This version: 2024 06 11
Latest version: https://urisohn.com/46

GAMify Spotlight & Floodlight: A More Robust and Informative Approach to Probing Interactions

Andres Montealegre*
Cornell University
am2849@cornell.edu

Uri Simonsohn*
ESADE Business School
urisohn@gmail.com

Abstract.
Interactions, where a variable moderates the association between other variables, are central to marketing research. The authors re-analyze data from four recent marketing papers to demonstrate that the traditional approach used for probing interactions—linear regression followed by Simple-Slopes ("spotlight") and/or the Johnson-Neyman procedure ("floodlight")—is invalid when true relationships are not linear. They propose relying on GAMs (generalized additive models) as an alternative to linear regression, showing that GAM Simple Slopes and GAM Johnson-Neyman are just as straightforward, interpretable, and statistically powerful as are their linear counterparts, but GAM leads to more informative, robust, and sensical results. The accompanying new R package "interacting" makes GAM probing as easy as "*interprobe(x,z,y)*".

Keywords: interactions, spotlight, floodlight, GAM, simple slopes, Johnson-Neyman, regression, misspecification

* Acknowledgements: We thank Aleksandra Lyubarskaja for excellent research assistance.

Data and code to reproduce all results are available from:
https://researchbox.org/2859 (use code QAMDS)

Interactions, where a variable modifies the association between two (or more) other variables, are commonly examined in marketing papers. Researchers seek to answer questions like, does this consumer trait, or that product characteristic, make people more price sensitive? After testing interactions, it is common to 'probe' them, assessing the estimated relationship between the focal predictor x (e.g., price) and the dependent variable y (e.g., quantity purchased) at different values of the moderator (e.g., for participants at +1 SD on a "spendthrift-tightwad" scale), or assessing for which moderator values the effect of the focal variable is positive vs. negative, or statistically significant vs not. These approaches to probing interactions are known as Simple Slopes (Aiken and West 1991) and the Johnson and Neyman (1936) procedure respectively, and in marketing articles they are sometimes referred to as "spotlight" and "floodlight" analysis (Spiller et al. 2013).

The technology that is currently relied upon for studying interactions involves estimating a linear regression that includes the product of the two variables as a predictor (e.g., $y = a + bx + cz + dx \cdot z$), and then probing the interaction based on the regression coefficients that include the focal predictor x ($\hat{b}$ and $\hat{d}$). This technology is about 90 years old, dating back to Johnson and Neyman (1936). It is nevertheless still popular enough in marketing articles to make the tutorial by Spiller et al. (2013), the second most cited JMR paper published since 2010.[1]

In this article we propose that marketing researchers cease to rely on these nonagenarian (linear) Simple Slopes and (linear) Johnson-Neyman procedures to probe interactions. The reason for this proposal is that the estimates obtained with these techniques are too sensitive to the arbitrary and often false assumption that all effects in the model are linear. Coming back to our opening example, the linear regression assumes that the relationship between price and quantity

[1] A query on the Web-of-Science on April of 2024, for articles published in JMR since 2010 shows, the article by Berger and Milkman (2012) at the top, with 1526 citations, followed by Spiller et al. with 1223.

bought is constant, forming a perfectly straight line. However, both psychology (e.g., through diminishing sensitivity and satiation) and economics (e.g., through diminishing marginal utility) suggest that this relationship might not be linear. For instance, if the price of tacos drops from $9 to $7 Alex may buy 5 instead of 4 tacos. Yet, if the price drops further to $5 and then $3, we don't expect Alex to continue buying one more taco for every $2 decrease (e.g., buying 6 instead of 5 at $5, and 7 instead of 6 at $3). At some point Alex is ready for dessert regardless of the price of the next taco. We should be skeptical of a perfect straight line connecting price and quantity purchased. We should be skeptical of perfect straight lines connecting any two variables in general.

Fortunately, there is an easy way to relax this linearity assumption, it involves relying on a "Generalized Additive Model" (GAM) instead of a linear regression to model the relationships of interest. GAMs have been around for nearly four decades (Hastie and Tibshirani 1987), but have not been used much in social science. GAMs estimate rather than (arbitrarily) assume the functional form of all variables in the model. We propose that henceforth marketing researcher probe interaction relying on GAM Simple Slopes and GAM Jonson-Neyman procedures (Simonsohn 2024). GAM based procedures are closely related, similarly interpretable, similarly (statistically) powerful, but immensely more robust alternatives to linear Simple Slopes and linear Johnson-Neyman procedures.

We make the case for the switch to GAMs by revisiting four recently published articles in top marketing journals (JMR, JCR, and JCP), where authors relied on linear Simple Slopes and/or linear Johnson-Neyman procedures to probe interactions. We reanalyze the data the authors posted, and upon reproducing the published results, we demonstrate how the inferences based on the (linear) interaction results are partially or entirely reversed once the arbitrary linearity assumption is relaxed by using GAM instead of linear procedures.

Example 1 involves a case where an effect that supposedly is present for low values of the moderator, is actually only present for high values of it. Example 2 involves a case where the linear estimates bear almost no resemblance to reality. Example 3 involves a case where the linear model dramatically underestimates the magnitude of the interaction effect and obtains a statistically significant effect of the wrong sign. Example 4 involves a case where an effect that supposedly reverses merely attenuates.

Part of the appeal of the currently universally-used linear approach is its simplicity. For example, the highly cited tutorial by Spiller et al. (2013) explained how authors could probe interactions by subtracting a constant from the focal variable of interest, and then re-estimating the same regression model used to *test* for the interaction. Fortunately, switching to GAM-based procedures will not increase the difficulty of probing interactions for researchers. We have created an R package, `*interacting*`, which computes both GAM Simple Slopes and GAM Johnson-Neyman curves, producing a ready-to-publish figure with both plots. All researchers need to do is run this command: *interprobe(x,z,y).*

In what follows we begin by briefly reviewing the literature on linearity and interactions. We then provide a brief introduction to GAMs, and then proceed to the core of the paper, contrasting linear with GAM probing procedures for studies in four recently published marketing papers.

Prior Work on Nonlinearities And Probing Interactions

The concern that motivates this article, nonlinearities invalidating the probing of interactions, is not new. Even in the original paper by Johnson and Neyman (1936) they write "We do not think it entirely correct to assume that the regressions . . . are represented by planes. On the

contrary we see good reasons to assume that these regressions are skew. However, we assume linearity as a first approximation, hoping that **in the future** we shall be able to consider the problem more fully following the same method of approach." (p.83; bold added). We may not be Johnson and Neyman, but fortunately, we do live in the future.

