First draft: 2021 11 25
This draft: 2021 11 25
Newest draft: http://urisohn.com/42

Interactiongate: Testing and Probing Interactions with Linear Models in the Real (Nonlinear) World is Scandalously Invalid

Uri Simonsohn
ESADE Business School
urisohn@gmail.com

Abstract

Hypotheses involving interactions, where one variable modifies the association between another two, are common in experimental and observational social science. Interactions are typically tested relying on models that assume effects are linear, e.g., with a regression like $y=\mathbf{a}+\mathbf{b}x+\mathbf{c}z+\mathbf{d}x\cdot z$. In the real world few effects are linear, and this invalidates inferences about interactions. For instance, in realistic situations, the false-positive rate for an interaction can be 100%, the interaction can have the wrong sign with certainty, and results from probing the interaction, computing the effect of interest for different levels of the moderator, be entirely disconnected from reality. This paper proposes a revised toolbox for studying interactions which is curvilinear-robust, giving correct answers 'even' when effects aren't linear. It's applicable to most study designs, while introducing minor modifications to current analytical and reporting practices. The presentation combines statistical intuition, examples of published results, and simulations.

Data and code to reproduce all results are available from:
https://researchbox.org/313 (use code **SNGGCU**)

Studying interactions, where a variable moderates the relationship between two other variables, is common in social science. For instance, inspecting the March 2020 issues of three prestigious empirical journals in psychology (JPSP, JEP:G, and Psychological Science), I found that 71% of papers reported statistical analyses examining possible interactions.

The general approach to studying interactions is, for our purposes, the same for the majority, perhaps almost all statistical frameworks commonly used by social scientists; whether they consist of linear regression, ANCOVA, probit/logit model, multilevel models, meta-regression, structural equation models, etc. The problems discussed in this article, and their solutions, are applicable to all of these forms of analyses; therefore, for ease of exposition, I focus on linear regression. Consider as an example studying how age and gender (female = 1,0) are associated with peoples' weight, through this regression:

$$\text{weight} = \mathbf{a} + \mathbf{b}\,\text{female} + \mathbf{c}\,\text{age} + \mathbf{d}\,\text{age}\cdot\text{female} + \varepsilon \qquad [1]$$

To *test* the interaction, that is, to assess whether the association of age and weight is different for men vs. women, one would usually examine whether the estimate $\hat{d}$ is significantly different from zero (or equivalently, assess whether its confidence/credibility interval excludes 0). To *probe* that interaction, to assess how big the gender difference is for a given age, one would carry out a procedure that gets different names in different fields (e.g., simple slopes, Johnson-Neyman procedure, conditional marginal effect, spotlight, pick-a-point, floodlight), and which involves combining $\hat{b}$ and $\hat{d}$. For instance, at age=10, the estimated marginal effect of being female on weight, from equation [1], is $\hat{b} + \hat{d}\cdot 10$.

For ease of exposition, I use causal language throughout this article, talking of the 'effect' of variables, but obviously this language is more warranted in some situations than in others. As shorthand I refer to the predictors in an interaction as *x* and *z,* and to the dependent variable as *y,* arriving at a generic version of equation 1:

$$y = \mathbf{a} + \mathbf{b}\,x + \mathbf{c}\,z + \mathbf{d}\,x\cdot z + \varepsilon \qquad [1']$$

The validity of interaction testing, and interaction probing, hinges on the arbitrarily made, and seldom verified assumption that the effects of *x*, *z,* and *x·z*, are all linear, that is, that *y* changes in a constant proportion to their changes (see footnote 1 for an additional implicit assumption).[1] There is a contradiction in researchers simultaneously collecting data to assess *whether x* & *z* are associated at all with *y*, while acting as if they knew for certain that any association must be linear.

Figure 1 illustrates how nonlinear relationships invalidate our interpretations of linear interactions; it depicts results for estimating equation [1] with actual data. Probing the interaction, we erroneously conclude that baby girls are (substantially) heavier than baby boys.

