or they may be unaware of the features of the link function their model uses. A thorough consideration of the assumption one is willing to make may lead to a suitable statistical model; and testing whether the interpretation changes under different assumptions becomes a natural robustness test.

Third, if multiple scaling assumptions could be justified or are familiar to readers, there is little reason not to show the data multiple ways, perhaps even in an interactive plot.

Results regarding both main effects and interactions should be reported for all of them. For example, the coefficients from logistic (or probit) regression models should be supplemented by probabilities of the outcome. As demonstrated above, this is easily done for the simple 2x2 case without additional covariates; Ai and Norton (2003) and McCabe et al. (2020) provide guidance for more complex scenarios.

Box 1: Floor- and Ceiling-Effects

There is one particular issue of scale dependence that stands out because (1) it is common, in particular in subfields relying on rating scales and because (2) it requires less nuanced consideration than other cases, because a flawed measurement model is at the root of the interpretation issue.[6]

individual has the same value, *all* interactions are undone (and so are all main effects). So there exists a monotonic transformation that removes nonremovable interactions. Of course, the constant function may not be particular relevant for empirical research, but there are other cases of monotonous but not strictly monotonous transformation that are of interest: ceiling- and floor-effects, in which all individuals past the scale boundaries are assigned the same value. Ceiling- and floor effects may remove nonremovable interactions when the reversal of the direction of the effect occurs outside of the scale boundaries.

[6] In this section, we focus on floor and ceiling effects that result from flawed measurement of quantities that are assumed to be unbounded. Of course, there are also variables that are inherently bounded (e.g., one cannot have fewer than zero children in a household).

Researchers who use rating scales will often notice that values are not nicely distributed across the scale range, but bunch up at the lower or upper scale end. Such scales can easily induce spurious interactions. Consider the simulated data in Figure 2, Panel A, which shows the relationship between a predictor and an outcome in two groups marked by color. As we can see by the equal slopes of the two regression lines, there is no interaction between predictor and the color-coded group membership.

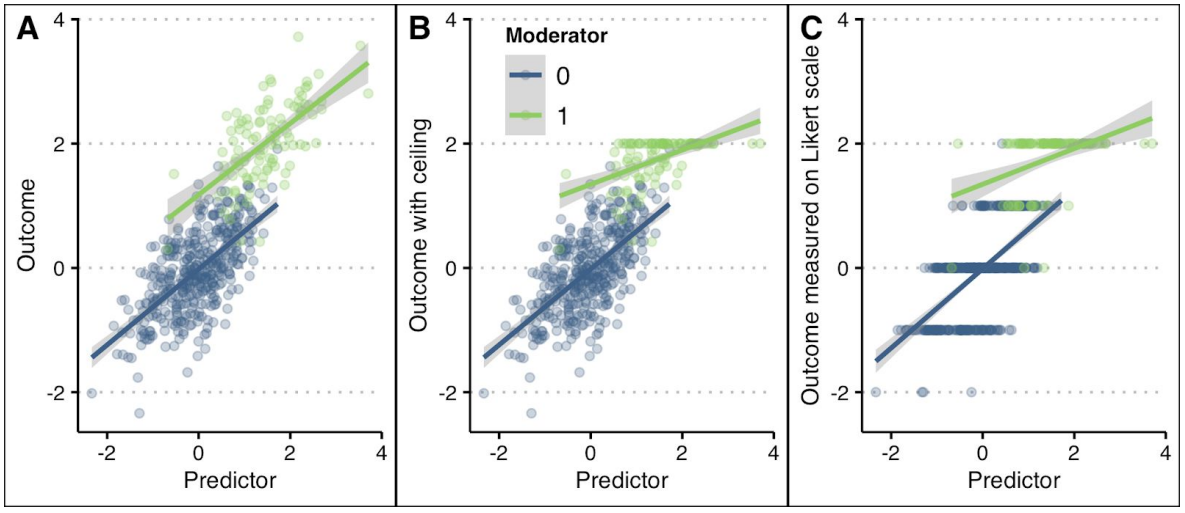

Figure 2. Effect of a predictor variable on an outcome in two different groups. Panel A: no interaction. Panel B: ceiling effect induces a spurious interaction. Panel C: an ordinal variable induces a spurious interaction.

However, assume that the measurement device that we were using suffered from a ceiling effect as shown in Panel B. Suddenly, the slopes between the regression lines differ, and a regular regression analysis would indicate an interaction between the predictor and group membership, with the slope in the group closer to the ceiling being flatter. This interaction is dependent on the assumption that there is a linear mapping between the measured outcome and the unobserved metric of interest across the whole range of observed values. In other words, the regression model assumes that all individuals at the ceiling do indeed have the same value on the outcome, rather than exhibiting some variability that got censored by the scale. If we want to be open to the possibility that such variability exists, we can instead run a regression model with censoring. All observed outcomes at the upper limit of the scale are labeled as

right-censored (i.e., we assume that we only know that they have the observed value *or higher*) — and suddenly, the spurious interaction disappears.[7]

Panel C shows a scenario that is even more common in psychology. The outcome variable from panel A has been mapped to an ordinal scale, such as a Likert scale. Again, values can only occur within the scale's bounds. If we model the outcome as with a regular linear regression model, we would infer an interaction. If we instead estimate an ordinal regression, for instance a cumulative model with equidistant thresholds (Bürkner & Vuorre, 2019), the interaction disappears.

Problems caused by ceiling- or floor-effects are exacerbated the stronger the main effect of the moderator on the outcome is, as this will push individuals with certain moderator values closer to the boundaries of the scale. In such scenarios, we suggest that the best course of action is more clear cut than in more subtle ones described above. Many measures in psychology aim to yield approximately normal distributions. But normal distributions cannot always be ensured across research settings and moderator categories, so if measures are bounded, but the latent quantity is not, we need to account for this flaw of our measure in our model. Not doing so can result in interactions that should not be interpreted in a substantive manner, as they can be explained by more realistic measurement assumptions—variability beyond the range of the scale; a lack of symmetry in the mapping between the metric of interest and the observed response; rating scales being ordinal rather than metric.

Models building on such more realistic assumptions are not routinely used within psychology, and they are not implemented in many common software packages. In our experience, many researchers think that such models do not make much of a difference for results anyway, and thus only complicate analyses. The former intuition may draw on experiences regarding the estimation of main effects, and we concede that these may often be surprisingly robust to measurement issues, but this is not the case for interactions. The latter concern may be justified: more complex models are more complicated to run and to interpret. However, as mentioned above, the R package *brms* affords a lot of flexibility. Appropriate models for censored and skewed data can be implemented in a straightforward manner, and (Bürkner & Vuorre, 2019) have written an excellent tutorial for ordinal regression models.

[7] Running such a model only requires minor modifications of the model syntax in the R package *brms*, see details on the Open Science Framework (osf.io/apxtv).

Same Slope does not Imply Same Correlation

Motivating Example

A group of researchers predicts that satisfaction with one's job matters more for overall well-being among singles. To test their hypothesis, they collect within-subject daily diary data on both job satisfaction and overall well-being from both singles and individuals in relationships. Once they are ready to analyse the data, they discover that they had two different tests in mind. One researcher wants to compute intra-individual correlations between job satisfaction and overall well-being and compare their averages between singles and non-singles. Another researcher wants to estimate an interaction between singlehood and job satisfaction in a multilevel regression on overall well-being. They run both analyses and, to their surprise, they find that there is a substantial difference in the correlations between singles and non-singles, but the interaction effect is close to zero.

