This verion: 2025-03-21
Curent version: https://urisohn.com/46

Johnson-Neyman 2.0: More Flexible, More Robust, and More Interpretable Probing of Interactions

Andres Montealegre
Yale University
aandres.montealegremoreno@yale.edu

Uri Simonsohn
ESADE Business School
urisohn@gmail.com

Abstract
We re-analyze data from four recent papers to demonstrate that the traditional approach for probing interactions—linear regression followed by Simple-Slopes ("spotlight") or the Johnson-Neyman procedure ("floodlight")—can lead to qualitatively incorrect conclusions when true relationships are not linear. Building on proposals to use GAMs (Generalized Additive Models) to probe interactions, we introduce the Johnson-Neyman 2.0 procedure (JN2). This procedure involves: (1) using GAMs to probe an interaction of interest, and then (2) verifying the key conclusion of interest with a t-test or linear regression run on the subset of relevant data. For example, if a GAM determines an experimental manipulation has a negative effect for moderator values 0.67 SD below the mean, a t-test is run only among those observations to verify the negative effect.

Keywords: interactions, probing, Johnson-Neyman, 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 two (or more) other variables, are commonly examined in marketing research. Researchers ask questions like: Does a consumer trait or a product characteristic make people more price sensitive? After testing interactions, after determining whether there is or there isn't one, it is common to 'probe' them, to assess the effect size of the variable of interest for different moderator values, in one of two ways: (1) estimating the 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 (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; see Spiller et al. (2013).

The technology that is currently relied upon 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 (linear) probing 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 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 their future.

Returning to our opening example, 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 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 do not 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.

A natural solution to avoiding the bias that arises from assuming linearity is to not assume linearity, relying on more flexible models like Generalized Additive Models (GAMs). Though GAMs are rarely used in social science, they have existed for nearly four decades (Hastie and Tibshirani 1987). Unlike linear regressions, GAMs estimate rather than assume the functional form of the relationship between variables in the model—allowing them to capture many kinds of non-linear effects while incorporating a penalty for overfitting.

The added flexibility obtained with GAMs, however, comes at a cost to interpretability. While linear regressions produce a single interpretable coefficient for each predictor, GAMs provide many uninterpretable coefficients for each predictor that, when combined, produce the estimated functional form for such predictor. GAMs require visualizations to be interpreted. This means that there are no "regression tables" with GAMs, no simple numerical summaries that allow

researchers and readers to determine whether an expected pattern is or is not observed. In addition, from questions we have received when presenting this work, and conversations with colleagues more generally, we know that there are concerns about possible overfitting, and sensitivity of black box results to unknown assumptions.

In this paper, we propose addressing these limitations in the GAM probing of interactions with a procedure that combines the interpretability and transparency of linear models, with the robustness and flexibility of GAMs. We refer to it as *Johnson-Neyman 2.0* (JN2). It involves three steps: First one probes an interaction using GAM, rather than linear regression, generating a figure that depicts either GAM Simple Slopes or the GAM Johnson-Neyman procedure (Simonsohn 2024). Second, one identifies the critical patterns in the figure that support the main conclusions of the probing exercise (e.g., regions of moderator values where the effect of the manipulation shows qualitatively different magnitudes or reverses sign). Third, one analyzes the subset of the data for which a qualitative conclusion is obtained, relying on traditional tools like the t-test or linear regression.

For example, let's say the GAM Johnson-Neyman procedure indicates that the effect of a randomly assigned manipulation becomes negative for moderator values lower than 0.67 standard deviations below the mean. With JN2, one then runs a t-test on that subset of the data, obtaining a single point estimate (the difference of means) and associated *p*-value. JN2 provides a familiar and evaluable confirmation of the identified pattern. Similarly, if the key conclusion were that as the moderator values keep dropping past 0.67 SDs, the effect of the manipulation keeps getting larger in magnitude, one runs a *regression* on that subset of data and establishes whether the slope for the moderator is significantly negative as expected, obtaining again a single interpretable point estimate (the slope) and *p*-value.

