First draft: 2024-06-17
This draft: 2026-07-27

Johnson-Neyman 2.0: Probing Interactions Without Assuming Linearity, Four Demonstrations and One Tutorial

Andres Montealegre
Yale School of Management
andres.montealegremoreno@yale.edu

Uri Simonsohn
Esade Business School
urisohn@gmail.com

Abstract
Interactions, where one variable moderates the association between other variables, are central to marketing research. They are typically probed using linear regression followed by Simple Slopes ("spotlight") or the Johnson-Neyman procedure ("floodlight"). We reanalyze four recently published marketing papers to illustrate how these procedures can lead to qualitatively incorrect conclusions when true relationships are nonlinear. We propose relying on GAMs (Generalized Additive Models) instead of linear regressions to probe interactions. We provide a tutorial for conducting and reporting GAM Simple Slopes and GAM Johnson-Neyman procedures. They are just as easy to run and interpret, much more robust, and involve minor power losses.

Keywords: interactions, probing, Johnson-Neyman, spotlight, floodlight, GAM

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

INTRODUCTION

Interactions, where a variable moderates the association between other variables, are commonly examined in marketing research. Researchers ask questions like: Does a consumer trait or a product characteristic increase price sensitivity? After documenting a statistically significant interaction, it is common to 'probe' it, to assess the effect size of the variable of interest for different moderator values through two common techniques: (1) estimating the effect of the focal predictor x (e.g., price) on the dependent variable y (e.g., quantity purchased) at different values of the moderator z (e.g., for participants at +1 SD on a "spendthrift-tightwad" scale) and (2) assessing for which moderator values the effect of the focal variable is positive vs. negative, or statistically significant vs. not. These approaches are respectively known as Simple Slopes (Aiken and West 1991) or "spotlight" analysis, and the Johnson-Neyman procedure or "floodlight" analysis (Spiller et al. 2013).

The technology currently used for studying interactions involves estimating a linear regression that includes the product x·z as a predictor (e.g., y = a + bx + cz + dx·z), and then probing the interaction based on the regression coefficients associated with x ($\hat{b}$ and $\hat{d}$). This technology is about 90 years old, dating back at least to Johnson and Neyman (1936).

A significant limitation of this approach to probing interactions is that the results are too sensitive to the arbitrary and often false assumption that all effects in the model are linear. This concern is not new. Even in their original paper, Johnson and Neyman (1936) wrote "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 their future.

Returning to our opening example, a linear regression assumes that the relationship between price and quantity 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 is unlikely to be linear. For instance, if taco prices drop from $9 to $7, Alex may buy 5 instead of 4 tacos. Yet, if the price drops by additional $2s, to $5 and then $3, we do not expect Alex to continue buying one more taco for every $2 decrease. At some point Alex is ready for dessert regardless of how cheap the next taco is. We should be skeptical of a perfect straight line connecting price and quantity of tacos purchased. We should be skeptical of perfect straight lines connecting any two variables.

A natural way to avoid the problems that arise from assuming linearity is, quite simply, not to assume it, relying on more flexible models like Generalized Additive Models (GAMs), which have existed for about four decades (Hastie and Tibshirani 1987). GAMs are so broadly accepted in statistics today that `mgcv`, the package for estimating them in R, is one of only 15 packages that come recommended with R (there are nearly 25,000 other packages that did not make the cut). Wood's (2017) textbook on GAMs has over 22,000 Google Scholar citations. So while GAMs may be new to many of us in the social sciences, they are not new per se.

A recent methodological paper by Simonsohn (2024) is most relevant to this article. It uses simulation studies to compare the performance of GAM-based versus linear-based probing of interactions, finding that GAMs vastly outperform linear regressions for this purpose. For example, in realistic scenarios, linear regressions can hit a 100% false-positive rate when probing interactions (e.g., concluding with certainty that an effect reverses sign when it does not). In

contrast, the GAM false-positive rate stays at or below the nominal 5% (see his Figure 6). Conversely, when the underlying relationships are truly linear, GAM and linear regression have very similar levels of power; see his Figure 8. (We discuss these simulation results in more detail later.) GAMs, therefore, buy us a large reduction in Type 1 errors, while costing us very little in Type 2 errors.

GAMs outperform linear regressions because they estimate, rather than assume, the functional form of the relationship between the variables in the model, allowing them to capture many kinds of nonlinear effects while incorporating a penalty for overfitting. Another reason GAMs outperform is that they barely extrapolate from one region of the data to another. When summarizing the effect of dropping the price of tacos from $5 to $3, the linear regression gives an answer that averages the response across all prices, both much higher prices, say $10, and much lower, say $1. GAMs, in contrast, make local summaries based on local data. The summary of what happens between $5 and $3 is based primarily on data between around $5 and around $3, almost entirely independently of what happens when the price is $10 or $1.

Estimating the functional form locally has the further benefit of producing more calibrated and thus valid confidence bands. Because the linear model assumes a single line fits everywhere, it does not take into account how much data is in a region to determine the width of the confidence band for its predictions in that range of data; a linear regression could report narrow bands in regions with little or even no data. GAMs' confidence bands, in contrast, appropriately widen in regions with less data.

These two shortcomings of regression: (1) assuming rather than estimating functional form, and (2) using data from one region of possible values to inform estimates in distant other regions,

can, as shown by Simonsohn (2024) via simulation studies, lead researchers to qualitatively incorrect conclusions.

In this paper we make two main contributions. First, we provide an accessible introduction to GAMs for non-methodologically oriented researchers. We provide an intuition for how GAMs work, for why they outperform linear regressions, and give concrete guidance for readers to start using them in their own research. Second, we show that the scenarios simulated by Simonsohn (2024), where GAMs provide qualitatively different answers, are not just a theoretical possibility but something that happens in practice. We show this by reanalyzing the probing of interactions in four recent marketing articles. In each case, the qualitative conclusions from linear probing are partially or entirely reversed when one no longer arbitrarily forces linearity on the data.

Example 1 shows that the linear analysis in the published paper misplaced the locus of the interaction; it occurs only among extreme observations, but the linear model reports incorrectly that it occurs for most observations. Example 2 shows that forcing linearity on cyclical data produces estimates that are simply implausible. Example 3 shows that forcing linearity on the data produces a spurious reversal, the interaction merely attenuates the focal effect, but the linear model incorrectly indicates that it reverses. Example 4 shows that forcing linearity on the data produces a statistically significant effect of the wrong sign.

We want to make clear that while probing of interactions is central to our paper, it was not necessarily central to some of the papers we revisit. Indeed, in three of these four papers, the paper's main conclusions do not hinge on the specific results we reanalyze, either because the papers include other studies that do not exhibit the issue we flag, or because the issue we flag impacts a secondary result in the paper. We should also stress that we did not select these papers because they are substandard in quality; in fact, we chose them for the opposite reason. These papers are

useful to us precisely because the authors followed what is standard in the field: relying on linear models to probe interactions. Moreover, all four papers are *above* standard on open science: the authors posted their data, allowing us to reanalyze them with a different strategy.

Part of the appeal of the current approach, the linear probing of interactions, is its simplicity. Most of us are familiar with Simple Slopes (spotlight) and Johnson-Neyman (floodlight) plots. Fortunately, switching to GAM will not make probing interactions any harder for researchers. The R package `statuser`, available on CRAN, includes the function `interprobe()`, which produces ready-to-publish GAM Simple Slopes and GAM Johnson-Neyman plots. Both can be produced with a single line of code: `interprobe(x,z,y)`. We also made an online app available (webstimate.org/interprobe) so that readers who do not use R can run the same analyses by pointing and clicking on a website.

We next provide an intuitive introduction to GAMs, focusing on stylized scenarios that showcase how and why GAM outperforms linear models. We then present the four demonstrations reanalyzing recently published marketing papers. We end with a tutorial on how to conduct, interpret, and report interactions relying on GAM.