Figure 1. Nonlinear Effects Lead to Misleading Interaction Results
Panel A plots the reported median weight by age as reported by the Centers for Disease Control. Panel B depicts the vertical differences between the lines in panel A, the true weight difference between genders across age groups. Panel C shows misleading results obtained from estimates the true model from B using a linear regression with an interaction. The confidence band corresponds to the Johnson Neyman (1936) procedure.
R Code to reproduce figure: https://researchbox.org/313.2 (use code **SNSGCU**).

There are many reasons to suspect that the effects of variables studied by social scientists are generally not linear. Consider just these three reasons: (1) many dependent variables social scientists study depend to some extent on human perception of physical or numerical stimuli, and such perception exhibits diminishing rather than constant sensitivity (Stevens, 1957); (2) because of satiation, among other

[1] There is also the separate assumption that the interaction between x and *z* is well captured by their product. We are so used to this assumption that we think of "interaction" as synonymous with a product, but variables can interact in many ways that are poorly captured by a product, the interaction need not even be symmetric (e.g., x=4 and z=1 need not have the same associated effect as x=1 and z=4.). To be concrete, consider these functional forms involving interactions but not products: $y=x/z$, $y=x^z$, $y=x\cdot abs(z)$, y=x if z>4 but y= 2x if z<4, etc.

reasons, people's enjoyment of consumption and experiences exhibits diminishing rather than constant marginal benefit. And conversely, the marginal cost of effort is increasing rather than constant. And, (3) bounded scales are commonly used as dependent variables in social science; as the value of a predictor increases, inevitably at some point the magnitude of the marginal measured effect decreases as some observations reach the highest/lowest possible value on the scale (Loftus, 1978). In sum, psychophysics, diminishing marginal returns, and reliance on bounded scales, are three of the many powerful reasons to expect nonlinear relationships in social science.

Figure 2 provides some concrete examples of the kinds of nonlinear relationships we observe in real data. Panel A depicts perhaps the most intuitive nonlinearity, a smooth, monotonic, but concave function; older people are more conservative, but age at a decreasing rather than constant rate. Panel B shows an association where the weakening of the effect is more dramatic, almost entirely flattening out. I refer to such patterns as 'canopy effects', corresponding to a floor/ceiling that impacts the effect of a single, rather than of all, predictors. Canopy effects are also present in Panels C & D.

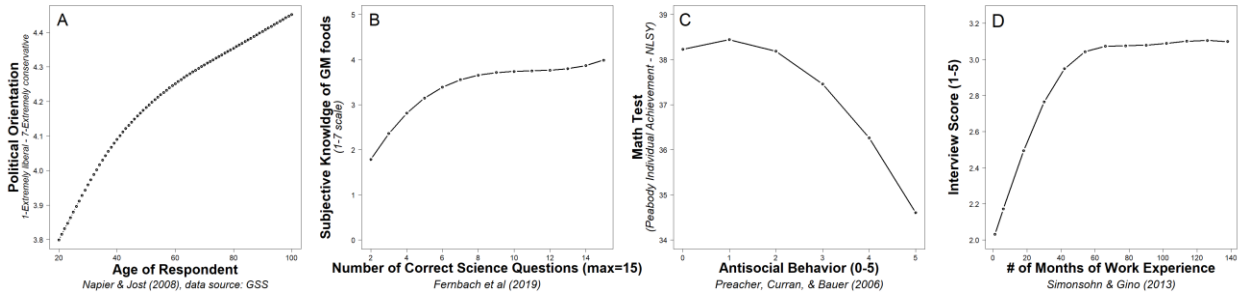

Figure 2. ***Examples of nonlinear associations in real data.***
In all panels, the lines are formed by fitting the data with flexible models (GAM for A & D, third degree polynomials for B & C). **A,** N=47729 survey responses from the General Social Survey. **B**, N=501 respondents to survey run by authors examining correlates of attitudes towards Genetically Modified (GM) foods. The science question were true/false statements such as "All plants and animals have DNA". **C,** N=956 survey respondents from the National Longitudinal Study of Youth; the paper does not indicate how antisocial behavior was measured. **D,** N=12427 interviews of applicants to an MBA program, the interviewer provides an overall assessment of the candidate on a 1-5 scale, work experience is obtained from the applicants' files.
R Code to reproduce figure https://researchbox.org/313.7 (use code **SNSGCU**).