JN2 provides three valuable contributions to the existing proposal of relying on GAMs for probing interactions (Simonsohn 2024). First, interpretability: JN2 provides a single interpretable numerical summary instead of intricate graphical depictions, making empirical results straightforward to communicate. Second, accessibility: JN2 serves as a transition tool that allows researchers to switch from linear to GAM probing while still relying on a tool that all readers can understand and critically evaluate. Third, robustness: JN2 addresses the skepticism toward unfamiliar black-box tools like GAMs by verifying results using a familiar tool.

The most notable limitation of the JN2 approach we propose arises with non-experimental data, or more specifically, when x and z in the x·z interaction could be correlated. In that case linear regressions can be invalidated if the underlying data are not linear (see e.g., Ganzach 1997), and this is also true when the regression is run on a subset of the data, as is done with JN2.

We make the case for GAM probing in general, and JN2 in particular, by revisiting four recently published marketing articles, where authors relied on linear Simple Slopes and/or linear Johnson-Neyman procedures to probe interactions. After reproducing the published results with the original data, we demonstrate how the inferences based on linear probing of the interactions are partially or entirely reversed when the arbitrary linearity assumption is relaxed.

In Example 1 an effect that supposedly is present for most values of the moderator, is actually only present for (extreme) high values of it. In Example 2 the linear estimates bear almost no resemblance to reality. In Example 3 an effect that supposedly reverses for low moderator values appears to merely attenuate. Finally, in Example 4 the linear model underestimates the magnitude of the overall interaction, and obtains a statistically significant effect of the wrong sign. This fourth example is useful also for illustrating the limitations of JN2 with observational data.

Part of the appeal of the currently universally-used linear approach is its simplicity. Fortunately, switching to JN2 will not increase the difficulty of probing interactions for researchers. We have created an R package, `*interacting*`, which probes interactions with GAM and produces ready-to-publish figures in just one line of code. The follow-up JN2 can then be conducted using familiar tools—traditional t-tests and linear regression commands—performed on data subsets.

An intuitive description of GAM estimation

In this section, we explain GAMs in broad terms so that readers can gain a basic understanding. We do not delve into details that are aptly covered in various statistics articles.[1]

GAM vs. Regression

GAMs resemble linear regressions in that they 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 for the relationship with each independent variable.[2]

In a 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 x and z and their interaction, as in $y=f_1(x)+f_2(z)+f_3(x,z)$. A GAM 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

[1] Readers wishing to learn more about GAMs may consult the introduction by Simonsohn (2024), the textbook by Wood (2017), the seminal article on GAMs by Hastie and Tibshirani (1987), or the introductory article by Beck and Jackman (1998).

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

different x and z 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 in the estimated function, 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. Indeed, later in the article (Figure 7) we show that if the true model with an interaction is fully linear, probing the interaction with a linear regression or with GAM leads to similar results.

As mentioned in the introduction, the main downside of GAMs is that they do not produce few interpretable 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 the social science toolbox so far. This challenge can be tackled in two ways.

First, as proposed by Simonsohn (2024), one can *probe* estimated GAMs in a manner analogous to probing interactions in linear models. This involves using the fitted GAM model to calculate predicted values for the dependent variable and predicted marginal effects for a given combination of predictor values. These predicted values can then be visualized, similar to how linear Simple Slopes and linear Johnson-Neyman are currently plotted for linear models—except that lines are replaced with curves. Second, building on this proposal, and following the JN2 procedure we propose here, the key findings identified in these figures can be verified and numerically summarized into a single estimate and *p*-value using traditional statistical tools such as t-tests and linear regression.

Illustration of GAM estimation

We report results from a simple simulation below to illustrate the greater accuracy 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 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.