AN INTUITIVE INTRODUCTION TO GAMs

The best way to introduce GAMs is to contrast them with linear regression. In both cases we wish to characterize the association between a dependent variable, y, and one or more predictors. In the simplest case, with just one predictor, let's say the true relationship between x and y is captured by the function f(), so y = f(x). When we run a regression we assume we *know* the functional form of f(), we *know* it's linear, and all that's left to figure out are the intercept and slope. So, we assume we know y=f(x)=a+bx, and all that's left is figuring out *a* and *b.* With GAM,

instead, we don't assume that. We instead *estimate* f(). GAMs carry out the estimation by (1) combining a set of functions (linear, log, quadratic, etc.) so that they fit the data well, while (2) allowing the function combination to be different for different portions of the x values (e.g., f(x) is log(x) for low values, but flat for high values of x), and finally (3) avoiding overfitting by preferring results that produce less wiggly predictions.

In practice, when the true function f(x) is linear, both the regression and the GAM will produce a linear estimate, but when f(x) is not linear, the regression will continue producing a linear estimate, and the GAM will approximate the true functional form with a combination of nonlinear functions. This of course all happens in the back end, in the front end the experience for the researcher running the analysis is remarkably similar. For example, to estimate a regression in R you run lm(y~x), to estimate a GAM you run gam(y~s(x)).

As an introductory illustration of the capability of GAM to recover functional form, in Figure 1 we report results from a simulation with 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 model 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.

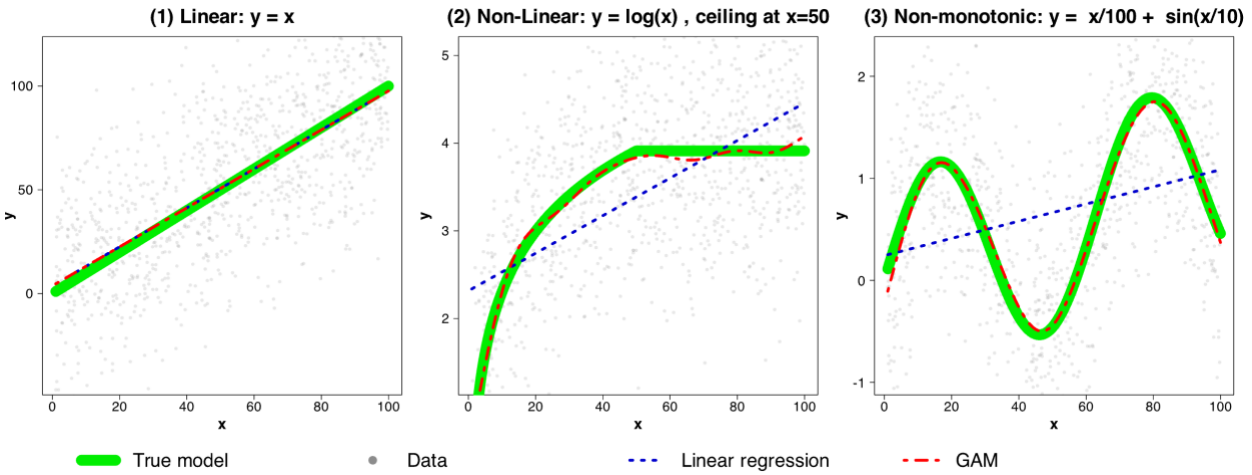

Figure 1. *Comparing Model Adequacy: GAM vs. Linear Regression.*
Notes. The figure depicts a single simulation of 1000 observations where x~U(0,100). The figure shows that when the true model is linear GAM recovers a linear model, but when it is not, it accurately captures the alternative functional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

Turning now to the specific context of interest, probing interactions, we illustrate in Figure 2 concrete and likely scenarios where a linear model is expected to arrive at qualitatively incorrect conclusions when probing an interaction, while GAM is expected to arrive at qualitatively correct conclusions. The figure has four scenarios. We begin with Scenario 1. Here there truly is an interaction, and everything is linear. Both a linear regression and a GAM do a good job of recovering that linear association. This is a scenario textbooks and methodological papers consider when advocating for linear probing. It is, actually, the *only* scenario they consider. But it is useful to also consider the rather likely alternative that not everything is linear.

In Scenario 2 we still assume an underlying linear association, but we consider the impact of a ceiling on the response variable, a common occurrence in research that relies on scales. Here both groups have the same slope below the ceiling, but one group hits the ceiling earlier. Because linear regression cannot accommodate this nonlinear pattern, it incorrectly projects a *reversal* at high moderator values. In contrast, GAM correctly recognizes that the interaction merely attenuates, and thus provides an adequate qualitative summary of the data.

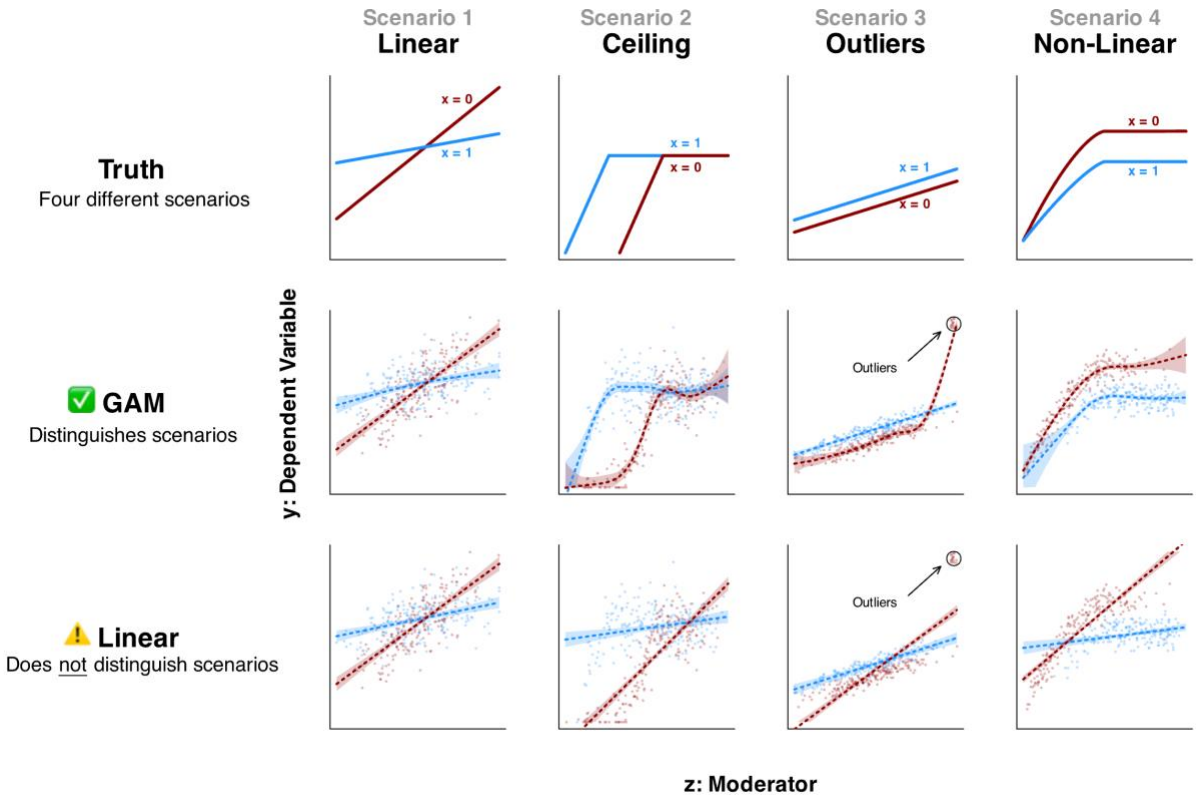

Figure 2. *GAM Can Handle the Nonlinear Truth, Regression Cannot.*
Notes. Comparison of linear regression and GAM Simple Slopes across four simulated scenarios (N=400 per scenario). Each column shows one scenario with three rows: true functional form (top), GAM estimates (middle), and linear estimates (bottom). Blue and red lines represent the two experimental groups (x=1 and x=0), and shaded regions represent the 95% CI. In Scenario 1, where the true relationship is linear, both methods perform similarly. In Scenarios 2-4, linear regression produces misleading results: projecting an erroneous reversal (ceiling effect), creating a spurious interaction (outliers), and misplacing the locus of interaction (nonlinear).