What is going on here? In the simple bivariate case—one outcome (overall well-being), one predictor (job satisfaction)—the correlation coefficient equals the standardized regression coefficient, which equals the unstandardized regression coefficient multiplied with the ratio of the standard deviation of the predictor to the standard deviation of the outcome. The variability in the outcome may be further decomposed into variability that can be attributed to the predictor X, and variability that remains unexplained — i.e., residual variability.

Eq. 1:

$$r = \beta = b \frac{\sigma_x}{\sigma_y} = b \frac{\sigma_x}{\sqrt{\sigma^2_y}} = b \frac{\sigma_x}{\sqrt{\sigma^2_{y\,explained\ by\ X} + \sigma^2_{y\,not\ explained\ by\ X}}} = b \frac{\sigma_x}{\sqrt{b^2\sigma^2_x + \sigma^2_{residual(y)}}}$$

So if a correlation varies between groups, it can have multiple reasons. The underlying unstandardized effect, i.e., the slope *b* may vary, or the standard variability of the predictor σ_x may vary, or the variability of the outcome σ_y may vary. The standard procedure to test for an interaction effect only considers changes in the slope. Thus, a comparison of correlations and a comparison of slopes (i.e., a standard test for an interaction) will result in

different patterns whenever the ratio of variances (predictor to outcome) varies between the groups that are compared (e.g., Smithson, 2012). Sometimes, we may find a difference in correlations but no difference in slopes, as in the example above. However, it is also possible to find a difference in slopes but no difference in correlations, as illustrated in Box 2.

The results observed above are hence no statistical surprise, but we still do not know which analysis (and which conclusion) should be preferred. The group of researchers thinks about the issue more deeply, and they realise that their verbal prediction—"job satisfaction *matters more* for overall well-being among singles"—was too vague. In fact, two of them had thought of quite different scenarios. One of them thought of a scenario in which, when singles have a bad day on the job, their overall well-being drops by more points, but good days give them a bigger boost (Figure 3A). This scenario corresponds to group differences in *b*, and it could be directly tested with the regression model including the interaction term. All else being equal, a difference in slopes will also result in a difference in correlations. But if all else is not equal (i.e., if variances differ between groups), a difference in slopes may not result in a corresponding difference in correlations.

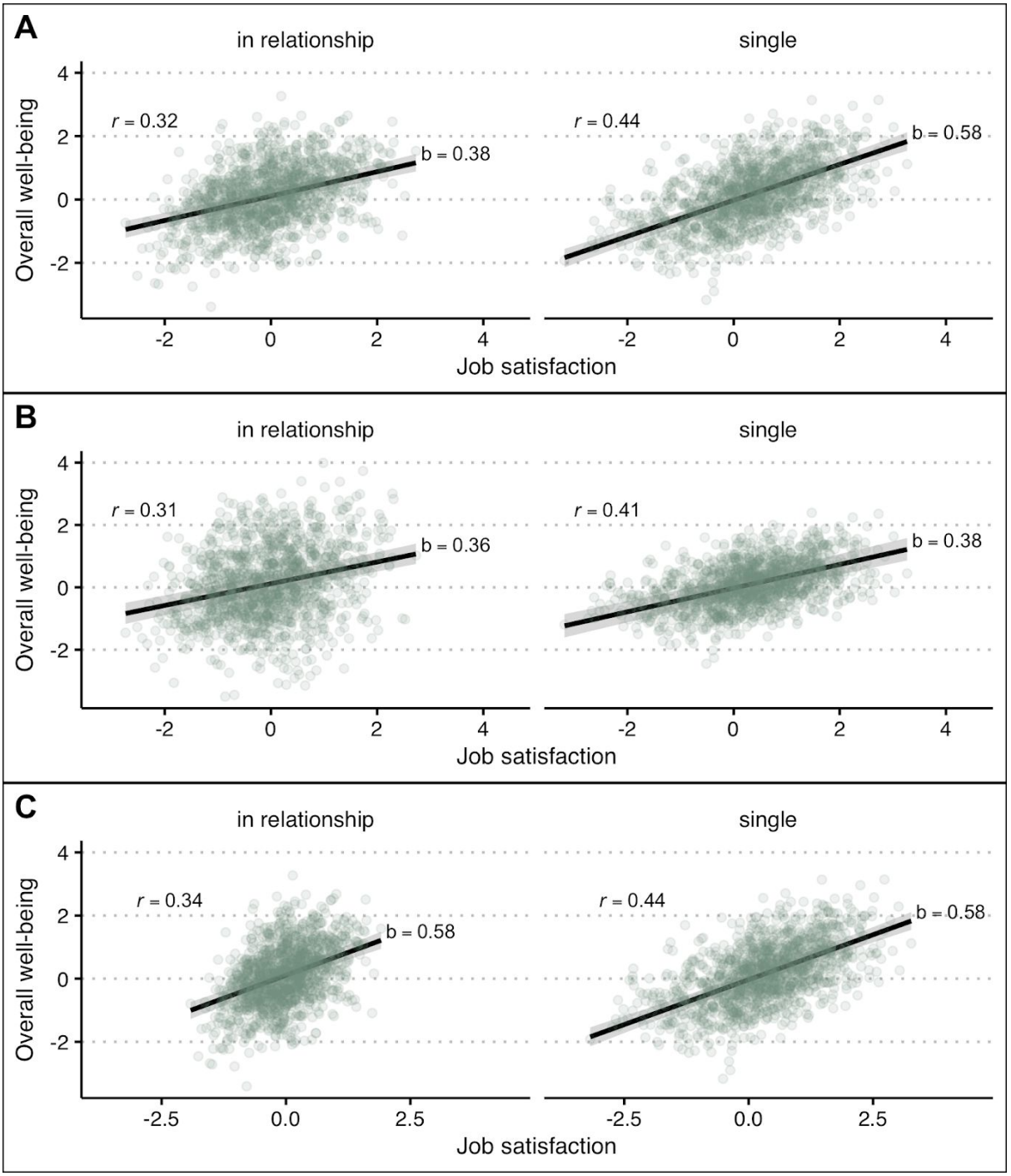

Figure 3. Relationship between job satisfaction and overall well-being for individuals in relationships and singles. Scenario (A): larger effect of job satisfaction among singles. Scenario (B): same effect, but there is less unexplained variability in well-being among singles. Scenario C: same effect, but there is less variability in job-satisfaction among singles. N=50 people sampled on 50 days each, data have been simulated.

Another one thought about things in a different way. A one-point change in job satisfaction may have the same effect among both singles and individuals in a relationship. But being in a relationship is an additional source of variation in well-being which is independent of job satisfaction, and thus, partnered individuals will have more variance in their overall well-being that cannot be explained by job satisfaction. In this scenario, we have group differences in the residual variance $\sigma^2_{residual(y)}$, which result in group differences in the variability of the outcome σ_y and all else (slope, variance of the predictor) being equal, it would result in a higher correlation for singles than for partnered individuals, but no differences in the moderation analysis difference in correlations (Figure 3B).

As we mentioned above, another thing that could differ between groups is the variance in the predictor. For example, let us imagine that in our study, individuals who are in a relationship tend to have "settled down" with jobs that are overall more steady-going. Some of the singles, in contrast, have exciting jobs that come with more ups and downs. This corresponds to group differences in the variance of the predictor. If the slopes are the same in both groups, this will also lead to group differences in the variance of the outcome; but the differences in the variance of the predictor will be larger and thus, the correlation will increase (Figure 3C).