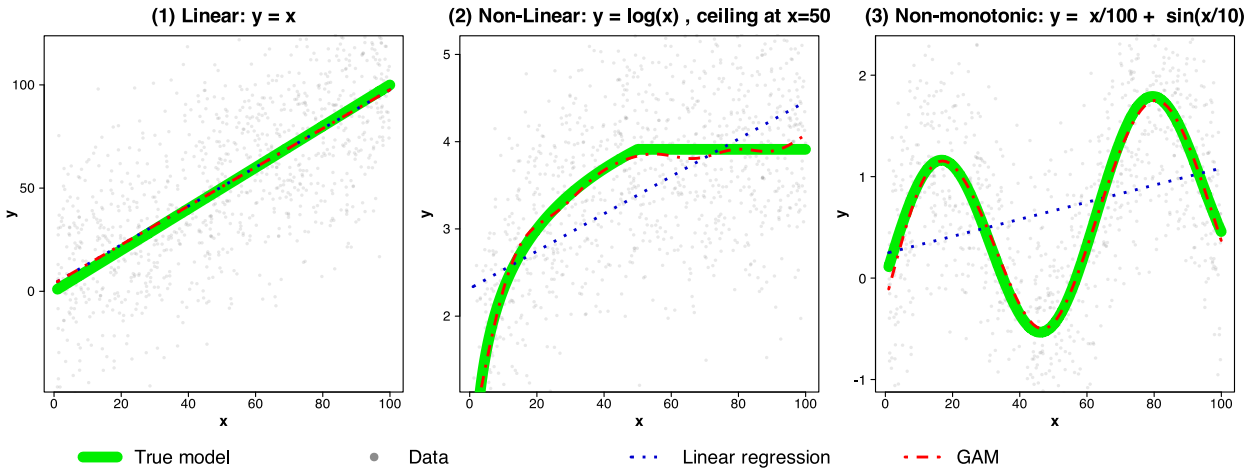

Figure 1. *Comparing Model Adequacy: GAM vs. Linear Regression.*
The figure depicts a single simulation of 1000 observations where x~U(0,100). The figure shows that when the true model is linear GAM recovers a linear model, but when it is not, it accurately captures the alternative functional forms. 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. In our first example we also provide an explanation for how

one obtains Simple Slopes and Johnson-Neyman estimates from a GAM, aided by the concreteness of the data at hand.

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).[3]

In their Study 3, 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), which was used as a moderator in the analysis.

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.[4] In Figure 2, we reproduce their linear Simple Slopes, followed by our GAM based probing.

Before interpreting these results, we use the figure 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 2A, are computed by estimating a linear

[3] The Académie française (2020) determined that the proper grammatical term is "La COVID" rather than "Le COVID". After this determination, formal sources tended to use "Le COVID" while "La 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.
[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, but the effect is also significantly negative for moderator values below -1.65.

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 2B).