In the 1990s a few articles in psychology and management journals discussed the problem that if the x and z predictors in x·z are correlated, the interaction test is biased in the presence of nonlinearities (Cortina 1993; Ganzach 1997; Ganzach 1998; Lubinski and Humphreys 1990). These authors proposed adding quadratic terms (x^2 and z^2) to fix the *test* of the interaction. These authors did not assess, however, the impact of nonlinearities on *probing* interactions, nor the potential for quadratic controls to improve on such dimension.

A recent political science article by Hainmueller et al. (2019) discusses two problems with how interactions are probed in that field: projecting estimated effects to regions with little or no data, and the potential presence of nonlinear interactions. They propose assessing the robustness of Johnson-Neyman type estimates with what they term a "binned estimator" and a "kernel estimator". Unfortunately, Hainmueller et al. (2019) do not build on (or cite) the 1990s articles that explained the problem with correlated x and z predictors and it turns out that the problems explained in those articles also apply to, and also invalidate, the estimators proposed by Hainmueller et al. Finally, and most relevant for this article, a recent paper in psychology covers both the testing and probing of interactions in the presence of nonlinear effects, proposing relying on GAM to test and GAM Simple Slopes to probe interactions (Simonsohn 2024). The current article builds on that work by (i) developing the GAM Johnson-Neyman procedure,[2] (ii) providing

[2] The GAM Johnson-Neyman is vaguely mentioned and only in an online supplement by Simonsohn (2024). What to plot in the axes, how to handle categorical vs continuous predictors, and how to compute standard errors, is not discussed in that article but necessary for implementation.

an R package ('interacting') to compute and visualize GAM Simple Slopes and GAM Johnson-Neyman, and (iii) demonstrating the practical relevance of making the switch from linear to GAM probing for marketing researchers by re-analyzing four recently published marketing papers.

A Brief Introduction To GAM

In this section we provide a basic introduction to Generalized Additive Models (GAMs), seeking only to explain what it does in broad terms so that readers can conceptualize what GAM Simple Slopes and GAM Johnson-Neyman involve, but without covering technical details. Readers wishing to learn more about GAMs may consult the longer introduction by Simonsohn (2024), the textbook by Wood (2017), the original paper introducing GAMs by Hastie and Tibshirani (1987), or the introductory article by Beck and Jackman (1998).

GAM vs. Regression

As mentioned in the introductions, GAMs resemble linear regressions in that they also estimate the relationship between a dependent variable and a set of independent variables. But, while regressions assume all predictors are linearly associated with the (sometimes latent) dependent variable, GAMs *estimate* the functional forms of each independent variable.[3]

In our canonical interaction model, for instance, instead of estimating four coefficients, a,b,c,d in the linear regression, $y = a + bx + cz + dx \cdot z$, a GAM *estimates* functional forms for each predictor and possibly their interaction, as in $y=f_1(x)+f_2(z)+f_3(x,z)$. GAMs estimates f_1, f_2, and f_3 combining a series of base functions (e.g., log, polynomials, etc.), and allowing that combination to change for different x values. So, for instance, the GAM estimation may result in fitting a log function in the lowest third of the x-range, fitting a combination of polynomials for higher values of x, and a flat horizontal line for higher values still.

To avoid over-fitting, GAMs include a penalty for wiggliness, seeking smooth rather than rugged fits to the data. In practice this means that if a function is summarized "well enough" by a linear model, GAM will output a linear model, but if it requires a polynomial, or a log, it will output that instead. The main downside of GAMs is that they do not produce few and easy to interpret coefficients, instead they produce many uninterpretable coefficients (loosely speaking, the weights given to each underlying base function to form the estimated functional form).

This interpretability challenge seems to have prevented GAM from becoming a main tool in our social science toolbox so far, but the challenge is rather easy to circumvent by *probing* the models estimated by GAM in a manner that is analogous to how we probe interactions from linear

[3] Note that in a regression that includes non-linear terms, e.g., $y=ax+bx^2+e$, the regression is still linear in the sense that the effect of an increase of x^2 by 1 is always $\hat{b}$.

models. That is, by producing predicted values for the dependent variable, and predicted marginal effects from the model, and then plotting those predicted values and marginal effects, just like linear Simple Slopes and linear Johnson-Neyman curves currently do for linear regressions.

Illustration of GAM estimation

We report results from a simple simulation below to illustrate the greater accuracy and identical interpretability of GAM models. We produced 1000 observations for the random variable *x* (uniform 0 to 100) and consider three possible functional forms, ranging from linear to non-monotonic. Figure 1 shows that while the regression line is perhaps a useful summary of the average association, only GAM provides an adequate characterization of the relationship between x and y overall, and a precise estimate of the effect of x on y for specific values of x. For example, in the second panel, the linear regression misses the fact that once x reaches 50, it no longer impacts y, and in the third panel it misses the fact that the effect of x on y switches sign within the observed data.

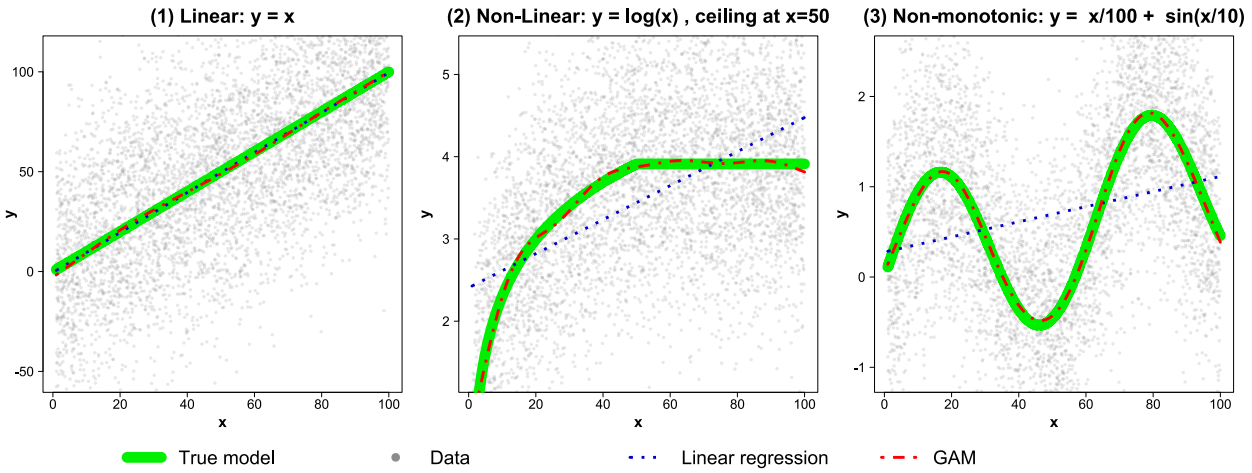

Figure 1. *Comparing Model Adequacy: GAM vs. Linear Regression.*
The figure depicts a single simulation of 1000 observations where x~U(0,100). The figure shows that when the true model is linear GAM recovers a linear model, but when it is not, it accurately captures the alternative functional forms. In all models, normal random error with the same SD as that produced by x on y, is added to y.
Code to reproduce figure: https://researchbox.org/2859/1 (enter code QAMDS)

These kinds of shortcomings for the linear model are not just a theoretical possibility. In the sections that follow, we present examples from recently published marketing papers exhibiting precisely these kinds of shortcomings.