In Scenario 3, instead of a ceiling effect, we consider outliers. While there is no true interaction at all, the outliers cause the linear regression to generate a spurious interaction purportedly for all moderator values. The GAM, in contrast, correctly recognizes these as isolated extreme values and detects no interaction across most moderator values.

Scenario 4 shows a nonlinear interaction such that the difference across conditions is near zero for low moderator values, grows as the moderator increases, and then levels off. The linear regression correctly detects an interaction but mischaracterizes its shape, and in doing so also

projects an erroneous reversal of the effect at low moderator values. GAM's estimates, in contrast, qualitatively mirror the true association.

Look back at the bottom row of Figure 2, and note how the linear Simple Slopes look similar across all four scenarios, even though the underlying realities are radically different. In contrast, GAM's estimates accurately capture four different underlying realities.

GAM also does a better job of conveying the uncertainty about functional form in different regions of the data. To illustrate, we created a scenario where the true relationship is an inverted-U, but for mid values the data is sparse. The left panel of Figure 3 shows the true effect. The linear model (middle panel) not only gets the shape wrong, estimating a flat line instead of an inverted-U, but its confidence band is narrowest near the center, precisely where the data is sparsest (thus, the linear model exhibits the Dunning-Kruger effect here: it's most confident where it knows the least). The GAM (right panel), in contrast, captures the right shape and communicates through a wider confidence band that there is more uncertainty where there is less data. GAM not only gives better knowledge, it also gives better meta-knowledge.

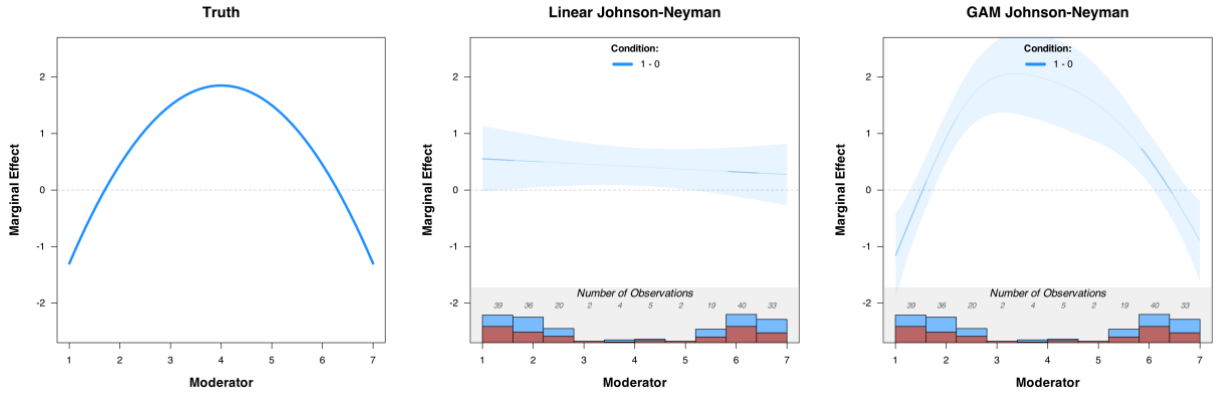

Figure 3. *Linear vs. GAM Johnson-Neyman with Sparse Data.*
Notes. Comparison of Linear (with robust SEs) and GAM across a simulation with inverted-U interaction (N=200) where observations are concentrated at the extremes of the moderator, with very few in the middle. The left panel shows the true marginal effect. The linear model (middle panel) estimates a near-flat effect and shows narrow confidence bands in the data-sparse middle region, while GAM (right panel) recovers the inverted-U shape with confidence bands that widen where data is scarce.

Simonsohn (2024) relied on simulation studies to systematically evaluate the performance of GAMs vs. linear regression for probing interactions. Most relevant here are his simulations of experiments, where the focal predictor is randomly assigned. In scenarios where the underlying associations are nonlinear, linear probing produces false-positive rates—statistically significant effects of the wrong sign—that reach 50% and even 100%, while GAM probing stays at or near the nominal 5% (see his Figure 6). Moreover, when the true relationships are not linear, GAM has greater statistical power and is more informative about the nature of the relationship of interest (a clear example is presented in our Figure 2). Conversely, when the true relationships are linear, GAM has lower power than the linear model, but the drop is generally inconsequential, between roughly 0 and a few percentage points (see his Figure 8). In addition, when the true model is linear, model fit (mean squared error) is essentially identical in the linear and GAM models (see his footnote 15). Given that GAM protects us when linear regression fails, while performing nearly identically when linear regression succeeds, GAM is the safer default for probing interactions.

The scenarios we depict in Figures 2 and 3, and the simulations reported by Simonsohn (2024), establish that it is *possible* for the linear model to arrive at qualitatively incorrect conclusions. We next demonstrate that this is not merely a hypothetical possibility, presenting examples from recently published marketing papers that exhibit precisely these kinds of shortcomings. We also use our first example to walk through the mechanics of running GAM probing.

Before we proceed, we want to be clear that our goal is not to critique the four individual papers or their authors; they all defensibly followed standard procedure. Our goal is to critique the standard procedure, so that everyone stops following it.

EXAMPLES

Example 1: Misplaced Moderation

Mecit, Shrum, and Lowrey (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. For instance, in French one may say "la" COVID (feminine) or "le" COVID (masculine).[1]

After documenting main effects in Studies 1 and 2, in Study 3 they examine moderation by participants' chronic gender stereotyping. In that study, N=305 French speakers were randomly assigned to have the disease described with the feminine vs. masculine terms and indicated how dangerous COVID-19 seemed to them. They also completed a 24-item gender stereotype questionnaire to measure "chronic gender stereotyping" (p.321); the moderator.

The authors find a significant interaction between the manipulation and chronic gender stereotyping (p = .0015), which they then probed with linear Simple Slopes and the Johnson-Neyman procedure. The latter indicated that the effect of "Le" vs. "La" was significant for participants with a chronic gender stereotype average above 0.53.[2] In Figure 4, we reproduce their linear Simple Slopes, followed by our GAM based probing.

[1] The Académie française (2020) recommended that the proper grammatical term is "La COVID" rather than "Le COVID". After this, formal sources tended to use "La COVID" while "Le COVID" remained in use in informal settings. Thus, Le COVID is a masculine *and informal* term, and La COVID is a feminine *and formal* term.

[2] 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, but the effect is also significantly negative for moderator values below -1.65.

Before interpreting these results, we use Figure 4 to remind readers of what Simple Slopes and Johnson-Neyman involve for linear procedures, and explain how they extend to GAM procedures. Linear Simple Slopes, like those in Figure 4A, are computed by estimating a linear regression like y=**a**+**b**x+**c**z+**d**x·z, and then predicting y (in this example, *Danger Perception*), for a given x (in this example, "*Le*" vs "*La*"), for different z values (in this example, *Chronic Gender Stereotypes*). The way GAM Simple Slopes are computed is analogous. One estimates a GAM, and uses it to predict y, for a given x, for different z values. The key difference is that GAM does not force the lines to be straight (Figure 4B).