In what follows, I begin reviewing prior methodological papers that have dealt with the problem of interest here, highlighting where the current work departs from existing work. I then explain three main ways in which linearity may be relaxed (dichotomizing the moderator, adding flexibility to the estimated direct effect of x and z, while keeping *x·z* linear, and estimating a model where also the interaction can have nonlinear effects). I then report results of the performance of these three alternative approaches when testing and probing interactions for experimental data (where the two predictors in the interaction are uncorrelated), followed by doing the same for nonexperimental data. Throughout this article, starting with the title, I use the term "invalid" to label shortcomings with statistical procedures. In statistics, the term 'validity', unlike terms like 'bias' and 'consistency', lacks an agreed upon definition. I use the term 'invalid' to refer to a procedure which has a sufficiently elevated false-positive rate (e.g., >20%), or sufficiently low power (e.g., 0%), or sufficiently large bias, that it would not seem advisable to *ever* rely on such procedure to make inferences from data. Validity is thus at least somewhat subjective.

Prior work on nonlinearities and interactions

Just a handful of books and peer-reviewed tutorials appear to account for the vast majority of references social scientists use to guide the testing and probing of interactions (Aiken & West, 1991; Brambor, Clark, & Golder, 2006; Cohen, Cohen, West, & Aiken, 2003; Preacher, Curran, & Bauer, 2006; Spiller, Fitzsimons, Lynch Jr, & McClelland, 2013). Aiken and West (1991) alone accumulate, as of September 2021, over 49000 Google citations, and Preacher et al., another 4900. These go-to references do not discuss how strong the linearity assumptions are (i.e., how at odds they are with what we should expect real world data to look like), nor how consequential the violation of such assumptions is. Possibly for this reason, few empirical papers consider the impact of nonlinearities on the interpretability of the interactions they report. While largely ignored by these tutorial pieces and most empirical work, some prior methodological articles have been concerned with the issues raised here.

Focusing on testing interactions, at least three papers (Cortina, 1993; Ganzach, 1997; Lubinski & Humphreys, 1990) have warned that if the effect of *x* on *y* is not linear, and *x* is correlated with the potential moderator, *z,* the estimate of the interaction is biased, and its false-positive rate is inflated (the intuition for, and evidence of this bias, is presented later in this paper). The authors of these earlier articles assume that all nonlinearities are essentially quadratic (see footnote 2 for documentation of this claim), and thus consider true models only of this form: [2]

$$y = \mathbf{a} + \mathbf{b}x + \mathbf{c}z + \mathbf{d}\,x \cdot z + \mathbf{e}x^2 + \mathbf{f}z^2 + \varepsilon \qquad [2]$$

Upon assuming a true model that is quadratic, these articles naturally propose researchers estimate quadratic regressions, i.e., including x^2 and z^2 as covariates. In relation to this work, I relax the assumption that all nonlinear relationships are (even approximately) quadratic, and relax the assumption that when *x* and *z* are correlated, *their* relationship is linear. I empirically evaluate how well the proposal of including x^2 and z^2 as covariates performs for testing the *x·z* interaction under nonlinearity, finding it performs surprisingly well, but not quite best. [3]