Example 1: Misplaced Moderation

Mecit et al. (2022) propose that describing a disease with feminine grammatical terms, rather than masculine grammatical terms, leads to lower danger-perception of that disease. They focus on Spanish and French speakers' perceptions of dangers associated with COVID, because in those languages COVID can be described with either grammatical term. Specifically, "the name of the disease resulting from the virus (COVID-19) is grammatically feminine, whereas the virus that

causes the disease (coronavirus) is masculine." (their abstract). In Spanish, for instance, it is "el" coronavirus but "la" COVID.

In their Study 3, N=305 French speakers were randomly assigned to have the disease described with the feminine vs masculine terms and "answered five questions concerning their current danger perceptions about COVID-19 (e.g., how threatened do you feel, how difficult is it to eradicate)" (p.320). After a filler task they completed a 24-item gender stereotype questionnaire to measure "chronic gender stereotyping" (p. 321), which was used as a moderator in the analysis.

The authors find a significant interaction between the manipulation and chronic gender stereotyping (p = .0015), and probing this interaction linearly they find that the effect of the manipulation is significant for participants with a chronic gender stereotype above 0.53.[4]

We obtained the data posted by the authors (https://osf.io/9437y) and successfully reproduced the results.[5] The original paper plots (linear) Simple Slopes for the subset of data with less extreme values of the moderator, ranging between 0 and 1.5 (between the 27th and 82nd percentiles of the moderator; its full range in the data spans from -1.75 to +4.75). See left panel of Figure 2.

[4] The authors report p<.001 for the interaction (p. 321), but re-analyzing their data we obtain p = .0015. The authors report only the point at which the effect is significantly positive. It is significantly negative for moderator values below -1.65.

[5] The authors also report results for another dependent variable: (gender) stereotypical judgments about the virus e.g., in a bipolar scale how weak/strong, passive/aggressive it is. We reproduce the regression results for it as well and for this variable the GAM and the linear models arrive at more consistent conclusions, although, the linear Johnson-Neyman produces a significant reversal for low enough (and quite rare) values of the moderator while the GAM Johnson-Neyman does not.

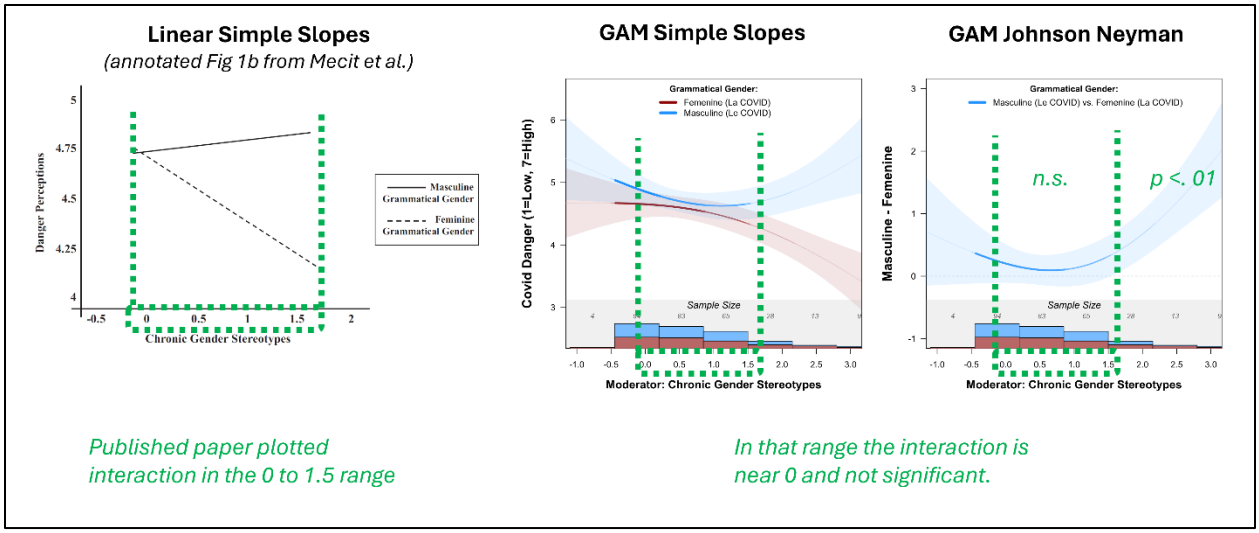

Figure 2. *Interaction is non-significant in range of values plotted in original paper.*
Notes. Reanalysis of Study3 by in Mecit et al. (2022) on danger perception of COVID by French speakers (N=302) when relying on masculine "le" COVID vs feminine "la" COVID grammatical terms. Left panel shows the (linear) Simple Slopes plot included in the original paper, it only depicts the interaction for moderator values around 0 to 1.5. Panels B and C show that after relaxing the linearity assumption there is no interaction in that range. The overall interaction is driven by participants with more extreme moderator values, above 1.5.
Code to reproduce middle and right panels in figure: https://researchbox.org/2859/7 (enter code QAMDS)

The middle and right panels of Figure 2 report results when GAM probing the data. The confidence band for the GAM Johnson-Neyman curve includes zero in the aforementioned range [0-1.5], and it is associated with a significant effect only in the [1.53 to 4.75] range. In other words, the effect of the manipulation is only significant for participants with moderator values that are more extreme than those plotted in the original paper. [6] The original linear analysis over-estimates the effect among participants with moderate values by taking the larger effect produce by the more *extreme* participants, and distributing that effect linearly across *all* observations. A linear model cannot accommodate a nonlinear effect. GAM, in contrast, allows for changes of functional form

[6] To be clear, we are not interpreting the non-significant effect below 1.5 as accepting the null. For instance, when the moderator, chronic gender stereotyping, equals 1, the estimated effect of the manipulation is a drop of the dependent variable by -0.13, with a confidence interval which does not rule out values up to -0.41. Because the SD of the dependent variable is about 1, that's a Cohen-d of about .4. The data, then, do not rule out effects of a considerable magnitude. At the same time, they do not rule out zero.

across moderator values, and thus does minimal projection and misplacement of effects from one region of values to the other.