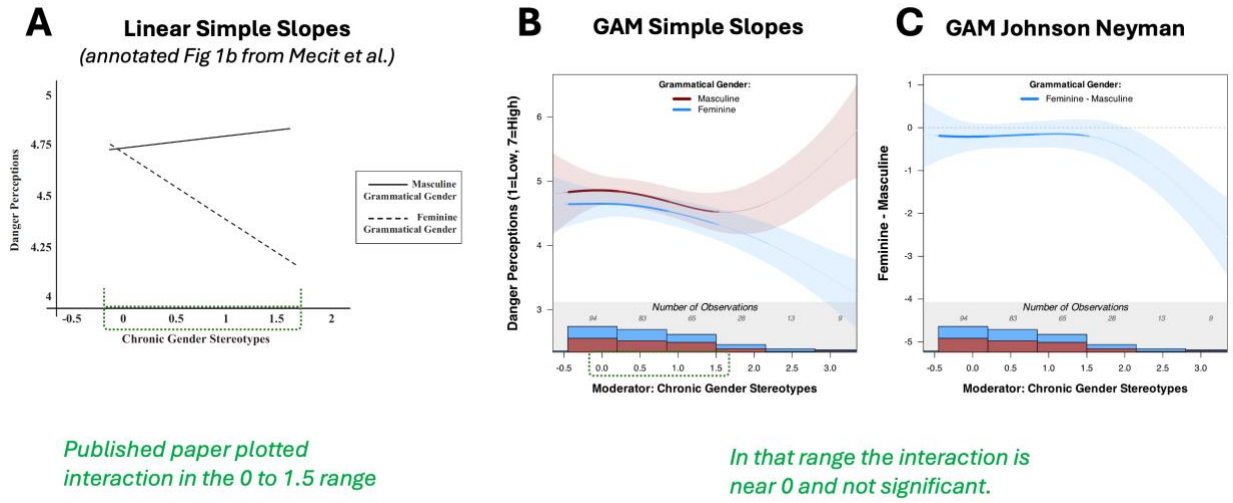

Figure 4. *The Interaction Is Non-Significant in Range of Values Plotted in Original Paper.*
Notes. Reanalysis of Study 3 by Mecit et al. (2022) on danger perception of COVID by French speakers (N=302) when relying on masculine "le" COVID vs. feminine "la" COVID grammatical terms. Panel A shows the (linear) 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. Note that the y-axes are not aligned across the three figures.

In turn, the Johnson-Neyman curve involves putting in the y-axis the *effect* of the focal predictor, for all values of the moderator.[3] In the context of a two-cell experiment, it is the vertical

[3] One sometimes plots the effect of the moderator values, for all possible values of the focal predictor.

difference between the two Simple Slopes. This applies to both linear and GAM Johnson-Neyman; the only distinction is whether the difference is between straight lines (linear regression) or lines that are not necessarily straight (GAM; Figure 4C).

Having reviewed how interactions are probed, we now return to the interpretation of the results from our first example. Recall that the authors found a significant interaction between the manipulation and chronic gender stereotyping ($p = .0015$), which they probed linearly. We obtained the data posted by the authors (https://osf.io/9437y) and successfully reproduced the results.[4] Importantly for our purposes, the original paper plots Simple Slopes for only the *subset* of the data with less extreme moderator values, between 0 and 1.5 (the data spans from -1.75 to +4.75). This is a sensible and defensible choice (focus readers' attention to where most of the data lie), but, as we shall see, it has interesting consequences for our purposes of illustrating the shortcomings of linear probing.

When we probed the interaction with GAM (Figures 4B and 4C), we found that, in this 0 to 1.5 range, there actually is neither an average effect nor moderation of the effect of interest, and that the interaction, that $p = .0015$, is entirely driven by more extreme values excluded from the original figure. As shown in Figure 4C, the confidence band for the GAM Johnson-Neyman curve includes zero in the entire aforementioned range [0-1.5], and it is associated with a significant effect only for more extreme moderator values.[5]

---

[4] 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.

[5] 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 -0.15, with a confidence interval which does not rule out values up to -0.46. Because the SD of the dependent variable is about 1, that's a Cohen's-d of about .46. The data, then, do not rule out effects of a considerable magnitude. At the same time, they do not rule out zero.

Intuitively, the original linear analysis over-estimates the effect among participants with moderate values by taking the larger effect produced by the more *extreme* participants, and distributing that effect linearly across *all* observations. This is because a linear model cannot accommodate a nonlinear effect. GAM, in contrast, allows for changes of functional form across moderator values, and thus does minimal projection and misplacement of effects from one region of values to the other.

When we relax the linearity assumption, the study's conclusions are updated in three interesting ways. First, the effect is driven by people with extreme values of chronic gender stereotyping, which hints at a different type of effect than if it were driven by more typical values. This is especially relevant if one is theoretically interested in more common and possibly implicit forms of gender stereotyping, as opposed to more extreme and overt forms. Second, from a purely descriptive perspective, a smaller share of the participants showed the effect of interest. While 52% of participants exhibited chronic gender stereotyping scores high enough to imply they showed the effect with linear Johnson-Neyman (moderator above 0.53), only 10% did for the effect implied by GAM Johnson-Neyman (moderator above 2.06). Third, because the effect is driven by more extreme observations, a closer look at potential measurement issues with those participants (inattention, demand effects, tendency to give high answers to every question, etc.) may be justified.

For ease of exposition, we focused the above discussion on significant vs. non-significant regions. The same qualitative contrast arises if we were to instead focus on estimated effect size, or their confidence interval.

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" (p.473) 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 eight magazine accounts, four belonging to "virtue magazines" (The Atlantic, Forbes, Health and The New Yorker) and four to "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).[6] 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.[7] We obtained data posted by the authors (https://osf.io/hya8z) and successfully reproduced the key results.[8]

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 sensibly and defensibly define the start of the day at 6 AM, we reproduce their findings. See left panel in Figure 5. However, if we define the start of the day at midnight, an alternative but in our mind equally sensible and defensible definition, a completely different pattern—one that lacks an interaction—is obtained; see middle panel in

[6] Functional form aside, the published analyses do not take dependence into account, treating observations as statistically independent.
[7] The authors do not indicate how they handled time zones, we use the posted data as is.
[8] We do obtain slightly different estimates for reasons we were not able to 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 authors find that before 12:01 PM liking was lower for vice magazines, and with our results it is 11:59 AM, just two minutes apart.

Figure 5. And if we define the start of the day at 4 PM, yet another completely different pattern arises (one that also lacks an interaction). See right panel in Figure 5.

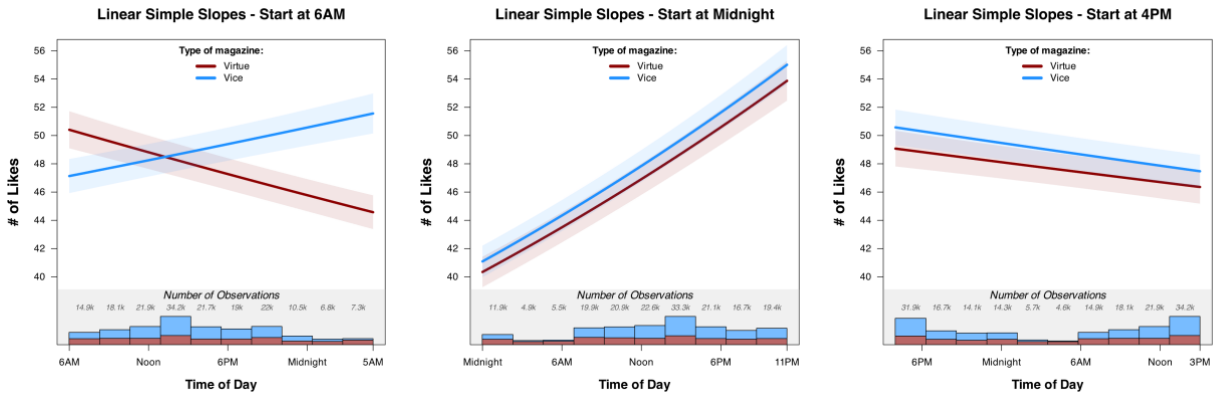

Figure 5. *Linear Probing Gives Inconsistent Answers for Different Definitions of Start of Day.*
Notes. Reanalysis of Study 1A by Zor, Kim, and Monga (2022) on number of *likes* (N=176,390) received by tweets posted at different times of day. Left panel reproduces the published results, middle and right panels depict how the probed interaction differs when using different start-day definitions with the same linear model.

You can start getting an intuition for why this happens, and for why *neither* model in Figure 5 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 rises discontinuously from the right end of the figure (which corresponds to 5:59 AM) to the left end of the figure (which corresponds to 6:00 AM). The increase for that single minute is (necessarily) of the same magnitude as the change for the remaining 1439 minutes. For example, for Virtue tweets, at 5:59 AM 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.[9]

[9] The original authors worried about linearity and present a quadratic regression as a robustness check. In Web Appendix 1, we show that this analysis does not address the fact that the results are sensitive to how the start of the day is defined.