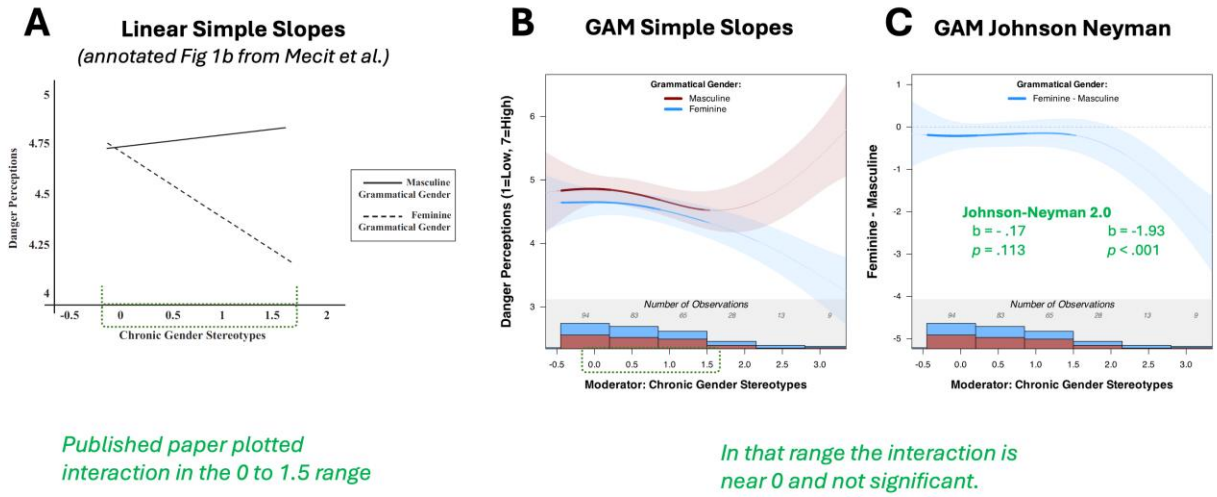

Figure 2. *The Interaction Is Non-significant in Range of Values Plotted in Original Paper.*
Notes. Reanalysis of Study 3 by in Mecit et al. (2022) on danger perception of COVID by French speakers (N=302) when relying on masculine "le" COVID vs. feminine "la" COVID grammatical terms. Panel A shows the (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.
Code to reproduce Panels B &C: https://researchbox.org/2859/30 (enter code QAMDS)

In turn, the Johnson-Neyman curve involves putting in the y-axis the *effect* of the focal predictor, for all values of the moderator.[5] In the context of a two-cell experiment, it is the vertical 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 2C).

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

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.[6] It is worth noting that the original paper plots Simple Slopes for only a *subset* of the data with less extreme moderator values, between 0 and 1.5 (the data spans from -1.75 to +4.75). When we probed the interaction with GAM (Figures 2B and 2C), we found that, in this range, there actually is neither an average effect nor moderation of said effect, and that the interaction, that $p = .0015$, is entirely driven by more extreme values excluded from the original plot. As shown in Figure 2C, 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.[7]

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.

---

[6] 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.

[7] 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 -0.15, 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.

GAM finds a significant effect only for moderator values above 2.06. We follow-up with Johnson-Neyman 2.0 (JN2). We split the data into two subsets, one above and one below 2.06, and test the effect of the manipulation in each range with simple t-tests. Consistent with the GAM results, for moderator values below 2.06, the effect of the manipulation is small and not significant, $M_{La} = 4.56$ vs. $M_{Le} = 4.73$, $t(267.34) = 1.59$, $p = .113$, while for moderator values equal to or above 2.06 the difference is more than 10 times larger and highly significant, $M_{La} = 3.54$ vs. $M_{Le} = 5.47$, $t(26.32) = 3.72$, $p < .001$.

In terms of the implications of these differences in results, the conclusions from the study are interestingly updated when we stop imposing the linearity assumption. First, the effect is driven by people with extreme values of sexism. This is especially relevant if one were theoretically interested in more common and possibly implicit levels of sexism, as opposed to more extreme and overt levels. That an overall finding is driven by extreme observations hints at a potentially different type of effect than if the overall finding were driven by more typical observations. 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 linear Johnson-Neyman (moderator above 0.53), only 10% did for the effect implied by GAM Johnson-Neyman (moderator above 2.06). 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, tendency to give high answers to every question, etc.) may be justified (we do just that in Supplement 1).

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. For example, JN2 could include not only the range of moderator values with a significant effect, but with a directional positive 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" (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).[8]

In their Study 1A they analyze 176,390 tweets from eights 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).[9] 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.[10] We obtained data posted by the authors (https://osf.io/hya8z) and successfully reproduced the key results.[11]

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

[8] The phrase "when morning turns to evening" appears 7 times in the paper.
[9] Functional form aside, the published analyses do not take dependence into account, treating observations as statistically independent; we do not think the reported significance results are valid.
[10] The authors do not indicate how they handled time zones, we use the posted data as is.
[11] 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 11.59AM, just two minutes apart.

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 —one that lacks an interaction— is obtained; see middle panel in Figure 3. And if we define the start of the day at 4PM, yet another completely different pattern arises (one that also lacks an interaction). See right panel in Figure 3.