Box 2: Likewise, same correlation does not imply same slope

A company wants to find out whether working from home boosts productivity. To this end, they randomise employees to work between zero and five days a week from home and measure productivity using their in-house tracking of productive hours per week. Because employees with children tend to be less productive, the company leadership are especially interested whether parents benefit from remote work more than non-parents. The parents have lobbied for remote work as a solution, but the leadership is skeptical. The trial is run and the results show that working from home benefits productivity, though the correlation is small. To test whether parents' productivity benefitted more substantially, the correlations between productivity and days worked from home are computed for parents and non-parents. There is no difference between the correlations. The leadership says they will take the results under consideration. However, one mother, worried that they will keep the status quo, requests the trial data. Armed with Figure 4, she marches into the head office.

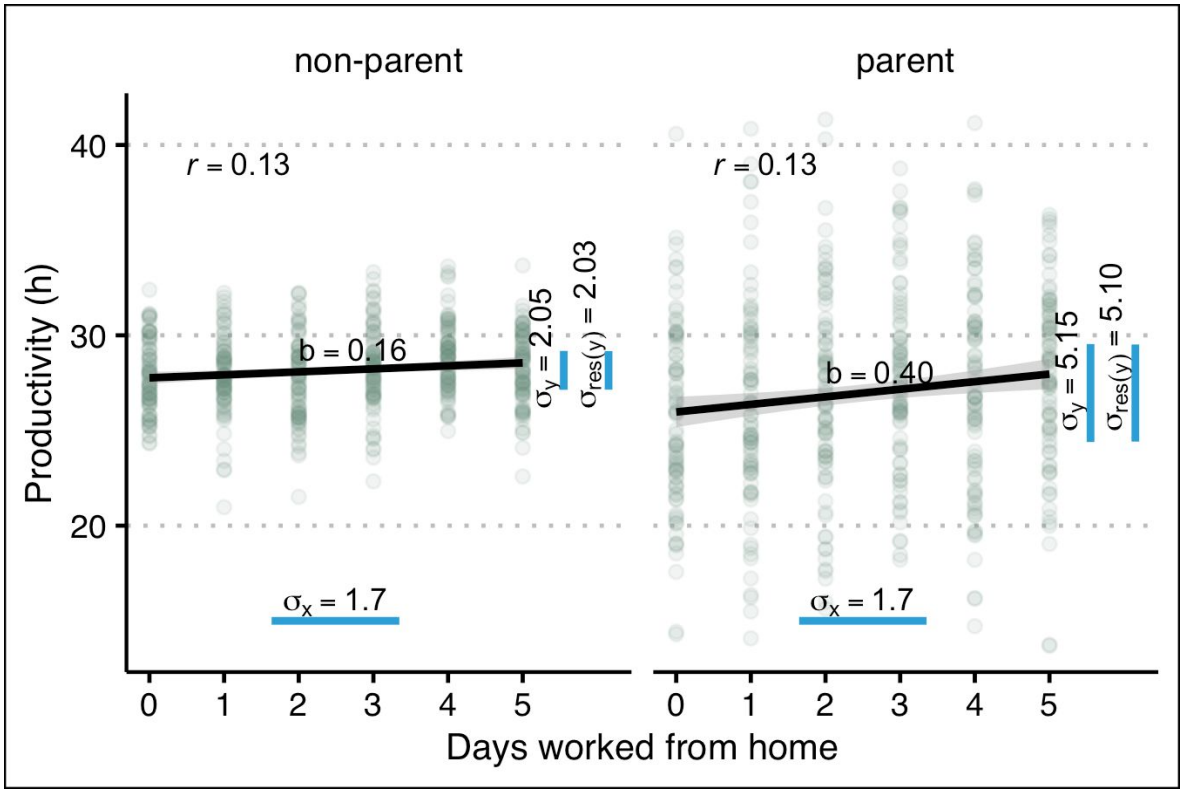

Figure 4. The association between days worked from home and productivity (measured in hours) for parents and non-parents. The blue vertical lines show the standard deviations of productivity and of the residual variation. The blue horizontal line shows the standard deviation of days worked from home.

As she patiently explains, the black line is steeper for parents. This means that, on average, their productivity was boosted more by working from home, almost enough to close the parent productivity gap. But should there not be a difference in the correlations as well then, the leadership asks? Yes, *if* all else were equal. But as the raw data shows, parents' productivity fluctuates much more than that of non-parents (i.e., productivity is *heteroskedastic* across parenthood). Although she cannot pinpoint the exact reason, lost sleep and the perennial infections brought in from daycare seem like good candidates. She has days where she gets almost no hours of productive work done, but—after the occasional good night's sleep—also some days of eight hour focus. Childless colleagues exhibit much stabler productivity and rarely drop below 4 hours a day. Working from home allowed her to manage her time better, and reduced scheduling conflicts, but the fundamental factors of sleep and health remain the same. The leadership cares about optimising average productivity. That the productivity of parents fluctuates so much in ways that are unrelated to their company policy is not under their control. Hence,

comparing correlations was clearly not the right test for their question--instead, they should have looked at the slope, the effect of days worked from home on productive hours. While this example may seem contrived, weaker versions of these patterns occur frequently and can cause over- and underestimation of the effects of interest.

Choosing the Appropriate Model

Phrases which we routinely use to formulate verbal predictions, for example, "matters more", "stronger influence", or "more important for" could refer both to steeper slopes or larger correlations. In psychology, where many constructs are measured in abstract quantities like points on a Likert scale, standardised effect sizes (e.g., Pearson's product-moment correlations or standardised regression coefficients) are so widespread that the distinction between slopes and correlations may often be lost. However, as we have shown above, the distinction matters when we want to ensure that we are actually testing our substantive hypotheses.

As a first step, we need to be more specific about what our theory actually says (Hand, 1994). Verbal theory specification leaves room for ambiguities; formalizing our theories with the help of equations or computational models can remove these ambiguities (Smaldino, 2017) and force us to think more carefully about slopes and variances, along with many other assumptions and predictions. Going even further, the resulting statistical hypotheses could be reported in a machine-readable format which results in a maximum of clarity and transparency (Lakens & DeBruine, 2020). But even just a simple data simulation, such as the examples above, can help clarify the relationship between substantive hypothesis and patterns in the data. Once we have a better understanding of what we want to test, we can start thinking about the appropriate statistical model.

If a difference between slopes is of interest, we are firmly in the territory of interactions and may, for example, simply include the product term between the variables of interest. But what should we do if our hypothesis concerns the variance components? It may be tempting to simply compare correlations, but as we have seen above, correlations are sensitive to differences in slope, differences in the variance of the predictor, and differences in the variance of the outcome. Thus, depending on our hypothesis of interest, the correlation coefficient may be too “coarse” and miss important patterns in the data, and we should instead explicitly model the variance of interest.

Once again, the flexibility of *brms* pays off—it allows us to implement distributional models in a straightforward manner (Bürkner, 2020; Umlauf & Kneib, 2018). To understand how these models work, let's quickly think about what a simple "regular" regression model does. In a standard regression model, we are predicting an outcome which is normally distributed. This normal distribution is described by two parameters: its mean and its standard deviation (which captures the residual unexplained by our model). When we include predictors, these variables explain the first parameter, the mean of the normal distribution; the standard deviation is assumed to be constant across all observations. A distributional model additionally allows us to include predictors to explain the second parameter, the residual standard deviation.