When we analyze the data with GAM, see Figure 6, the results are not consequentially altered by whether the start of the day is defined at 12 AM, 6 AM, or 4 PM.[10] 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 10 PM and 4 AM 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 (of course, a linear model cannot estimate a cyclical relationship).

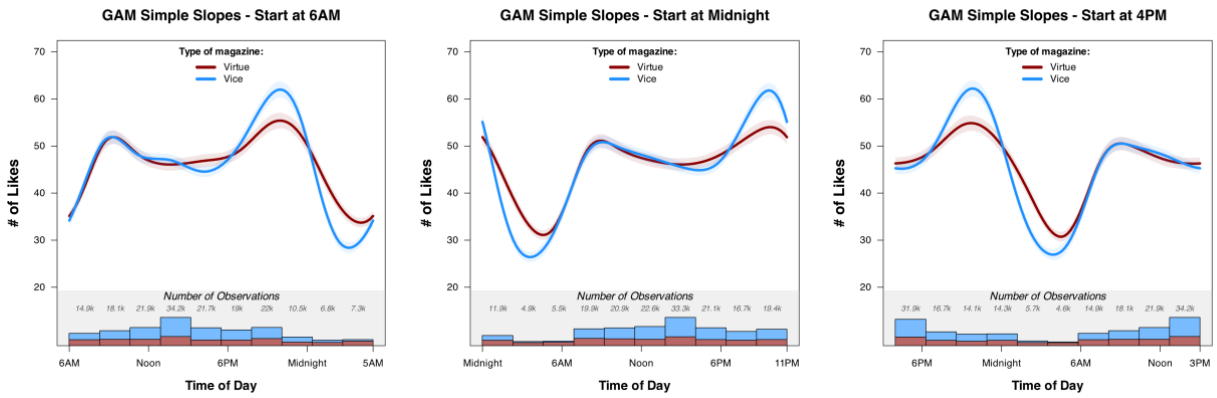

Figure 6. *GAM Probing is Robust to Different Definitions of Start of Day.*
Notes. Reanalysis of Study 1A by Zor, Kim, and Monga (2022) on number of *likes* (N=176,390) received by tweets posted at different times of day. In contrast to the linear model (see Figure 5) GAM probing leads to similar results no matter when the day is defined to start. All results are inconsistent with those in the published paper.

Example 3: 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

[10] In the GAM estimation we defined the data as a 'cyclical time series' so that it would take into account that 23:59 comes right before 01:00, but even without making this specification GAM recovers very similar hourly patterns regardless of what time of day is chosen.

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 measures aggregated into an index. Participants also completed a multi-item scale measuring how inter- vs. independent they were. The authors report a significant interaction, such that the effect of the frequently bought vs. loved manipulation was moderated by the degree of interdependence, p = .003. Re-analyzing the original data (https://researchbox.org/108) we successfully reproduced this result.

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

Concretely, as shown in Figure 7, left panel, the *linear* Johnson-Neyman curve implies that for moderator values below -1.31 (n = 11) there is a statistically significant reversal of the effect (the authors predict no effect among them). However, with the GAM Johnson-Neyman, right panel, 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 0.14 points, p = .704, in contrast to 0.65 points, p = .035 with the linear model).

Interestingly, at the lowest value of cultural orientation, the confidence interval of the GAM barely includes the point estimate of the linear model. The confidence band of the GAM is quite wide and does not reject a sizeable effect of either sign (nor zero of course). Our read is that the

surprising sign reversal obtained in the original analysis is a side-effect of the linearity assumption, and that *the data* do not conclusively support nor contradict such reversal. This example highlights that GAM can produce results that align better, not just worse, with the authors' hypotheses.

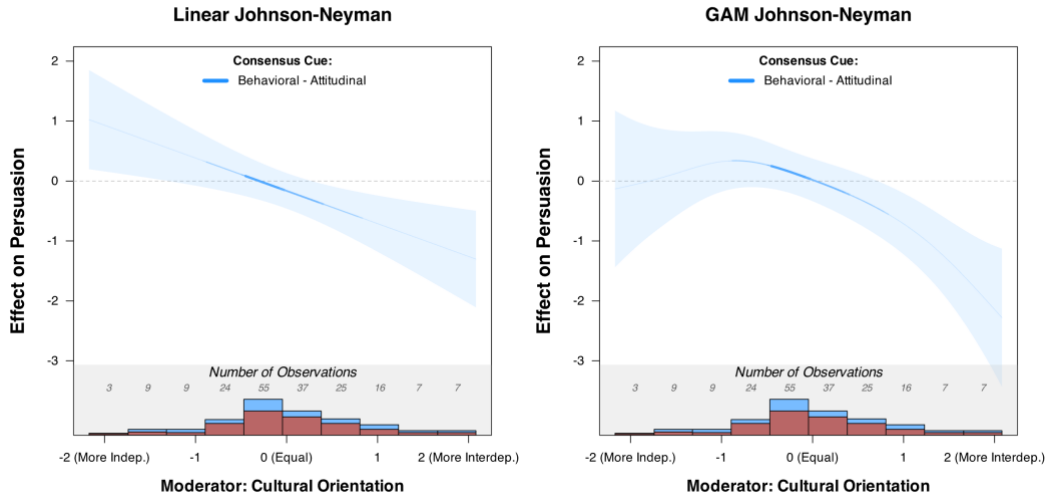

Figure 7. *Effect Reverses Only If We Assume Linearity*
Notes. Reanalysis of Study 5 by Barnes and Shavitt (2023) on how behavioral vs. attitudinal cues (being told that 81% of people bought vs. loved a set of headphones) affect product interest depending on participants' (N=128) cultural orientation. The significant reversal among more independent individuals is only obtained when forcing linearity on the data by estimating a linear model.

Example 4: Significant Estimate of the Wrong Sign

Woolley, Kupor, and Liu (2023) propose that consumers prefer high-tech products made by larger companies and low-tech products made by smaller companies. The paper includes one observational study (Study 1) and five experiments (Studies 2-6). Here, we focus on Study 1.

In their Study 1, 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 predicted "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).

Company size was measured by averaging the number of employees and revenues per company, and the tech-intensity of each company was based on an MTurk survey where participants evaluated companies on a 7-point scale (from 1=low tech to 7=high tech). The NPS data was obtained from "Customer Guru" (p.433).

The authors estimate a regression predicting NPS with company size, tech intensity, and the interaction, which was statistically significant, $p < .001$. The authors probed the interaction 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/vtfb2) and successfully reproduced these results (see Figure 8A).

What's interesting for us, given our focus on the probing of interactions, is that the reversal for low-tech firms, the 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, Figure 8B 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.

This result contradicts the abstract which states that "For low-tech products . . . quality evaluations and choice [move] in favor of smaller companies." (p.425). It may seem surprising that the linear regression shows a positive slope if the true functional form is U-shaped, especially when the negative slope among low moderator values is more pronounced than the positive slope for high moderator values. The explanation lies in the distribution of the moderator: there are very few observations with moderator values below 3. When minimizing squared errors in linear regression, the model sacrifices fit in regions with fewer observations (lower moderator values) to

better fit regions with more observations (higher moderator values), resulting in an overall positive slope.[11]