We checked the robustness of these conclusions by running a *linear* regression that allowed for different slopes for moderator values below vs above 1.5 (this is in the spirit of the two-lines test for U-shapedness by Simonsohn (2018)).[7] Consistent with the GAM results, we found that for moderator values below 1.5 the interaction of the moderator and the gender treatment is small and not significant, while above it is (five times) larger and significant ($\hat{b}_{\text{below 1.5}}$ = -.323, p = .417, and $\hat{b}_{\text{above 1.5}}$ = -1.80, $p < .001$).[8] In other words, the GAM conclusions are robust to not relying on GAM.

In terms of the implications of these differences in results, we begin by stressing that the overall conclusion of the original study in particular, and the paper more generally, holds when linearity is relaxed: that is to say, also in our analysis the same virus was perceived as less dangerous when referring to it with feminine rather than masculine grammatical terms. The evidence, however, is interestingly updated when we stop imposing the linearity assumption. First, the effect is driven by people with more extreme values of sexism. This is especially relevant if one were theoretically interested in more common and possibly implicit levels of sexism, vs more extreme and overt levels. It hints at a potentially different type of effect. Second, from a purely descriptive perspective, a smaller share of the participants exhibit the effect of interest. While 52% of participants exhibited sexism levels high enough to imply they showed the effect with the linear model (moderator above 0.53), only 17% did when relying on the GAM model (moderator above

[7] The regression we estimated, in R syntax, is, lm(y~x*z*z_high) where x is the focal predictor, z is the moderator, and z_high is a dummy which equals 1 when z>=1.5 and 0 otherwise.

[8] We restricted the sample for this regression to include only positive moderator values to avoid biasing the estimate of the interaction by including observations with potentially opposite bias. If we include all observations, our point only gets stronger, as the interaction with moderator values in the original articles is even closer to zero: $\hat{b}$ = -.013, p = .955. See R Script https://researchbox.org/2859/30/ (Section #7.1).

1.53). This may be important if the goal is to understand typical rather than atypical effects. Third, because the effect is driven by more extreme observations, a closer look at potential measurement issues with those participants (inattention, demand effects, tended to give high answers to every question, etc.) may be justified. We are not proposing those issues are indeed a concern, we don't know, we are suggesting the data more clearly suggest checking whether they are, once one knows the locus of the effect.

Example 2: Timing is Everything (Except Linear)

Zor, Kim, and Monga (2022) propose that time of day predicts the kinds of tweets that people engage with, specifically, that "as morning turns to evening" engagement in social media shifts from virtue (e.g., liking a tweet from "The Atlantic") to vice (e.g., liking a tweet from "Vanity Fair").

In their Study 1A they analyze 176,390 tweets from eights magazine accounts, four accounts of "virtue magazines" (The Atlantic, Forbes, Health and The New Yorker) and four accounts of "vice magazines" (Cosmopolitan, Entertainment Weekly, People and Vanity Fair).

The authors analyze the number of likes a tweet obtains within its first hour, across tweets posted at different times of day. They report a significant (time of day × virtue vs vice) interaction, "$z=12.17$, $p<.001$" (p.479). Probing the interaction with the (linear) Johnson-Neyman procedure, they conclude that before 12.01 PM, virtue tweets get more likes than vice tweets do, and that starting at 2:26 PM, the opposite is true.[9] We obtained data posted by the authors (https://osf.io/hya8z) and successfully reproduced the key results.[10]

[9] The authors do not indicate how they handled time zones, we use the posted data as is.

[10] We do obtain slightly different estimates for reasons we did not determine. It is possible they are explained by different defaults in STATA (used by the original authors), vs R. For example, the authors report Z=12.17 for the interaction, and we obtain Z=12.525. The differences are inconsequential however, for example the

What's interesting for us, given our focus on the probing of interactions, is that when relying on the linear model, the published result hinges entirely on the hour at which we define the start of the day. If we do as the authors did, and define the start of the day at 6AM, we reproduce their findings. See left panel in Figure 3. However, if we define the start of the day at midnight, which is how many people define the start of the day, a completely different pattern (that lacks an interaction) is obtained; see middle panel in Figure 3. And if we define the day as starting at 4PM, which is desirable for statistical reasons we explain later, yet another completely different pattern arise (one that also lacks an interaction). See right panel in Figure 3.

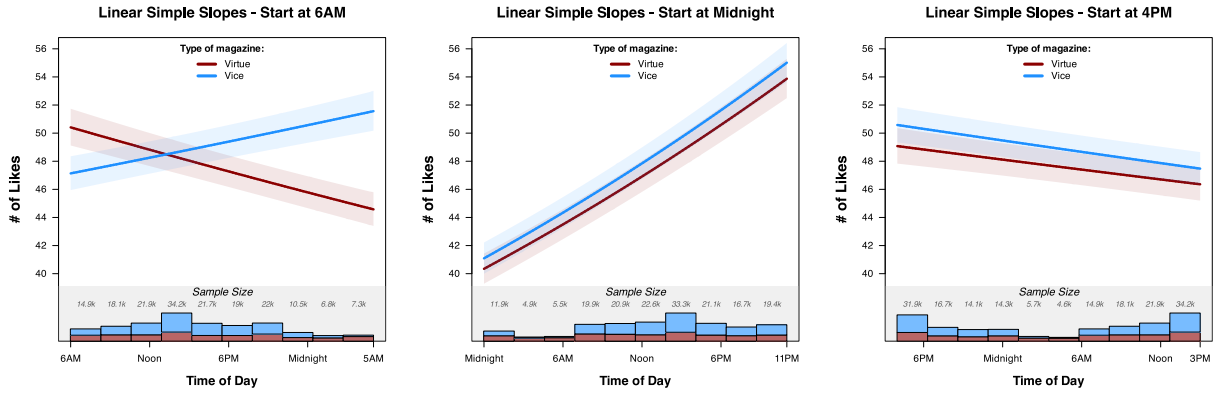

Figure 3. *Linear Probing Gives Inconsistent Answers for Different Definitions of Start of Day*
Notes. Reanalysis of Study 1A by in Zor, Kim, and Monga (2022) on number of *likes* (N=176,390) received by tweets posted at different times of day. Left panel reproduces the published results, middle and right panels depict how the probed interaction differs when using different start-day definitions with the same linear model.
Code to reproduce figure: https://researchbox.org/2859/7 (enter code QAMDS)