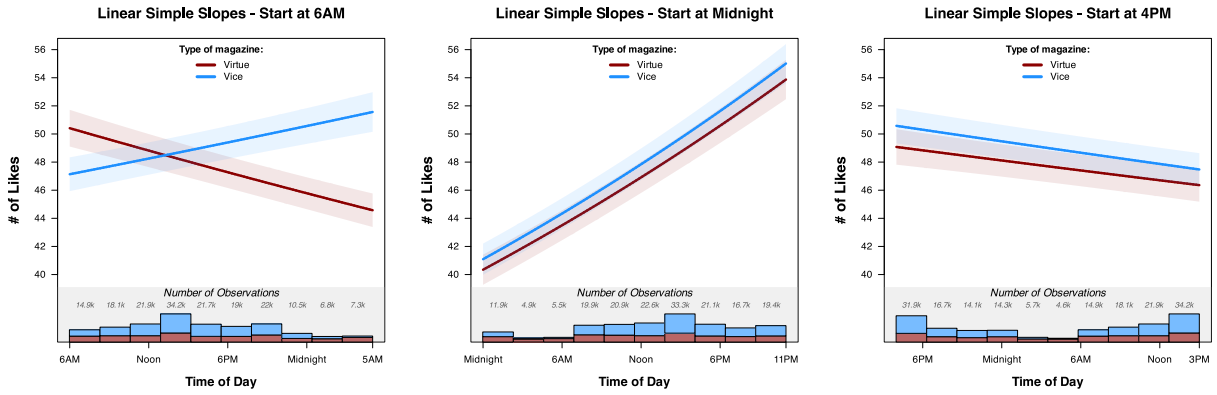

Figure 3. *Linear Probing Gives Inconsistent Answers for Different Definitions of Start of Day*
Notes. Reanalysis of Study 1A by in Zor, Kim, and Monga (2022) on number of *likes* (N=176,390) received by tweets posted at different times of day. Left panel reproduces the published results, middle and 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 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.[12]

When we analyze the data with GAM, see Figure 4, the results are not consequentially altered by whether the start of the day is defined at 12AM, 6AM, or 4PM.[13] 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 (of course, a linear model cannot estimate a cyclical relationship).

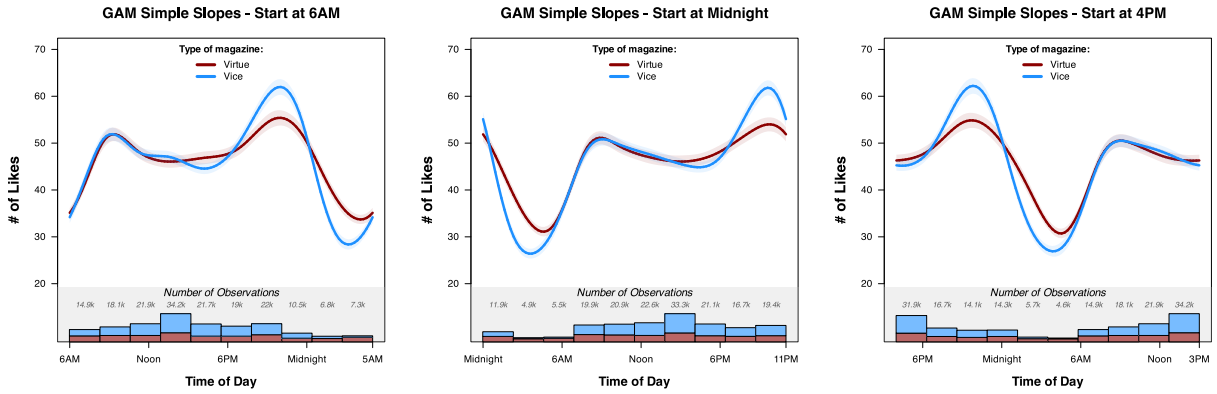

Figure 4. *GAM Probing is Robust to Different Definitions of Start of Day.*
Notes. Reanalysis of Study 1A by in Zor, Kim, and Monga (2022) on number of *likes* (N=176,390) received by tweets posted at different times of day. In contrast to the linear model (see Figure 3) GAM probing leads to similar 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)

The GAM results suggest four distinct periods in terms of how time correlates with engagement. We follow up our description of the GAM results with JN2, estimating separate linear