We can use this approach to simply test, for example, whether a continuous variable has an effect on the standard deviation of another variable. If we additionally include predictors for the mean, we can then test whether the residual standard deviation—the variability not explained by the predictors included for the mean—varies depending on the level of some third variable. Concerning our example above, we could *simultaneously* test for an interaction in the narrower sense (difference in slopes) and for group differences in the residual standard deviation (differences in unexplained outcome variability) by running the following model in *brms*:

```
overall well-being ~ job satisfaction * single
sigma ~ single
```

The first line represents the standard interaction analysis which can be interpreted in the usual manner. The second line allows the residual standard deviation to vary depending on whether or not individuals are single. Here, a negative coefficient would indicate that the residual standard deviation is smaller among singles, which captures the idea that there is less unexplained variability among singles, as they lack a romantic relationship which constitutes a major source of variability in well-being. On OSF, we provide more detailed code examples (osf.io/apxtv).

Substantive Examples

Individual importance weighting. Rohrer and Schmukle (2018) investigated importance weighting — the idea that individuals judge the overall quality of their lives by aggregating their satisfaction with various life domains. In this model, more important domains receive higher weight, and in the literature, this is normally captured by including

interactions between domain satisfaction and domain importance ratings (e.g., by including the product term *importance rating of health*satisfaction with health*). However, after publication, we discussed whether importance weighting could not also be interpreted as improved prediction—for example, when regressing overall life satisfaction on health satisfaction, there should be less unexplained variance among individuals who consider health very important for their health. The substantive literature on the topic does not take a clear stance on what exactly is meant by "weighting" as it mostly relies on vague verbalizations, which probably go unchallenged because the underlying notion ("things that are important matter more") is so intuitive.

Partner preferences after entering a relationship. Gerlach et al. (2017) investigated the stability of partner preferences in a sample of singles. At time 2, a substantial proportion of participants had entered a new relationship. Gerlach et al. were interested in whether partner preferences changed more among those who found a partner. In a standard moderation analysis, the slope of partner preferences at time 1 on preferences at time 2 was highly similar among singles and those who found a partner. For example, on average, people who expressed a strong preference for attractiveness at time 1 did so again at time 2, regardless of whether they had found a partner in the meantime. However, among singles, the individual data points were closer to the regression line—at time 2, they deviated less from their preference at time 1—which resulted in a higher correlation across time. Individuals who had found a partner had adapted their preferences to better match the traits of their actual partners (e.g., if their partner fell short of their preference for attractiveness at time 1, they reported a weaker preference for attractiveness at time 2). This adaptation to the partner reduced the correlation with preferences at time 1, but not the slope.

Moderation of heritability by socioeconomic status: Initial studies reported that the heritability of intelligence was moderated by the family's socio-economic status (Turkheimer et al., 11/2003). However, heritability refers to the *proportion* of variance in the population explained by genetic differences. Hancscombe et al. (2012) re-investigated the question with unstandardized variance components. They found that the result was driven by differences in the environmental variance component: among high-status individuals, there was less variability in the environment. At the same time, the unstandardized amount of variance explained by additive genetics was similar across the range of socioeconomic status. While this pattern results in a higher heritability estimate, it should not be interpreted as increased importance of genes at higher levels of socioeconomic status—instead, it reflects how environmental *variation* is larger at lower levels of socioeconomic status.

Meta-analysis of standardised effect sizes. Meta-analysts commonly investigate standardised effect sizes, such as correlation coefficients, across studies. These may conflate differences related to the research question (differences in slope, in the magnitude of effects) with unrelated differences (differences in variances, e.g., due to measurement error in the outcome variable, range restriction in the predictor). This leaves researchers at risk of spurious inferences about effect size heterogeneity, publication bias, and between-study moderators (Wiernik & Dahlke, 2020). So should we instead aggregate unstandardised effects? Unfortunately, the lack of standardisation in psychology means that effects are difficult to bring to a single metric by means other than standardisation using the observed standard deviation and mean (but consider e.g., the percentage of the maximum possible, Cohen et al., 1999). Luckily, there are productive ways forward. Researchers can explicitly account for measurement error and selection effects (Wiernik & Dahlke, 2020). Furthermore, with well-validated instruments, such as IQ tests, standardisation with norm data (rather than with data from the sample which may vary in idiosyncratic ways) could avoid the problems discussed here. As a side effect, this would also allow researchers to quickly notice when a sample only covers a restricted range of values (e.g., because only psychology students were included).

Causal Identification of Interactions

Motivating Example

A group of researchers is interested in how a stress reduction program affects participants' subsequent subjective well-being. For this purpose, participants are randomly assigned to either participate in the treatment or in a control condition. Furthermore, the researchers are interested in how the intervention interacts with participants' personality, which they assessed before the intervention took place.

In their first analysis, they regress subjective well-being at the end of the study on (1) a binary indicator of whether or not participants were in the treatment condition, (2) participants' neuroticism, measured before the intervention took place and (3) the product of the two variables. Table 3, Analysis 1 shows the results from this analysis.

Table 3

Results from a Regression Analysis Predicting Subjective Well-Being from Treatment, Neuroticism and their Interaction, plus Gender (Analysis 2), plus the interaction between Treatment and Gender (Analysis 3)

	Analysis 1		Analysis 2		Analysis 3	
Variable	b	95% CI	b	95% CI	b	95% CI
Treatment yes/no	2.85	[2.68; 3.01]	2.88	[2.73; 3.03]	2.11	[1.92; 2.29]
Neuroticism	0.06	[-0.04; 0.15]	-0.09	[-0.18; -0.01]	0.05	[-0.04; 0.13]
Treatment*Neuroticism	0.38	[0.25; 0.51]	0.33	[0.21; 0.45]	-0.02	[-0.16; 0.11]
Female Gender	-	-	0.92	[0.77; 1.07]	0.06	[-0.14; 0.24]
Treatment*Female Gender	-	-	-	-	1.84	[1.57; 2.13]

Note. N = 1,000, data have been simulated.

These numbers suggest that the treatment interacts with the neuroticism of the treated individual, with bigger treatment effects among the more neurotic. However, the researchers are aware that women have reliably higher neuroticism than men, and one of them suggests that female gender should thus be statistically controlled for. The results from their second analysis can be found in Table 3, Analysis 2. The statistical evidence for the interaction remains mostly unaffected. Thus, they provisionally conclude that, even controlling for female gender, the treatment still has bigger effects among the more neurotic.

But then a colleague makes them aware of a blog post that highlights that interactions require “interaction controls” (Simonsohn, 2019; see also Yzerbyt et al., 2004), and so they dutifully run a third analysis. Lo and behold, their third analysis reveals an interaction between female gender and treatment, but the interaction with neuroticism that was initially of interest has disappeared (Table 3, Analysis 3).[8]

[8]Inclusion of the interaction between treatment and female gender also changed the main effects of female gender and neuroticism, which are not of central interest here. There is no simple interpretation for these coefficients—giving a substantive interpretation to the coefficients of confounders and modifiers in multiple regression analyses constitutes an instance of the so-called table 2 fallacy (Westreich & Greenland, 2013). However, given that these are simulated data, we can

The Question of Causality

The previous example hints at the existence of two different types of "interaction"—one that is causal in nature, and another one that is not. VanderWeele (2009) refers to them as interaction (the effect of one treatment is causally changed by another variable) and effect modification (the effect of the treatment co-varies with a third variable), but other terms have been used elsewhere. For example, psychologists may understand interaction to refer to both phenomena (as we did in this article, up to this point), and subsequently distinguish between (causal) moderation and statistical interaction. To us, the terminology seems less important than a clear understanding of the phenomena, so we are going to talk about "causal interaction" versus "effect modification" to maximize the distinction.