You can start getting an intuition for why this happens, and for why *neither* model in Figure 3 is a sensible depiction of reality, by noticing an incoherent pattern implied by modelling time of day linearly. In the left panel, for example, focus on when a day ends and a new day begins, going from the right-end to the left-end of the graph. The predicted number of likes drops

authors find that before 12.01 PM liking was lower for vice magazines, and with our results it is 11.59AM, just two minutes apart.

discontinuously from the end of the figure (which corresponds to 5:59 AM) to the left-end of the figure (which corresponds to 6:00 AM). The drop for that single minute is (necessarily) of the same magnitude as the change for the remaining 1439 minutes. For example, for Virtue tweets, at 5:59AM the minimum is obtained, at about 44 likes per tweet; a minute later, at 6:00 AM the maximum is obtained, at about 50 likes per tweet.[11]

When we analyze the data with GAM, see Figure 4, the results are not altered by whether the day is defined to begin at midnight, 6AM, or 4PM. For example, in all panels we see that virtue and vice tweets follow similar patterns through the day, and that the vice peak and trough are more pronounced (at 10PM and 4AM respectively). More generally, GAM estimates a daily pattern that is completely different from the pattern obtained with the linear model; it is cyclical rather than linear (and of course, a linear model cannot estimate a cyclical relationship).

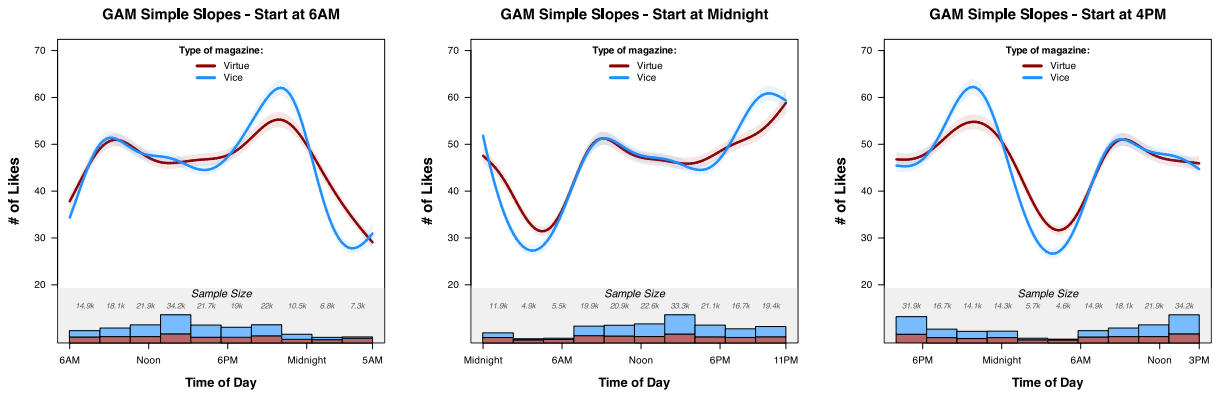

Figure 4. *GAM Probing is Robust to Different Definitions of Start of Day.*
Notes. Reanalysis of Study 1A by in Zor, Kim, and Monga (2022) on number of *likes* (N=176,390) received by tweets posted at different times of day. In contrast to the linear model (see Figure 3) GAM probing leads to similar results no matter when the day is defined to start. All results are inconsistent with those in the published paper.
Code to reproduce figure: https://researchbox.org/2859/7 (enter code QAMDS)

[11] The original authors worried about linearity and present in the supplement a quadratic regression. However, this it is still insufficiently flexible. In a nutshell, they allow the main effects but not the interaction to be non-linear.

We now return to why we believe that defining the day as starting at 4PM is reasonable. There is very little data at the beginning and end of a natural day (midnight). Fewer people liking tweets at 3AM than at 3PM. See histograms in the bottom of Figure 4. Defining the day as starting at midnight or 6AM leaves the hours with less data at the ends of the range of values, making it challenging to estimate functional form during those hours (e.g., GAM cannot use 7AM data to smooth 5AM data when they are on different ends of the graph). When defining the start of the day at 4PM, GAM has the most data at the ends of the range of values, increasing its ability to perform well at the tails, and it has more data surrounding the periods with few observations (midnight), increasing its ability to perform well at those times. Thus, we believe the estimates of the model that defines the day as 4PM-3:59PM to be most trustworthy, and indeed in that model the discontinuity from end to start of day is the smallest and the overall pattern the smoothest. We did not tell GAM that 3:59PM comes right before 4:00PM, but GAM effectively figured it out by flexibly estimating functional form.

Example 3: Significant Estimate of the Wrong Sign

Woolley et al. (2023) propose that consumers prefer high-tech products made by larger companies, and low-tech products made by smaller companies. Study 1 in the paper is observational, Studies 2-6 involve experiments. We focus on Study 1, which is the only one where interactions are probed. Here the authors use a company's Net Promoter Score (NPS) as a proxy for perceived quality of its products (it ranges from -100 to +100). They obtain data for 480 companies in the Fortune 500 list. The authors predict "an interaction between company size and industry type (low-tech vs. high-tech), such that a larger [company] size would negatively predict NPS for low-tech industries but would positively predict NPS for high-tech industries" (p.430).

The authors measured company size by averaging measures of number employees and revenues per company, and how tech intensive a company was based on an MTurk survey where participant evaluated the company on a 7-point scale (1-low tech, to 7-high tech). Finally, the NPS was obtained from "Customer Guru" (p. 433).

The authors estimate a regression predicting Net Promoter Score with company size, tech intensity of the company, and the interactions. That interaction proved significant, $\hat{b}$ =1.90, t(476)=3.67, p<.001. The authors probed it with the (linear) Johnson-Neyman procedure, finding that company size was negatively associated with Net Promoter Score when the tech index was below 1.94, and positively when above 3.57. We obtained the data posted by the authors (https://osf.io/hya8z/) and successfully reproduced these results (see left panel in Figure 5).