[12] The original authors worried about linearity and present a quadratic regression as a robustness check. In Supplement 2, we show that this analysis does not address the fact that the results are sensitive to how the start of the day is defined.
[13] 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.

models for the four periods with distinct overall slopes (reported in parenthesis following the description of each period). The four time trends being, (1) increasing from 5AM to 10AM, ($\hat{b}$ = 8.78), (2) flat from 10AM to 6PM ($\hat{b}$ = -0.39), (3) increasing from 6PM to 10PM ($\hat{b}$ = 5.68), and (4) decreasing from 10PM to 5AM ($\hat{b}$ = -10.00).[14]

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 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 onto 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 that is consistent with the abstract, the linear interaction reported by the authors implies

[14] We do not report statistical significance because the data are heavily dependent, a cyclical time series where there is correlation across tweets at the same time (due to time shocks), and across times. We thus report the coefficient for illustrative purposes only. We multiplied the coefficients by 100 to facilitate their comparison within the paragraph.

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 Figure 5, left panel, 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 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 .65 points, $p = .035$ with the linear model).

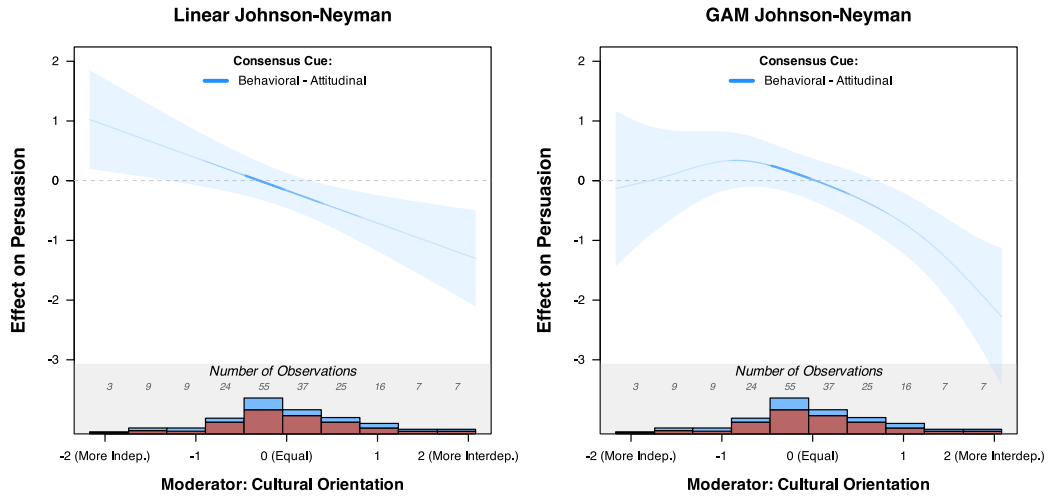

Figure 5. *Effect Reverses Only If We Assume Linearity*
Notes. Reanalysis of Study 5 by Barnes and Shavitt (2023) on how on how behavioral vs. attitudinal cues (being told that 81% of people bought vs. loved a set of headphones) affect product interest depending on participants' (N=128) cultural orientation. 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/2859/42 (enter code QAMDS)

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 arbitrary linearity assumption, and that *the data* do not conclusively support nor contradict such reversal.

We follow up these analyses with JN2 tests. First, we apply JN2 to the results reported in the original paper, which relied on *linear* Johnson-Neyman. Specifically, the paper reads "the effect of behavioral vs attitudinal consensus cues . . . was negative and significant . . . [when the moderator was] at or above 0.28 . . . [and] positive and significant . . . [with moderator values] at or below -1.31" (p. 11). Beginning with the former result, a simple t-test on the subset of data with moderator values above 0.28 verified the negative effect of the behavioral cue, $M_{\text{attitudinal}} = .41$ vs. $M_{\text{behavioral}} = -.18$, $t(37.21) = 2.47$, $p = .018$. Continuing with the latter result in the original Johnson-Neyman analysis, a simple t-test on the subset of data with moderator values below -1.31 did not verify the sign reversal, as the behavioral cue was directionally still *lower*, $M_{\text{attitudinal}} = -.87$ vs. $M_{\text{behavioral}} = -1.26$, $t(6.98) = 1.10$, $p = .309$. The JN2 results are thus consistent with GAM Johnson-Neyman but not with linear Johnson-Neyman.