A causal interaction refers to a scenario in which (hypothetically) intervening on the third variable would change the effect of the treatment (see VanderWeele, 2009, for formalized definitions). This may often be the intended meaning when psychologists *hypothesize* interactions. In the example above, the researchers may have speculated that there is something about the treatment that makes it more effective for neurotic individuals, not because of their gender, but because of their neuroticism—it may target particular cognitive processes such as anxiety and worries that affect them more frequently.

Experiments are the ideal design to identify and test causal interactions. If both the treatment and the third variable have been manipulated by the researcher, a causal interpretation is warranted. Factorial experiments, which are frequently evaluated with the help of ANOVAs, neatly illustrate the symmetric nature of causal interactions. If two variables A and B interact causally, it is appropriate to state that the effect of A depends on the level of B, just as it is appropriate to state that the effect of B depends on the level of A.

---

explain their behavior. The data were simulated in a manner that (1) the effect of the treatment depends on gender, (2) gender has no effect beyond that interactive contribution, and (3) neuroticism is an outcome of gender but of no further relevance. Because analysis 2 omits the interaction between the treatment and gender, part of this interaction ends up in the coefficient of gender. Why is the coefficient of neuroticism negative, even though neuroticism is not the cause of anything in our data generating model? Controlling for gender, the meaning of neuroticism changes to "anything in neuroticism that is not determined by gender"—in our particular example, as we simulated no other causes of neuroticism, this is simply a random variable which we may call U. The interaction term between treatment and neuroticism will systematically overpredict outcome values for participants who are high on neuroticism for their gender—i.e., participants which happen to have a high value on U. To "compensate" for the overprediction, U—or simply "neuroticism" in the output—gets a negative coefficient. In general, understanding non-focal coefficients from multiple regression analyses is a non-trivial endeavor.

Of course, experimental investigations are not always feasible or even just possible, and so researchers might sometimes want to consider an interaction in which one or both of the variables were not randomized. In such a scenario, inferring a *causal* interaction is equal to inferring causation from correlation—an endeavor that heavily depends on domain expertise and additional assumptions (see, e.g., Rohrer, 2018 for an introduction). One step into this direction, as illustrated in the example above, consists of the control of third variables that confound the association between the non-randomized variable and the outcome. Interactive control is necessary; in practice the inclusion of multiple interactions at once can lead to unstable estimates—here, variable selection procedures can help (Blackwell & Olson, 2020).

Sometimes, researchers might not be primarily interested in *causal* interactions, but rather concerned about effect modification. In a clinical trial, it might be of interest to see how effective the treatment is within different subpopulations. For example, researchers might find out that the treatment works best among individuals with certain comorbidities, and that information might be helpful for treatment planning, regardless of whether it is the other condition or one of its causes that (causally) interacts with the treatment. Unlike causal interaction, effect modification can be asymmetric. In our example above, it is correct to say that neuroticism modifies the effect of the treatment. However, this does not mean that the treatment modifies the effect of neuroticism — in fact, the data have been simulated so that neuroticism has no causal effect on subjective well-being at all.

Considering the example above, if effect modification had been the only question, researchers might have stopped after the first analysis and concluded that neuroticism does indeed modify the effects of the treatment. Of course, it is unclear to which extent that information would have been useful in this particular case: If the goal of the analysis is substantive understanding, we need to figure out why effect modification occurs; and if the goal is the identification of subpopulations which would benefit most, gender can be assessed more economically than neuroticism.

Substantive Examples

Personality moderates the effects of mindfulness on well-being. Much in line with our simulated example, de Vibe et al. (2015) investigated whether personality (neuroticism and conscientiousness) moderates the effects of mindfulness training among

students. They found that the intervention reduced mental distress particularly well among students with higher scores on neuroticism. While their analysis contained control variables (gender and baseline values of the outcome), they did not include the interaction between those control variables and treatment, and thus failed to account for alternative explanations (e.g., treatment effect may vary depending on gender, treatment effect may vary depending on initial level of mental distress). Thus, only effect modification should be concluded. Nonetheless, the article prominently discussed interpretations of a causal interaction between neuroticism and treatment, such as differences in emotional reactivity.

Personality moderates the effects of cultural tightness on cultural adaptation. Geeraert et al. (2019) investigated how students participating in intercultural programs adapted culturally to their host countries. They found that adaptation was lower for host countries with tighter cultures (i.e., cultures in which norms are more rigidly imposed). But this effect was moderated by personality; for example, students scoring high on honesty-humility showed high cultural adaptation even in tight cultures. Their longitudinal analyses do not account for potential confounders between personality and cultural adaptation and thus, only effect modification may be concluded. Nonetheless, the discussion section invoked the fit between personality and social norms as an explanation, which clearly assumes a causal role of personality (i.e., if we could intervene on personality, we would expect that this has subsequent effects on adaptation to tight cultures). Noticeable, the authors also suggested that poor fit between students and host countries may be very costly, and that selection with an eye for personality fit might hence be sensible. This conclusion would be justified even in a scenario of “mere” effect modification in the absence of interactions, since personality may be able to predict how a student adapts to a certain culture *regardless* of whether it is the cause of adaptation.

Country-level gender quality moderates the effects of gender x age on self-esteem. Bleidorn et al. (2016) investigated cross-sectional age trajectories of self-esteem in a sample spanning 48 nations. Overall, they found that on average, men reported higher levels of self-esteem (i.e., a gender gap), and so did older individuals, with no significant interaction between age and gender on average. However, these associations significantly varied across countries. Thus, in exploratory analyses, the authors investigated whether country-level characteristics moderated gender specific age-trajectories (i.e., they investigated the triple-interaction term gender x age x country characteristic). They found that, in countries with higher gender equality, the gender gap in self-esteem shrank with age. While it may be tempting to give a substantive causal interpretation to this pattern—women

who, through increased gender equality, have better access to high status jobs and the political sphere, end up with higher self esteem—the authors pointed out that gender equality is highly correlated with other country characteristics, such as GDP per capita and the Human Development Index. Thus, it may only be justified to conclude effect modification; the actual cause of gender differences in age trajectories of self-esteem may lie in other factors.

Recommendations

It is very well possible that a variable modifies the effect of another one without a causal interaction between the two (as in the example above), and even that there is a causal interaction without effect modification (although this requires that different effects cancel each other out, see VanderWeele, 2009). Thus, for researchers to arrive at the right conclusion, it is important that they can distinguish between the two—and determine which one is relevant in a given situation.

If the purpose is to test a specific hypothesis derived from a theory, a helpful question to consider is “Would I expect that an intervention on the third variable changes the effect of the other variable?” In the case of our example: “Would an intervention that reduces neuroticism, such as psychotherapy, reduce the treatment effect of our stress reduction intervention?” If the answer is yes, a causal interaction is of interest, and if the third variable cannot be manipulated, all the concerns of causal inference on the basis of observational data apply (see, e.g., Rohrer, 2018).

Of course, not all interaction questions arise a priori, and sometimes a researcher might be confronted with the coefficient of a product term and struggle to find the right interpretation. Here, helpful questions to consider could be: “Assuming that the main effect of the third variable was of central interest, would the present study design allow me to interpret it as a causal effect under reasonable assumptions?” If the answer is yes, a causal interpretation may be warranted. If the answer is no, subsequent interpretations should take into account that one cannot conclude that a manipulation of the third variable would change the effect of interest.