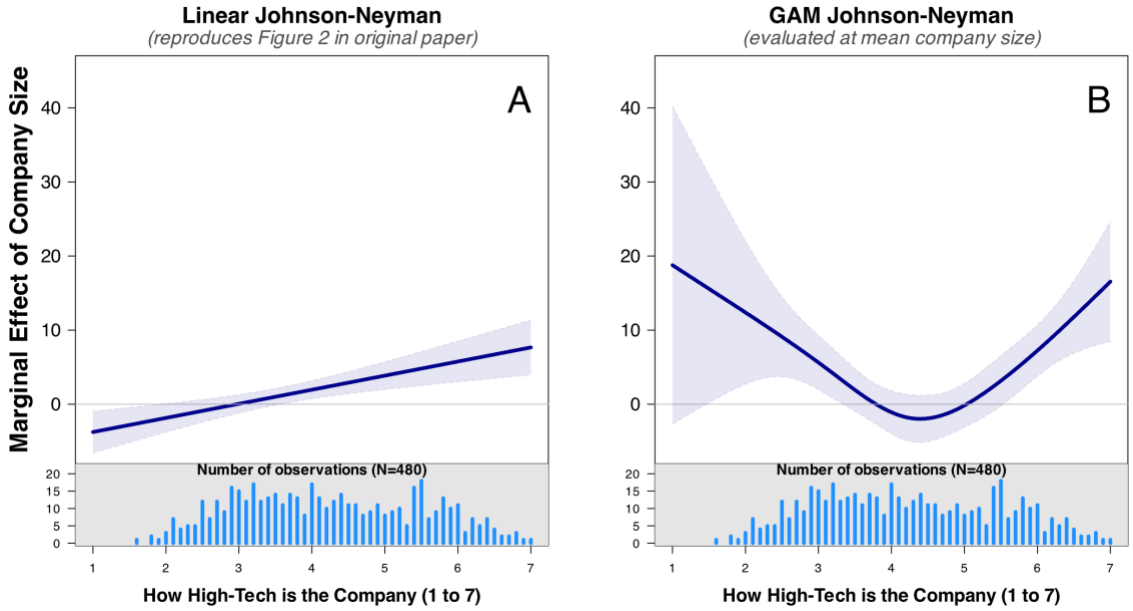

Figure 8. *Linear Model Estimates Spurious Reversal*
Notes. Reanalysis of Study 1 by Woolley et al. (2023) on how company-size (N=480 companies) predicts Net Promoter Score (NPS) for companies rated low- vs. high-tech (by sample of MTurk respondents). The pattern that among low-tech companies bigger companies get lower NPS is only obtained when forcing linearity in the model.

Note how the confidence band from the GAM adequately incorporates the uncertainty from the small sample size among low-moderator values, while the linear regression does not. The reason is that the linear regression assumes the true model is linear, and thus the slope in regions with little (or even no) data is inferred, with confidence, from the slope in distant regions with more data. The GAM, in contrast, uses only nearby data to make that inference and thus the confidence band widens much faster. 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

[11] In Web Appendix 2, we (i) further explore the data identifying significant outliers (e.g., Walmart is 15 SDs larger than the mean in company size), (ii) document that with robust standard errors the results change substantially, and (iii) report GAM probing results for the 25th, 50th, and 75th percentile of the moderator.

model (5.10 vs. 0.99 respectively). The GAM results are thus not only more accurate but also more honest about their uncertainty.

TWO FAQs ABOUT THESE FOUR EXAMPLES

In this section we discuss two questions raised by multiple members of the peer review team that we thought other readers may also wonder about.

On the Prevalence of Incorrect Linear Conclusions

In all four examples, relying on GAM instead of linear probing qualitatively altered the conclusions about the probed interaction. You may be wondering whether we are implicitly suggesting that all published interaction results would similarly change if the analyses were redone with GAMs. No. We are not suggesting that, and we do not believe that. Our examples are not a quantitative estimate of *how often* GAMs alter conclusions; they are a qualitative illustration of *how* GAMs *can* alter conclusions.

It is difficult to estimate the proportion of published interaction results whose conclusions would change if the data were reanalyzed with GAMs, for two main reasons. First, for most published marketing papers the data are not publicly available, precluding the necessary reanalyses, and the subset of papers whose data are publicly available is unlikely to be representative of all papers (open research practices are not randomly assigned to papers). Second, whether a change in results constitutes a qualitatively meaningful difference is inherently subjective (and, as we will explain, in our view, a qualitatively meaningful difference is always present).

To be concrete about this subjectivity point, let's return to our opening example of taco prices. Suppose that lowering the price increases demand, but at a decreasing rate. A linear regression would capture the monotonic relationship but not the decreasing rate part. How important is that part for authors or readers? Would identifying the decreasing rate, while not challenging the monotonic effect of price, count as a categorically different result where GAM outperforms the linear model?[12] In our view, the answer is yes. Learning about the functional form of documented associations is by itself an important aspect of every empirical investigation. Thus, even in cases where GAM returns a perfectly linear estimate—one indistinguishable from that obtained with the linear model—we still obtain a qualitatively better estimate, because we now *know* the association is (approximately) linear, instead of merely *assuming* that it is. This is similar to medical tests that come back negative; while they merely confirm one does not suffer from a feared condition, they do provide valuable information.

How Do We Know That GAM Is Right and Linear Is Wrong?

In our examples we show that conclusions change when relying on linear vs. GAM probing, but how do we know the GAM results are the correct ones? Well, we don't know this for certain, because nobody ever knows for certain that the results of any one analysis on any one dataset are correct. Statistical analysis does not deliver that kind of certainty. But we believe we can make a pretty strong case that the GAM results are much more likely to be correct. We make this case in two complementary ways.

[12] A related challenge is that reported interactions are often so noisily estimated that it is almost impossible to obtain results that are significantly different from them with another estimator, as the confidence interval around the point estimate includes all plausible values (of the same sign).

First, we follow the standard approach statisticians use to choose among competing procedures: relying on simulation results where (unlike real data) the true model is known. As discussed earlier, Simonsohn (2024) reports simulations showing that GAMs substantially and systematically outperform linear models when probing interactions. Absent a reason to expect an exception to this general result, the justified inference is that GAM is correct when it disagrees with linear. We can think of this first approach as deductive.

Our second approach to assessing whether the GAM or linear results in our examples are correct is more inductive. We return to the raw data and assess whether they better match the depiction implied by the GAM or the linear model. We apply this approach to the two examples involving experiments (Examples 1 and 3), where analyses of raw data are straightforward.[13]

The analyses involve focusing on subsets of observations. Specifically, on observations with moderator values for which the linear and GAM models disagree. We take the raw data, compute the condition means for those subsets of observations, and contrast the observed difference of means with that implied by the GAM and linear model.

In our first example, the linear results implied a statistically significant effect for moderator values above 0.53, while the GAM did so only for values above 2.06; the models thus disagree in the 0.53–2.06 range (43% of the sample). In our third example, the models disagree for moderator values below -1.31 (9% of the sample), where the linear results implied a statistically significant reversal of the effect, while the GAM an imprecisely estimated null effect. Figure 9 presents these comparisons. In both examples, we see that the GAM is much closer to the raw data than the linear model is (make sure to notice the very small sample, $n = 11$, in Example 3).

[13] Regarding the two examples with observational data: Figures 5 and 6 already attest to the relative superiority of the GAM model, for it survives alternative definitions of when a day starts while the linear model does not. We do not analyze the raw data for Example 4 in light of the presence of extreme outliers (15 SDs above the mean) discussed in our Web Appendix 2.

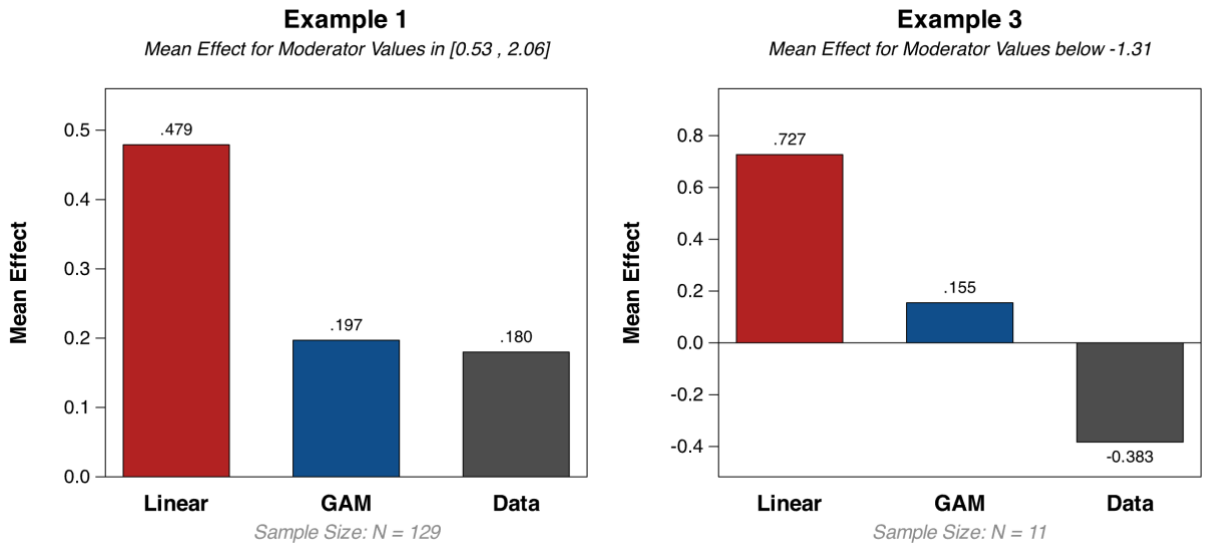

Figure 9. *Where Linear and GAM Disagree, the Data Side with GAM*

Notes. Bars depict the difference between conditions among participants in the region of moderator values where the linear model implies a significant effect but the GAM does not. The first two bars are the effects implied by the linear model and the GAM; the third bar is the raw mean difference.