What's interesting for us, given our focus on the probing of interactions, is that the reversal for low-tech firms, that estimated negative coefficient for company size among low-tech firms, appears to be spurious. In the GAM model, the association is actually *positive* also for low-tech firms. Specifically, Panel B in Figure 5 shows the GAM Johnson-Neyman curve, which in this case is U-Shaped. It shows that the association between company size and NPS is positive for both high- and low-tech firms. Moreover, the *biggest* positive estimate is for the low-tech firms (around +27 points) rather than high-tech firms (around +18.0 points).

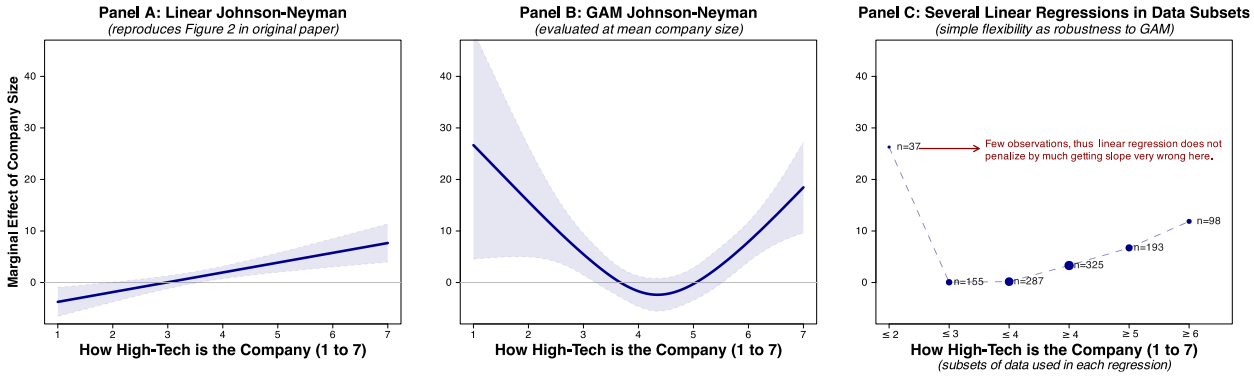

Figure 5. *Linear Model Estimates Spurious Reversal*
Notes. Reanalysis of Study 1 by Woolley et al. (2023) on how company-size (N=480 companies) predicts Net Promoter Score (NPS) for companies rated low- vs high-tech (by sample of MTurk respondents). The patterns that among low-tech companies bigger companies get lower NPS is only obtained when forcing linearity in the model. Both GAM and a flexible regression show, instead, a U-Shaped relationship.
Code for figure: https://researchbox.org/2859/4 (enter code QAMDS)

Note that this result contradicts the abstract which states that "For low-tech products . . . quality evaluations and choice [moves] in favor of smaller companies." (p.425). Note also, however, that there is experimental evidence of that prediction in other studies in the paper.

We checked the robustness of these conclusions by running a series of linear regressions predicting NPS only with company size, on subsets of data with different tech index values (see Panel C in Figure 5). For example, the right-most dot in that panel reports the estimated coefficient for company size predicting NPS among companies rated 6 or higher in tech. As we move left, the coefficients are based on subsets of the data with lower tech ratings. This simpler (but less powerful and less scalable) approach to estimating functional form, qualitatively reproduces GAM's U-shape.[12]

The confidence band among low tech firms is much wider with GAM vs. with linear probing (comparing the left side of panels A and B). The linear model confidently but erroneously

[12] We should note that in the data there is an outlier 14.75 SDs above the mean (Walmart), which leads the linear model to have a false-positive rate above and beyond the issues we discuss here.

estimates the negative effect for low tech firms relying on the linearity assumption. The GAM model appropriately places great uncertainty on an estimate that is based on a small number of observations (e.g., just n=37 with tech index below 2). For example, the standard error for the point estimate for a company with a tech index equal to 2 is about five times larger with GAM than with the linear model (5.64 vs 1.00 respectively). The GAM results are providing better knowledge and better meta-knowledge.

Example 4: A Spurious Sign Reversal

Barnes and Shavitt (2023) propose that 'interdependent' people prefer products that are frequently loved rather than frequently bought, while 'independent' people are not impacted by whether a product is frequently loved or bought; for them it "makes little difference" (their abstract).

In their Study 5, participants were presented with an image of a set of headphones which they were told 81% of people had either purchased or loved (there was also a control condition which is not relevant for our purposes). Participants then indicated their interest in the headphones through a few dependent variables (e.g., rating how appealing the headphones were, and what percentage of the retail price they would pay for them) that were aggregated to form a single index. Participants also completed a multi-item scale to measure whether they perceive themselves to be interdependent (e.g., "I will stay in a group if they need me. . .") vs. independent (e.g., "I enjoy being unique. . ."). The authors report a significant interaction, such that the frequently bought vs frequently loved manipulation was moderated by the degree of interdependence of participants ($t(186) = 3.00$, $p = .003$). We obtained the data posted by the authors (https://researchbox.org/108) and successfully reproduced the key results.

What's interesting for us, given our focus on the probing of interactions, is that while in a manner that is consistent with the abstract, the linear interaction reported by the authors implies that interdependent participants actively prefer frequently loved products, in a manner that contradicts the abstract it also implies that independent participants show the opposite pattern.

Concretely, as shown in the left panel of Figure 6, the *linear* Johnson-Neyman curve implies that for moderator values below -1.31 there is a statistically significant reversal of the effect (the authors predict no effect among them). However, with the GAM, see right panel of Figure 6, the effect for low values of the moderator is estimated as much smaller and far from significant (with the moderator at -1.5, the estimated effect is 0.16 points, $p = .630$, in contrast to .65 and p=.029 with the linear model).