Box 3: Interaction Issues Interact

The three issues we consider here cannot be considered in isolation. When the issues interact and models become more complex, formal models, data simulation, and visualization are even more helpful. To give an example of how these issues interact, if we want to correctly estimate effects on residual variance, we cannot ignore that our measure has a ceiling and a floor. If we did, we would underestimate the residual variance whenever a value is close to the scale's lower or upper limit, because values can only deviate in one direction. This could result in an illusory effect on residual variance, if, for instance, the investigated moderator has a strong main effect, driving values to the limit. New statistical software, such as brms, makes it easy to formulate the appropriate distributional models (Bürkner, 2020) with small changes (e.g., it only takes a small tweak to switch from predicting the residual standard deviation in a Gaussian regression to predicting the discrimination parameter in an ordinal regression, which inversely relates to the standard deviation of the latent variable).

To give another example, if we would adjust our interaction effect for a potential confounder to better fit our causal model, we should also adjust for the confounder when estimating effects of the moderator on residual variance. Again, it is easy to estimate multiple effects on the residual standard deviation in brms, but difficult to do the same in the more established framework of correlational analysis.

In the introduction, we wrote that questions of nonlinearity of interactions can be resolved by careful examination of the data, whereas the issues we discuss here often require us to interrogate our assumptions. However, if we observe only subsets of the data, for example, if we have an old and a young cohort in our study that were recruited in different ways, we could run into the question of causal interaction or effect modification. If we observe that age effects on our outcome of interest are flattened in the older group is it because recruitment methods act as a moderator? Or is the effect of age simply nonlinear but the two parts of the curve we can see look straight? Such problems are another common version of the question of causal interaction or effect modification.

Finally, questions of scale dependence and linearity of effects also come up when predicting distributional parameters other than the mean, such as the residual standard deviation. Again, brms allows for distributional assumptions and link functions to be changed and for

nonlinear effects on all distributional parameters to be estimated using, for instance, thin-plate splines (Bürkner, 2020).

Conclusion: Better Answers, Better Questions

In this manuscript, we have discussed three issues. First, conclusions about interactions depend on scaling decisions, and flawed measures can lead to spurious interactions. Second, moderation of slopes is not the same as moderation of correlations. Third, effect modification is not the same as causal interaction. If researchers are not aware of these distinctions, they might accidentally analyse the data in a manner that returns the technically correct answer to the wrong question (Hand, 1994). This disconnect between research questions and statistical analyses can result in misled conclusions.

These issues may seem daunting, and after considering how these issues interacted (see Box 3) one may be tempted to conclude that the best course of action is to stop investigating interactions altogether. Some weaker variation of this notion may be defensible. Many psychologists seem enthusiastic about increasingly complex claims about interactions ("boundary conditions"), mediation ("processes", "mechanisms"), and any combination of the two. This enthusiasm should be tempered: Complex claims require very large samples and strong designs; they come with methodological complications like the ones outlined above; and they require reliable knowledge about more basic aspects (e.g., measurement properties, response biases, main effects). Our enthusiasm for complex claims may actively hinder the quest for such reliable knowledge: Researchers are disincentivized from conducting "less exciting", "less novel" basic research, and thus it is quite possible that we end up building on sand.

At the same time, we concur with Lakens and Caldwell (2019) that there are benefits of examining interactions. For example, they entail risky predictions that allow for particularly informative tests of competing theories, and they may help address issues of generalizability across different populations. Thus, instead of putting interaction research on hiatus, we should strive for improved interaction research.

Approaching the issue from the empirical side, researchers should pay closer attention to the details of their data and their model. Classical regression diagnostics (Belsley et al., 1980), such as plotting fitted values against residuals, are often taught, occasionally practiced, and rarely reported. In a Bayesian workflow, these diagnostics can be seen as special cases of posterior predictive checks (generating data from the model and

comparing it to the real data, see Figure 5, (Gabry & Mahr, 2018; Gelman et al., 2020). Such checks can uncover where the model falls short, such as assuming homoskedasticity, or ignoring ceiling and floor effects. Research articles frequently only report estimates from linear models and select simple slopes graphs, but this leaves readers in the dark about the details of the data. Following the recommendations by (McCabe et al., 2018), raw data and estimates should instead be depicted in so-called small multiples with individual plots for several simple slopes. Online supplements also make it possible to habitually share diagnostic plots, especially for complex models where simply graphing the raw data is insufficient to evaluate the model.

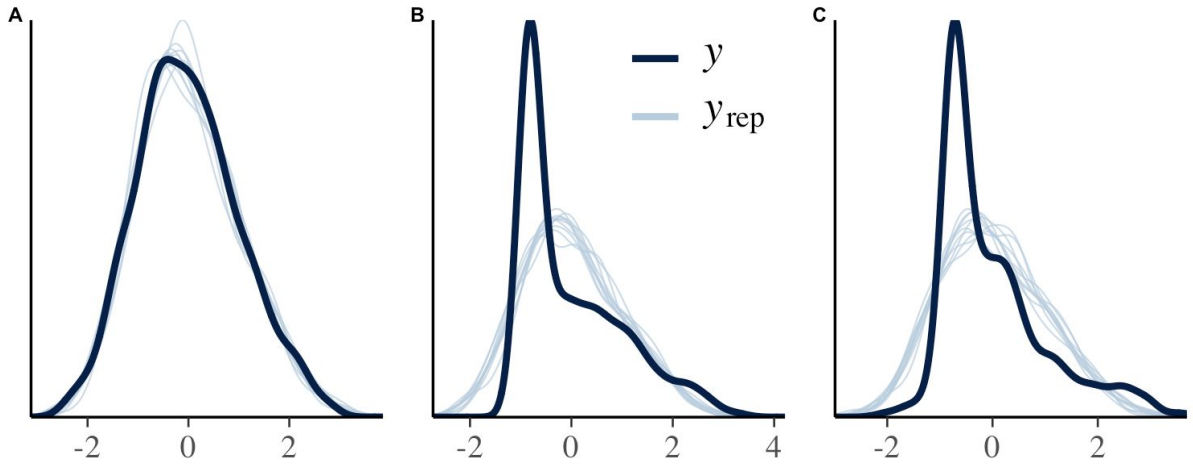

Figure 5. Posterior predictive checks for a simple interaction model based on three different data-generating processes. The distribution of the outcome is shown along with model-predicted distributions for ten samples. Panel A: The model distribution shows a good fit to the real data. Panel B: A floor effect is apparent in the real data, but not part of the model. Panel C: Heteroskedasticity makes for an awkward fit between real data and model predictions.

As a pedagogical tool, we have generated Figure 6, a triple triptych in homage to Anscombe's quartet (Anscombe, 1973). Every row represents a different interaction scenario resulting in identical simple slopes plots (black lines) and linear regression results. The first row shows the "ideal" scenario, which many researchers will naively assume: only the slopes differ by moderator level. The middle row shows how a floor effect in the measurement scale can cause the illusion of an interaction, even though only main effects were simulated at the latent level. The third row shows a scenario in which slopes and correlations exhibit reverse patterns: while the slope increases with higher moderator values, the correlation decreases

(i.e., in the low moderator panel the values scatter the least from the regression line). Importantly, a researcher who only runs a linear model and reports the resulting coefficients and simple slopes would not be able to distinguish between these scenarios, although they lead to different substantive conclusions.