PROBING INTERACTIONS WITH GAM - A TUTORIAL

In this section we provide step-by-step instructions for conducting, interpreting, and reporting analyses that probe interactions nonlinearly with GAMs.

Conducting the Analyses

Carrying out GAM probing naturally requires software. In this tutorial we provide instructions for two alternative software solutions. The first is the function `interprobe`, which we contributed to the R package `statuser`. The second is an online app we created: webstimate.org/interprobe . The app has a point-and-click interface (no coding required) and runs the same `interprobe` function, from the same `statuser` package, in the background. Thus, whether you run the analysis in R directly, or through the app, the results will be identical.

Relying on R, a single line of code carries out all the calculations and produces publication-ready figures. To be concrete, consider a hypothetical dataset named 'tacos_data' which contains three variables from a study on taco consumption: the spiciness of the taco (`spicy`), respondents' tolerance for spicy food (`tolerance`), and their enjoyment of the taco (`enjoyment`). To probe how tolerance moderates the effect of spiciness on enjoyment, one runs this single-line of code, where `x` designates the focal predictor, `z` the moderator and `y` the dependent variable:

```
interprobe(x='spicy', z='tolerance', y='enjoyment', data=tacos_data)
```

That single line generates publication-ready GAM Simple Slopes and GAM Johnson Neyman figures, and includes all statistics one would report in a paper. To run the same analysis with the web app, users (1) visit webstimate.org/interprobe, (2) upload the data, (3) point and click

to designate the variables `enjoyment`, `spicy` and `tolerance` as the dependent, focal, and moderator variables respectively, and (4) click the 'run' button.

The output produced by `interprobe` is different when the focal predictor takes only two or three values (e.g., when comparing conditions in an experiment) vs. more than three values (e.g., when assessing the 'effect' of a measured variable in an observational study). We thus discuss the presentation and interpretation of `interprobe` output separately for these two cases.

Case 1. Output for Dichotomous Focal Predictor (e.g., two-cell experiments)

Consider a stylized experiment where participants are randomly assigned to eat a taco whose salsa is either not spicy at all (1 on a 1-7 scale) or extremely spicy (7), and then report how much they enjoyed it (on a 1-100 scale). The moderator is participants' self-reported tolerance for spicy food (also on a 1-7 scale). After running the single line command in R, or relying on the web app, one obtains output like that shown in Figure 10 (which we manually annotated for this tutorial).

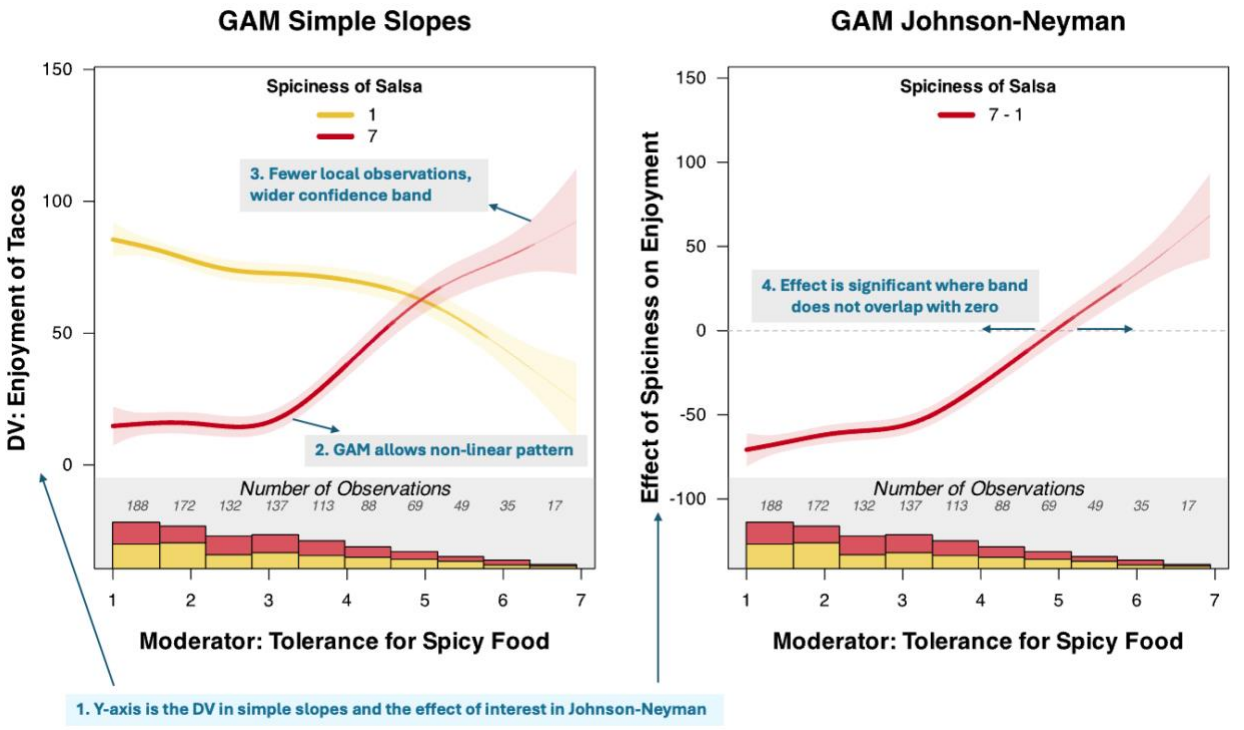

Figure 10. *Annotated Output of `interprobe` for a Binary Focal Predictor (e.g., experiment)*

Starting with the GAM Simple Slopes in the left panel, we observe that for people with low tolerance for spicy food (left end of the x-axis), the spicy taco is much less enjoyable than the non-spicy one. Whether the person has very low or simply low tolerance does not greatly impact this difference (in the 1-3 range), but then, as we cross the midpoint of 4, we see a marked increase in the relative liking of the spicy taco. Around 5 the pattern reverses, and the spicy tacos become more enjoyable than the non-spicy one. This is the kind of nonlinear relationship that a standard linear regression would mischaracterize. The pattern just described is, of course, an estimate, and thus contains uncertainty arising from sampling error. That uncertainty is reflected in the confidence band, which widens in regions of the x-axis where observations are scarce (see annotation 3), telling us where we can be more confident.

Moving to the right panel, the GAM Johnson-Neyman figure, note that the y-axis now shows the effect of the focal predictor. With a dichotomous predictor, that effect is simply the

difference between conditions, the vertical gap between the two simple slopes in the left panel (the effect of receiving a spicy vs non-spicy taco).

We see that the effect of receiving a spicy taco is negative for moderator values up to about 5, and positive beyond that. In other words, people with lower tolerance enjoy the spicy taco less than the non-spicy one, while people with higher tolerance enjoy the spicy taco more. The confidence band in this second panel reflects the uncertainty of both simple slopes being compared, and thus conveys the statistical significance of their difference. As annotation 4 states, the effect is statistically significant wherever the confidence band excludes zero.

Simple Slopes plots are useful for understanding general patterns in the data, while Johnson-Neyman plots focus our attention on the effect of interest, the difference between conditions. For experiments, where the focal predictor (the manipulation) is binary, both plots are easy to interpret and complement one another, so it is useful to report both. (We will argue that for continuous predictors, Case 2, Johnson-Neyman plots are harder to interpret and less informative.)