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 participant

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/hya8z/) and successfully reproduced these results (see Figure 6A).

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, Figure 6B 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.[15]

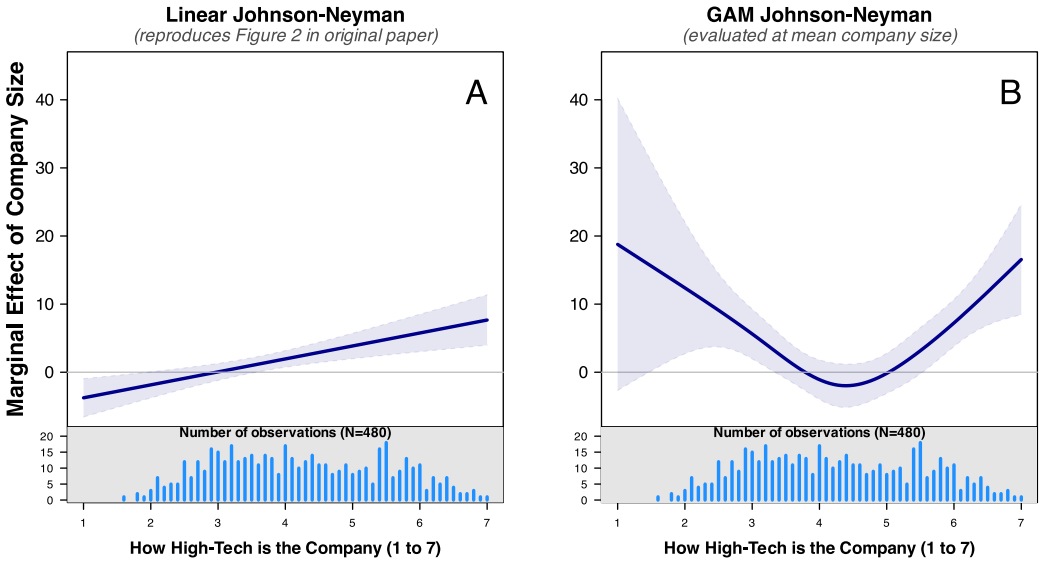

Figure 6. *Linear Model Estimates Spurious Reversal*
Notes. Reanalysis of Study 1 by Woolley et al. (2023) on how company-size (N=480 companies) predicts Net Promoter Score (NPS) for companies rated low- vs. high-tech (by sample of MTurk respondents). The patterns that among low-tech companies bigger companies get lower NPS is only obtained when forcing linearity in the model.
Code for figure: https://researchbox.org/2859/43 (enter code QAMDS)

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

[15] In Supplement 3, 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 providing better knowledge and better meta-knowledge.

In the previous three examples we relied on JN2 to complement results obtained from GAM probing. We do not do so for this example because, with observational data, and especially when x and z in the x·z are non-dichotomous, a linear regression, even when estimated on a subset of the data, can be biased by non-linearities (Ganzach 1997). In addition, in this particular dataset, there are severe outliers (e.g., one of the companies, Walmart, is 15 standard deviations larger than the mean), which make average regressions especially difficult to interpret (see Supplement 3).

Do GAMs Need Larger Sample Sizes?

Researchers often intuit that flexible models like GAMs, require much larger sample sizes than do linear models to be informative. This, however, does not seem to be the case. To illustrate, we present results from a simulation comparing the precision of linear regression and GAM when relying on small samples.

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 noise. We consider a researcher interested in estimating a model with the x·z interaction, and probing the effect of x at the 15th vs. 85th percentile of z.[16] We run 5,000 simulations adding 9 times as much variance from noise as there is variance produced by the x·z interaction, and another 5,000 simulations with 99 times as much noise as there is signal. This means that the highest possible R^2 is 10% and 1% respectively. Figure 7 depicts the distribution of estimated probed effects.