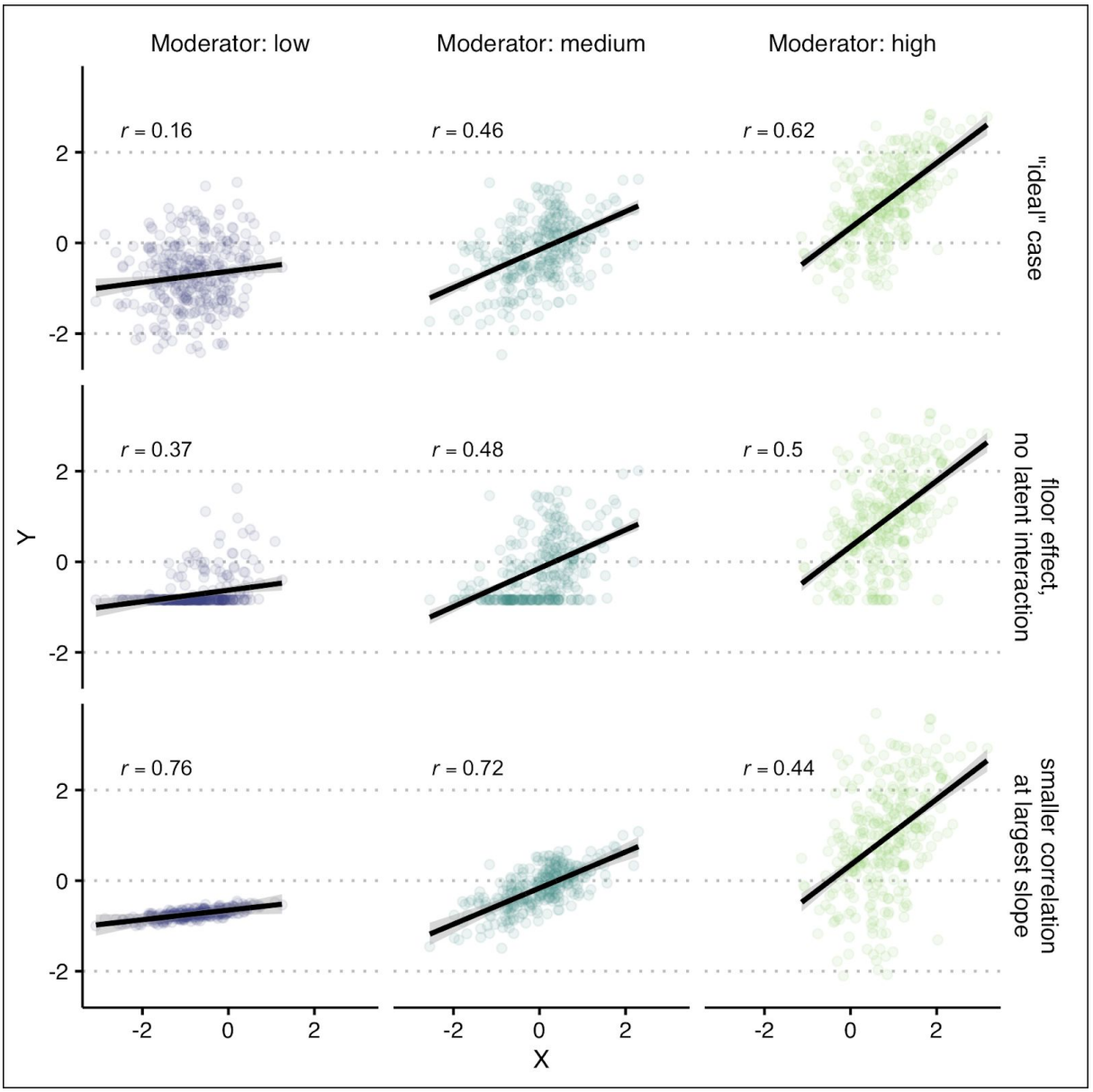

Figure 6. Columns show levels of the moderator variable (M). X and M are the same across rows, whereas Y has been generated according to different scenarios but maintains the same mean and variance. The slopes estimated according to a simple linear interaction model (black lines) are the same across rows, but graphing the raw data shows that the data-generating processes were quite different. Shaded regions around the regression lines show 95% credible intervals.

Careful consideration of the details of the data is important, but as we have noted in the beginning, it is not sufficient to solve the issues highlighted in this manuscript. No amount of graphing can answer questions about scaling assumptions or causality, or tell us what exactly our research question is and how it could be tested. This leads us to a broader underlying issue.

In an idealized scenario, one may start with a substantive research question and then choose the appropriate statistical analysis. One may mistakenly pick the wrong analysis, but course corrections are possible as the goal of the analysis is clear. If the research question was formulated as a generative model, finding flaws in the analysis strategy is even possible before data collection. In our experience, the actual research process often works quite differently. The research question is rather vague to begin with ("How do X and Y affect Z?"), and statistical analyses are chosen for a variety of reasons (e.g., domain norms, familiarity, publishability, implementation in popular statistical packages), but not for their capability of providing appropriate answers.

So the necessary course corrections may be much broader, as the underlying problem concerns our research questions and theories (Hand, 1994; Muthukrishna & Henrich, 2019). Psychological theories are often vague verbalizations that accommodate many different readings and corresponding statistical models. Researchers might "theorize" that one construct interacts with another one but leave open what pattern is to be expected. Arguments about the empirical support for such a vague hypothesis are futile as it is not even established what exactly is being predicted. Thus, ultimately, some broader rethinking of the field may be necessary, with a stronger focus on formal modeling (Guest & Martin, 2020; McElreath, 2020; Smaldino, 2017), more rigorous theorizing, and more precise research questions.

References

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, *80*(1), 123–129. https://doi.org/10.1016/S0165-1765(03)00032-6

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician*, *27*(1), 17–21. https://doi.org/10.1080/00031305.1973.10478966

Beck, E. D., & Jackson, J. J. (2020). *A Mega-Analysis of Personality Prediction: Robustness*

and Boundary Conditions. https://doi.org/10.31234/osf.io/vsm9y

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley. https://openlibrary.org/books/OL4415962M/Regression_diagnostics

Blackwell, M., & Olson, M. (2020). *Reducing model misspecification and bias in the estimation of interactions*. Working Paper available at https://mattblackwell. org/files/papers/lasso …. https://mattblackwell.org/files/papers/lasso-inters.pdf

Bleidorn, W., Arslan, R. C., Denissen, J. J. A., Rentfrow, P. J., Gebauer, J. E., Potter, J., & Gosling, S. D. (2016). Age and gender differences in self-esteem—A cross-cultural window. *Journal of Personality and Social Psychology*, *111*(3), 396–410. https://doi.org/10.1037/pspp0000078

Bond, T. N., & Lang, K. (2019). The Sad Truth about Happiness Scales. *The Journal of Political Economy*, *127*(4), 1629–1640. https://doi.org/10.1086/701679

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. In *Journal of Statistical Software* (Vol. 80, Issue 1, pp. 1–28). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. In *The R Journal* (Vol. 10, Issue 1, pp. 395–411). https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2020). *Estimating Distributional Models with brms*. CRAN. https://cran.r-project.org/web/packages/brms/vignettes/brms_distreg.html

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, *34*(3), 315–346. https://www.tandfonline.com/doi/abs/10.1207/S15327906MBR3403_2?casa_token=syB