The `interprobe` function reports, in addition to these figures, both in R and the web app, results for statistical inference. First, it reports two overall tests of the interaction, one relying on the linear model and the other on GAM. Then, it reports the "regions of significance", the subset of moderator values for which the effect of the manipulation is significantly different from zero. Those regions are visually depicted in the Johnson-Neyman figure, but the text provides their precise starting and ending points. Figure 11 shows this output for the same example.

```
Probing the interaction of 'spicy' * 'tolerance'
p-value for the interaction:
  linear model: t(996) = 18.21, p < .001 (robust errors: HC3)
  GAM: F(3.96, 4.82) = 87.46, p < .001

Johnson-Neyman Regions of Significance
1) The contrast for 'spicy' of 7 - 1 is negative in the range of 'tolerance': [1 to 4.72] (85.1% of the sample)
2) The contrast for 'spicy' of 7 - 1 is positive in the range of 'tolerance': [5.2 to 6.94] (9.5% of the sample)
```

Figure 11. *Statistical Results of `interprobe` for a Binary Focal Predictor*

A template paragraph describing these results in a paper may read: "The interaction between tolerance for spicy food and the spiciness of the salsa in the taco proved significant in both the linear, $t(996) = 18.21$, $p < .001$, and GAM model, $F(3.96, 4.82) = 87.46$, $p < .001$. Probing the interaction with GAM showed that the effect of spiciness was negative and significant for moderator values below 4.72 (the bottom 85% of the sample), and positive and significant for values above 5.20 (the top 10% of the sample)."

Case 2. Output for a Continuous Focal Predictor (e.g., observational data)

Instead of a hypothetical experiment where spiciness is randomly assigned to be either 1 or 7, consider a hypothetical study where salsa can take any level of spiciness between 1 and 7. This second case merits separate consideration because the results of probing the interaction are much richer. In Case 1 the focal predictor takes just two values (1 or 7), so for any moderator value there is one effect size to estimate: conceptually, the difference of means between the two conditions for the subset of participants with that value of the moderator. In Case 2, the focal predictor can take many values, so for any given moderator value there are many effects to estimate: e.g., the effect of a taco going from 1 to 2 in spiciness, another for going from 2 to 3, another from 3 to 4, and so on.[14]

For concreteness, imagine there are 50 possible moderator values. In Case 1, with a dichotomous predictor (taco with spicy level 7 vs 1), there are 50 effect sizes to estimate, one for each moderator value. In Case 2, where there are more than two possible values of spiciness, there are more effects to estimate. For example, say there are 100 possible spiciness levels; then there

[14] The variable is continuous so the effect can also be different going from 1 to 1.1 vs from 1.1 to 1.2, etc.

would be 100 × 50 = 5000 effects of spiciness to estimate. This is not literally what GAM does, it does not literally estimate 5000 effects independently (it smooths effects for proximate predictor values), but it does help visualize how much more data-demanding the task is, and how much richer the output will be.

Because a full function is now estimated for the focal predictor, not just a single difference of means, it becomes convenient to change what `interprobe` puts on the x-axis. Specifically, it puts the focal predictor (taco spiciness) on the x-axis instead of the moderator (tolerance for spicy food), and plots three curves, one for each of three moderator values. By default, the 15th, 50th, and 85th percentiles of the moderator (for a normal distribution, these correspond roughly to the mean and ±1 SD). For continuous focal predictors we believe the Johnson-Neyman curves, both linear and GAM, are difficult to interpret and we thus recommend relying only on Simple Slopes to probe interactions when a focal predictor is continuous (for details see Web Appendix 3).

All these changes happen in the background. From the user's perspective, the exact same single line-command is run in R, or the exact same point-and-click actions are taken on the web app, whether the focal predictor is binary or continuous, and `interprobe` produces the most informative figures by a (modifiable) default. Figure 12 shows the Simple Slopes plot for this example.

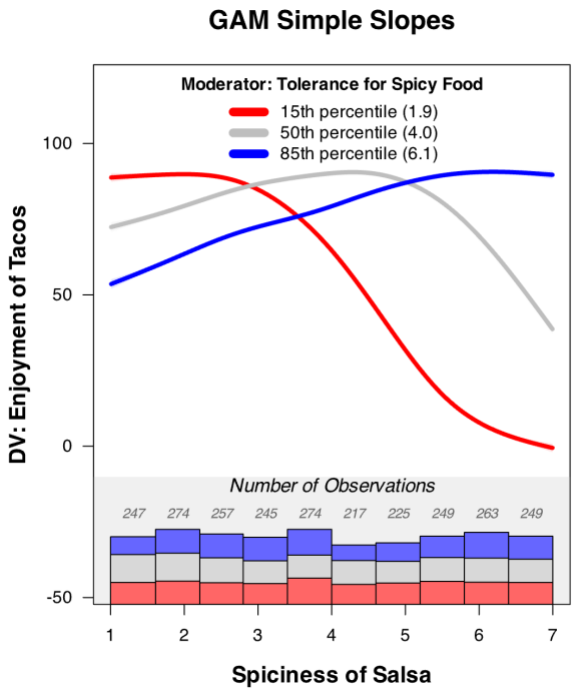

Figure 12. *Sample Output of `interprobe` for a Continuous Focal Predictor (e.g., observational data)*

The Simple Slopes in the figure depict the relationship between how spicy the taco is and how much people enjoy it, at the three default moderator values. The results are intuitive. The red line shows that people with low tolerance for spicy food much prefer mild tacos (it falls steeply down from left to right, with tacos rated 1 on spiciness enjoyed at around 90 and those rated 7 on spiciness enjoyed at about 0). The blue line shows the opposite for people with high tolerance, though their dislike of bland tacos is milder than the low tolerance group's aversion to spicy ones (the blue line never drops as low as the red one does). Finally, the gray line shows that people with moderate tolerance enjoy moderately spicy tacos the most.

A Caution When Interpreting Interactions With Observational Data

We caution that when the focal predictor is continuous, as in this second case, the data will typically be observational rather than experimental, and in such cases two important caveats apply. First, with non-experimental designs, an important question for the authors to address is how the association they report should be interpreted. If the authors lack a credible identification strategy—meaning they do not provide a specific mechanism that would justify treating the association as causal—a second question for the author arises: why does an association of uninterpretable causal origin provide valuable information?

Second, in addition to the issue of causality, which is a matter of interpretation, non-experimental data also poses a statistical challenge: when two predictors, x and z, in the interaction $x \cdot z$ are correlated ($r(x,z) \neq 0$), as they typically are in observational data, and either has a nonlinear effect on y, as is typical of any data, we expect the *linear* model's estimate of the interaction to be biased (Cortina 1993; Ganzach 1997, 1998; Lubinski and Humphreys 1990). If one "must" report a linear regression result, one should always include quadratic terms as controls for both factors in the interaction (Ganzach 1997; Simonsohn 2024). While GAMs are not biased by a correlation between factors in the interaction, one should rely on bootstrapping for statistical inference, i.e., computing p-values (see Figure 11 in Simonsohn 2024). The current version of `statuser` does not offer this bootstrap correction, but we will add it before this paper is published.

CONCLUSION

The main reason to probe interactions linearly is that it is what we have been doing for 90 years. The main reason to probe interactions with GAMs is that the results will be richer, allowing us to learn more from the data we collect, and more robust, reducing the chances that we arrive at wrong conclusions because of the arbitrary and usually false assumption that all relationships are linear.

REFERENCES

Académie française (2020), "Le Covid-19 Ou La Covid-19 ?," https://web.archive.org/web/20200515113159/http://www.academie-francaise.fr/le-covid-19-ou-la-covid-19.

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*.

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 Il Error in Detecting Interaction," *Journal of Management*, 24 (5), 615-22.

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), 1-22.

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*, Chapman & Hall.

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.

Zor, Ozum, Kihyun Hannah Kim, and Ashwani Monga (2022), "Tweets We Like Aren't Alike: Time of Day Affects Engagement with Vice and Virtue Tweets," *Journal of consumer research*, 49 (3), 473-95.