[16] For normally distributed data, the 15th and 85th percentiles are roughly the mean plus/minus one SD.

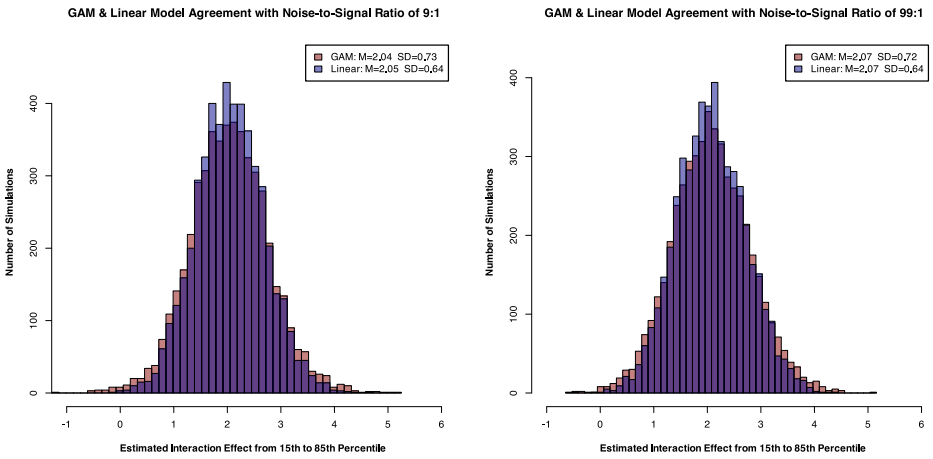

Figure 7. *When True Model is Linear, Linear Regression and GAM are Similarly Precise*
Note: Distributions of estimated changes in the effect of x on y when the moderator goes from 15th to 85th percentile. The true model is $y = x + z + x \cdot z + e$, where x is randomly assigned (1 vs. 0), z is a standard normal moderator, and *e* represents Normal noise. In Panel A, the error accounts for 90% of the variance, whereas in Panel B, it accounts for 99% of the variance. Sample sizes are n=100 and n=1000 respectively, achieving power around 90%.
Code to reproduce figure: https://researchbox.org/2859/6 (enter code QAMDS)

We see that the mean estimates for GAM and linear regression are nearly identical, and the variation in estimates of the probed interaction increase by about 15% with GAM. In Supplement 4, we show similar simulations for skewed distributions of the moderator.

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 and support erroneous conclusions from data. Our reanalysis of four recently published papers demonstrates that the problems associated with assuming linearity are not just a theoretical possibility, but a practical reality.

Our proposed Johnson-Neyman 2.0 procedure allows researchers to relax the linearity assumption by combining GAM's flexibility with the interpretability of traditional statistical tools. By first visualizing interactions using GAM and then confirming key findings with targeted tests on relevant data subsets, JN2 provides researchers with clear numerical summaries that are

straightforward to communicate and evaluate—and that serve as a valuable robustness check. We see no realistic circumstances where linear probing of interactions—the approach currently adopted in virtually all papers—would be preferable to GAM probing of interactions. We see the transition from linear to non-linear probing as unavoidable; JN2 should facilitate that transition.

JN2 has an important limitation that must be emphasized: it is only valid when the predictor and moderator of interest (x and z in x·z) are independent—a condition typically met in experiments but not in observational data. When x and z are correlated, non-linearities in their impact on the dependent variable lead to bias in the estimates of the linear interaction (Cortina 1993; Ganzach 1997, 1998; Lubinski and Humphreys 1990), making regressions run on subsets of data less compelling for evaluating a GAM model's robustness. For this reason, we do not rely on JN2 in Example 4, and in Example 2 we use it only to establish a main effect of tweets being liked throughout the day, rather than the interaction with vice vs. virtue magazines.

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*.

Beck, Nathaniel and Simon Jackman (1998), "Beyond Linearity by Default: Generalized Additive Models," *American Journal of Political Science*, 596-627.

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.

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.

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.

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.