4GlwxZyYAAAAA:6ts6yu90Qs8RO13_rp10WUBppQkg1FGLK4dh9SmE0QcDhvc9gI51M2Z0CsJE2bPEPDw7TVi5VTecuA

de Vibe, M., Solhaug, I., Tyssen, R., Friborg, O., Rosenvinge, J. H., Sørlie, T., Halland, E., & Bjørndal, A. (2015). Does Personality Moderate the Effects of Mindfulness Training for Medical and Psychology Students? *Mindfulness*, *6*(2), 281–289. https://doi.org/10.1007/s12671-013-0258-y

Gabry, J., & Mahr, T. (2018). *bayesplot: Plotting for Bayesian Models*. https://CRAN.R-project.org/package=bayesplot

Geeraert, N., Li, R., Ward, C., Gelfand, M., & Demes, K. A. (2019). A Tight Spot: How Personality Moderates the Impact of Social Norms on Sojourner Adaptation. *Psychological Science*, *30*(3), 333–342. https://doi.org/10.1177/0956797618815488

Gelman, A. (2018). *You need 16 times the sample size to estimate an interaction than to estimate a main effect [blog post]*. https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. In *arXiv [stat.ME]*. arXiv. http://arxiv.org/abs/2011.01808

Gerlach, T. M., Arslan, R. C., Schultze, T., Reinhard, S. K., & Penke, L. (2017). Predictive Validity and Adjustment of Ideal Partner Preferences Across the Transition Into Romantic Relationships. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000170

Giner-Sorolla, R. (2018). Powering your interaction. In *Approaching significance*. https://approachingblog. wordpress. com/2018/01/24/powering-your-interaction-2

Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science*. https://psyarxiv.com/rybh9/download?format=pdf

Hainmueller, J., Mummolo, J., & Xu, Y. (2016). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, 1–30. https://www.cambridge.org/core/journals/political-analysis/article/how-much-should-we-trust-estimates-from-multiplicative-interaction-models-simple-tools-to-improve-empirical-practice/D8CAACB473F9B1EE256F43B38E458706

Hand, D. J. (1994). Deconstructing Statistical Questions. *Journal of the Royal Statistical Society. Series A,* , *157*(3), 317–356. https://doi.org/10.2307/2983526

Hanscombe, K. B., Trzaskowski, M., Haworth, C. M. A., Davis, O. S. P., Dale, P. S., & Plomin, R. (2012). Socioeconomic Status (SES) and Children's Intelligence (IQ): In a UK-Representative Sample SES Moderates the Environmental, Not Genetic, Effect on IQ. *PloS One*, *7*(2), e30320. https://doi.org/10.1371/journal.pone.0030320

Johnson, W. (2007). Genetic and environmental influences on behavior: Capturing all the interplay. *Psychological Review*, *114*(2), 423–440. https://doi.org/10.1037/0033-295X.114.2.423

Lakens, D. (2020). *Effect Sizes and Power for Interactions in ANOVA Designs*. https://daniellakens.blogspot.com/2020/03/effect-sizes-and-power-for-interactions.html

Lakens, D., & Caldwell, A. R. (2019). *Simulation-based power-analysis for factorial ANOVA designs*. https://psyarxiv.com/baxsf/download?format=pdf

Lakens, D., & DeBruine, L. M. (2020). Improving Transparency, Falsifiability, and Rigour by Making Hypothesis Tests Machine Readable. *Advances in Methods and Practices in Psychological Science*.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312–319. https://doi.org/10.3758/BF03197461

McCabe, C. J., Halvorson, M. A., King, K., Cao, X., & Kim, D. (2020). *Estimating and interpreting interaction effects in generalized linear models of binary and count data*. https://psyarxiv.com/th94c/download?format=pdf

McCabe, C. J., Kim, D. S., & King, K. M. (2018). Improving Present Practices in the Visual Display of Interactions. *Advances in Methods and Practices in Psychological Science*, *1*(2), 147–165. https://doi.org/10.1177/2515245917746792

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. CRC Press. https://play.google.com/store/books/details?id=6H_WDwAAQBAJ

Murray, A. L., Molenaar, D., Johnson, W., & Krueger, R. F. (2016). Dependence of Gene-by-Environment Interactions (GxE) on Scaling: Comparing the Use of Sum Scores, Transformed Sum Scores and IRT Scores for the Phenotype in Tests of GxE. *Behavior Genetics*, *46*(4), 552–572. https://doi.org/10.1007/s10519-016-9783-5

Mustillo, S. A., Lizardo, O. A., & McVeigh, R. M. (2018). Editors' Comment: A Few Guidelines for Quantitative Submissions. *American Sociological Review*, *83*(6), 1281–1283. https://doi.org/10.1177/0003122418806282

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Open Science Collaboration, Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., … Zuni3, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 10. https://doi.org/10.1126/science.aac4716

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological*

Science, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rohrer, J. M., & Schmukle, S. C. (2018). Individual Importance Weighting of Domain Satisfaction Ratings does Not Increase Validity. *Collabra. Psychology*, *4*(1). https://doi.org/10.1525/collabra.116

Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020). Coronavirus pandemic (COVID-19). *Our World in Data*.

Rothman, K. J., Greenland, S., & Walker, A. M. (1980). Concepts of interaction. *American Journal of Epidemiology*, *112*(4), 467–470. https://doi.org/10.1093/oxfordjournals.aje.a113015

Simonsohn, U. (2017, February 23). *[57] Interactions in Logit Regressions: Why Positive May Mean Negative*. Datacolada. http://datacolada.org/57

Simonsohn, U. (2019). *[80] Interaction Effects Need Interaction Controls*. http://datacolada.org/80

Smaldino, P. E. (2017). Models are stupid, and we need more of them. *Computational Social Psychology*, 311–331. https://books.google.de/books?hl=en&lr=&id=gjwlDwAAQBAJ&oi=fnd&pg=PA311&dq=smaldino+models+are+stupid&ots=K1Lzti0em8&sig=yMgbjFYIDlINBLapHZ2Pmzr2v3w

Smithson, M. (2012). A simple statistic for comparing moderation of slopes and correlations. *Frontiers in Psychology*, *3*, 231. https://doi.org/10.3389/fpsyg.2012.00231

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. I. (11/2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science*, *14*(6), 623–628. https://doi.org/10.1046/j.0956-7976.2003.psci_1475.x

Umlauf, N., & Kneib, T. (2018). A primer on Bayesian distributional regression. *Statistical Modelling*, *18*(3-4), 219–247. https://doi.org/10.1177/1471082X18759140

VanderWeele, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology* , *20*(6), 863–871. https://doi.org/10.1097/EDE.0b013e3181ba333c

von Hippel, P. T., & Hamrock, C. (2019). Do test score gaps grow before, during, or between the school years? Measurement artifacts and what we can know in spite of them. *Sociological Science*, *6*, 43–80. https://www.sociologicalscience.com/articles-v6-3-43/

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: a survey of the field 33 years after Loftus. *Memory & Cognition*, *40*(2), 145–160. https://doi.org/10.3758/s13421-011-0158-0

Wiernik, B. M., & Dahlke, J. A. (2020). Obtaining Unbiased Results in Meta-Analysis: The Importance of Correcting for Statistical Artifacts. *Advances in Methods and Practices in Psychological Science*, *3*(1), 94–123. https://doi.org/10.1177/2515245919885611

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology*, *40*(3), 424–431. https://www.sciencedirect.com/science/article/pii/S0022103103001598