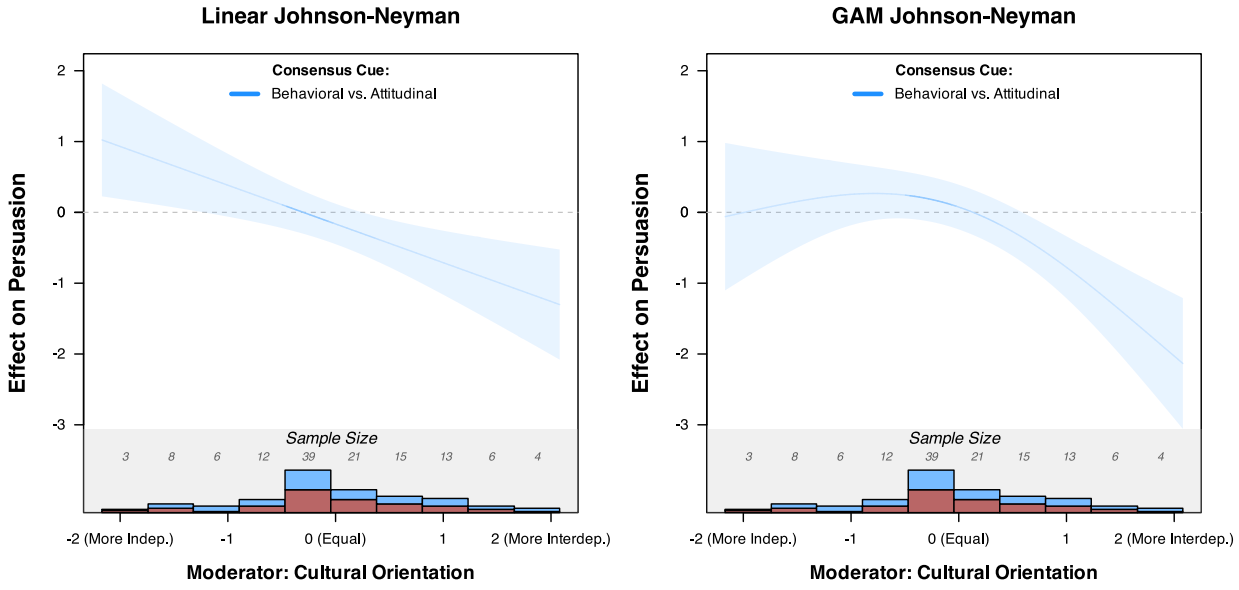

Figure 6. *Linear Model Incorrectly Reverses Interaction Effect, While GAM Accurately Recognizes the Effect Merely Attenuates.*

Notes. Reanalysis of Study 5 by Barnes and Shavitt (2023) on how on how behavioral vs. attitudinal cues (being told that 81% of people bought vs. loved a set of headphones) affect product interest depending on participants' (N=192) cultural orientation. The significant reversal among more independent individuals is only obtained when forcing linearity on the data by estimating a linear model.
Code for figure: https://researchbox.org/2859/5 (enter code QAMDS)

As we did in the previous examples, we checked the robustness of these conclusions by running a *linear* regression that allowed for a different interaction effect for moderator values below vs above 0, the point suggested by the GAM model in Figure 6. Consistent with the GAM results, the effect of the manipulation was small and not-significant for moderator values below 0 ($\hat{b}_{\text{below-zero}} = .01$, $p = .979$), while for moderator values above zero it was negative and significant ($\hat{b}_{\text{above-zero}} = -.997$, $p = .029$).

Interestingly, at the lowest value of cultural orientation, the confidence interval of the GAM rejects the point estimate of the linear model. The estimate, however, is quite noisy, in part because -as depicted in the figure- there are few participants with high enough independence in the region where the linear model predicts the reversal. We think of that as a feature rather than a bug. When there is little data around a moderator value, the GAM correctly conveys the uncertainty for the effect around such value, instead of (over)confidently misestimating it, by relying on the arbitrary and often false assumption that everything is linear.

GAM Non-limitations

We consider two non-limitations, that is, two arguments that are sometimes made against relying on GAMs which we do not believe are actually limitations: (i) interpretability of results, and (ii) sample size requirements.

In terms of interpretability. It is often argued that GAM results are less interpretable than regression results. We do not think this is an actual limitation, at least when it comes to probing interactions, because, as we have shown, computing Simple Slopes and Johnson-Neyman curves is as straightforward and interpretable for GAM models as for linear models.

In terms of sample size. Researchers often intuit that GAM models require much larger sample sizes than linear models to be informative. This, however, is not generally the case. To illustrate, we present results from a simulation comparing the precision of linear regression and GAM, focusing on the estimated change in the effect of the focal predictor when the moderator is low versus high.

We simulate a true model that is linear, giving the linear regression the advantage. Specifically, we generated a true model: $y = x + z + x \cdot z + e$, where x is randomly assigned (1 vs. 0), z is a standard normal moderator, and *e* is random noise (we calibrate the SD of e~N(0,SD) to give the linear model R^2=33%). We consider a researcher interested in estimating a model with the x·z interaction, and probing the effect of x when z=1 vs. z=-1. Because these are simulated data, we know the true effect of x; it is $\frac{\Delta y}{\Delta x} = 2$ when z=1 and $\frac{\Delta y}{\Delta x} = 0$ when z=-1. We thus know that the interaction, the change in $\frac{\Delta y}{\Delta x}$ when the moderator z goes from 1 to -1 is equal to 2-0=2. We run 5,000 simulations with n=20 per cell and 5,000 simulations with n=100 per cell, tracking the estimated interaction effect in each simulation, and then compared it to the true value of 2; see Figure 7.

When n=20 (left panel), unsurprisingly the estimates are quite noisy. While they are centered around the true value of 2, they fluctuated from -1 to 6 across simulations. Importantly, despite the small sample sizes, the estimate is about as precise with GAM as with linear regression. The standard deviation across simulations (the standard error of the estimator) is 0.96 with regression and 1.00 with GAM, only about 5% higher. The right panel shows that with n=100 per cell, the estimates are more stable from sample to sample, and the precision difference between GAM and the linear model becomes even smaller.

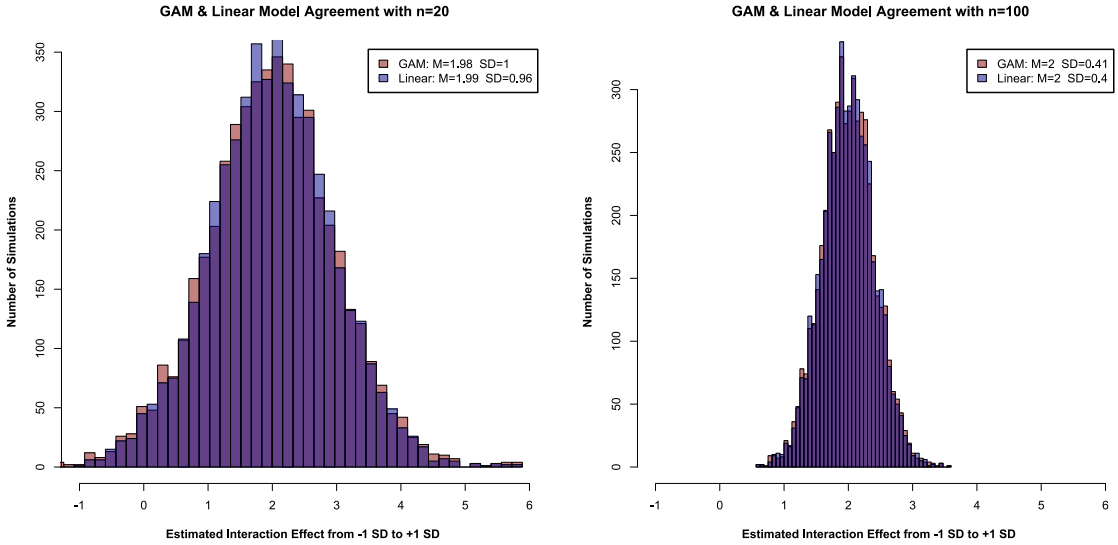

Figure 7. *When True Model is Linear, Linear Regression and GAM are Similarly Precise*
Note: This figure presents results from 10,000 simulations. The true model is $y = x + z + x \cdot z + e$, where x is randomly assigned (1 vs. 0), z is a standard normal moderator, and e represents random noise. The interaction effect was estimated for z=-1 vs. z=1 so the true change in effect is +2.
Code to reproduce figure: https://researchbox.org/2859/6 (enter code QAMDS)

GAM Limitations

While we have argued that neither interpretability nor sample size requirements are real GAM limitations, we do think GAM has limitations. In our view, the most important limitation is the inherent greater flexibility of a richer statistical approach like GAM, over a more primitive approach like linear regression. As is the case with mixed-models and meta-analysis, there are several options one can set when estimating a GAM (e.g., optimization routine and how many functions to rely on to fit individual predictors). Fortunately, the default options that are incorporated into R's most popular implementation of GAM, the 'mgcv' package, are sensible. We relied on those defaults for all examples and simulations in this paper; with those defaults we obtained results that seemed sensible to us and, more persuasively, were validated by estimating linear models that allowed for slope changes at the relevant points suggested by the GAM results. In our Example 1, for instance, we verified that the interaction effect of "*le"* vs "*la"* COVID was absent within the range of moderator values GAM identified as lacking an interaction, by estimating linear regression models that allowed the slopes in that range to be different. If the pattern identified by GAM were incorrect, then the linear model would not corroborate it (note that without GAM, we would not have known where to allow the slopes to change).

Nevertheless, sometimes the best fitting line obtained with GAM may overfit. It may for instance seem excessively wiggly for what one may reasonably expect to observe in social science data. For example, imagine an effect that is positive when a moderator is between 3 and 3.5, negative between 3.5 and 3.8 and then positive again between 3.8 and 4.3.

When worried that GAM may be overfitting data, there is an easy parameter to adjust: the number of different functions combined to fit each predictor's relationship with the dependent variable. That parameter is called 'k' in the mgcv package in R. For example, setting k=3, restricts

GAM to consider up to three functions to fit any one predictor. A practical solution when encountering apparent excessive wiggliness is to estimate a GAM with a low value of k, say k=3, and then increasing k up to the highest value producing a sensible (not excessively wiggly shape). When this proves necessary, we recommend authors include in the paper results for the alternative k values for transparency. We should emphasize we did not need to do this for any of the examples in this paper. The defaults worked well, when predicting net-promote-scores in fortune 500 companies, number of liked tweets posted by magazines through the day, danger perception of feminine vs masculine COVID, and interest in headphones. That's a rather broad range of variables, which also varied in the data structures that contained them.

Conclusions

Whenever we probe interactions with linear models, we rely on tools that assume, rather than estimate, the functional form of the effects of interest. In this paper, we have shown that this assumption can lead to highly misleading results, supporting erroneous conclusions from data. We offer an easy-to-use GAM-based alternative that relaxes the linearity assumption, imposes minimal cost on researchers, and allows us to better understand the phenomena we study.

References

Aiken, Leona S and Stephen G West (1991), Multiple regression: Testing and interpreting interactions: Sage.

Barnes, Aaron J and Sharon Shavitt (2023), "Top Rated or Best Seller? Cultural Differences in Responses to Attitudinal versus Behavioral Consensus Cues," *Journal of Consumer Research*.

Beck, Nathaniel and Simon Jackman (1998), "Beyond linearity by default: Generalized additive models," *American Journal of Political Science*, 596-627.

Berger, Jonah and Katherine L Milkman (2012), "What makes online content viral?," *Journal of marketing research*, 49 (2), 192-205.

Cortina, Jose M (1993), "Interaction, nonlinearity, and multicollinearity: Implications for multiple regression," *Journal of management*, 19 (4), 915-22.

Ganzach, Yoav (1997), "Misleading interaction and curvilinear terms," *Psychological methods*, 2 (3), 235.

---- (1998), "Nonlinearity, multicollinearity and the probability of type II error in detecting interaction," *Journal of Management*, 24 (5), 615-22.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu (2019), "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice," *Political Analysis*, 27 (2), 163-92.

Hastie, Trevor and Robert Tibshirani (1987), "Generalized additive models: some applications," *Journal of the American Statistical Association*, 82 (398), 371-86.

Johnson, P. O. and J. Neyman (1936), "Tests of certain linear hypotheses and their application to some educational problems," *Statistical Research Memoirs*, 1, 57-93.

Lubinski, David and Lloyd G Humphreys (1990), "Assessing spurious" moderator effects": Illustrated substantively with the hypothesized (" synergistic") relation between spatial and mathematical ability," *Psychological bulletin*, 107 (3), 385.

Mecit, Alican, LJ Shrum, and Tina M Lowrey (2022), "COVID-19 is feminine: Grammatical gender influences danger perceptions and precautionary behavioral intentions by activating gender stereotypes," *Journal of Consumer Psychology*, 32 (2), 316-25.

Simonsohn, Uri (2024), "Interacting with curves: How to validly test and probe interactions in the real (nonlinear) world," *Advances in Methods and Practices in Psychological Science*, 7 (1), 25152459231207787.

---- (2018), "Two Lines: A valid alterantive to the invalid testing of u-shaped relationships with quadratic regressions," *Advances in Methods and Practices in Psychological Science*.

Spiller, Stephen A, Gavan J Fitzsimons, John G Lynch, and Gary H McClelland (2013), "Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression," *Journal of Marketing Research*, 50 (2), 277-88.

Wood, Simon N (2017), Generalized Additive Models: An Introduction with R (Second ed.).

Woolley, Kaitlin, Daniella Kupor, and Peggy J Liu (2023), "Does company size shape product quality inferences? Larger companies make better high-tech products, but smaller companies make better low-tech products," *Journal of Marketing Research*, 60 (3), 425-48.