In terms of *probing* interactions, on estimating the effect of *x* on *y* for different values of *z,* the political science article by Hainmueller, Mummolo, and Xu (2019) warns against projecting linearly using estimated interaction coefficients. They propose, for robustness, estimating separately the slopes for the effect of *x* on *y*, for low, medium and high values of *z*, referring to this approach as the 'binning estimator'. In relation to their work, here I show that dichotomizing in general and the 'binning estimator' in particular

[2] The following quotes, with emphasis added, are used to document my claim that these paper equate nonlinearity with quadratic:
Lubinski and Humphreys (1990, p. 389): "Our interpretation of the positively accelerated trend corresponds to a similar curvilinear (**quadratic**) phenomenon observed within a variety of disparate behavioral domains"
Cortina (1993, p. 920): "Which nonlinear term or terms should be used? . . . psychological phenomena rarely display anything more complex than a **quadratic** trend"
Ganzach (1997, p. 236): "Note that a curvilinear relationship as defined above need not necessarily be quadratic. However, for the sake of simplicity, in the current article **I assume** that a true curvilinear relationship **is indeed quadratic**."

[3] Matuschek and Kliegl (2018) build on this work and propose using a two-stage estimation. In the first stage, one estimates a GAM (general additive model) of *y* on smooths of *x* and *z* separately, and on the second stage one estimates a linear regression of the residual of the first, on the *x·z* interaction. In supplement 1 I show this proposed solution does not perform well enough to merit consideration by behavioral scientists. For example, in scenarios where the tools proposed here obtain 80% power, the solution by Matuschek and Kliegl can suffer from 0% power.

is only curvilinear-robust if *x* and *z* are uncorrelated (e.g., in an experiment). When predictors are correlated (as they are outside experiments in general and in the most relevant example by Hainmueller et al in particular), dichotomization can lead to severely biased results. As biased, in fact, as the linear model where *z* is not dichotomized. In relation to this work I propose two alternatives for probing interactions which are curvilinear-robust for both experimental and observational data. [4]

Three approaches to relaxing linearity assumptions

In this section I overview three main approaches for relaxing the assumed linearity of effects: (1) discretizing the moderator, (2) adding flexibility to the functional form for the *independent* effects of *x* & *z,* and (3) estimating a fully flexible model where also the interaction has a flexible functional form.

Approach 1. Discretizing z, the moderator. Behavioral scientists have long retorted to testing interactions by dichotomizing moderators into high vs low values (e.g., median splits), and then contrasting the association between *x* and *y* for high vs low values of *z.* While this practice is popular (see the numerous examples cited by Iacobucci, Posavac, Kardes, Schneider, & Popovich, 2015), its criticisms are popular as well (see e.g., Cohen, 1983; DeCoster, Iselin, & Gallucci, 2009; Humphreys & Fleishman, 1974; Lubinski & Humphreys, 1990; Maxwell & Delaney, 1993; McClelland, Lynch, Irwin, Spiller, & Fitzsimons, 2015).

The traditional justification for discretization is "analytical ease and communication clarity" (Iacobucci et al., 2015, p. 652). The traditional argument against discretization is the expected reduction in statistical power; when there truly is an interaction between *x* and a continuous predictor *z,* dichotomizing *z* to test the interaction reduces the probability of obtaining a $p < .05$ for the $x \cdot z$ interaction term (Cohen, 1983; McClelland et al., 2015). There is a secondary benefit to discretization, however, beyond simplification of analyses, which appears to have gone largely unmentioned in the decades long

[4] Hainmueller et al discuss, in addition to the 'binning estimator', a kernel based estimation procedure. It is equally invalidated by correlated predictors (see Figure 11 in this paper).

debate: when we dichotomize a moderator, we no longer need to impose (arbitrary) linearity to *probe* the interaction.[5,6] In light of this benefit, I examine the performance of dichotomization for *probing* interactions. As a preview: it outperforms the linear model, but it is not the best solution, and it is only valid for experimental data, where the two predictors in $x \cdot z$ are expected to be un uncorrelated.

Approach 2. Adding flexibility to functional form for the independent effects of x & z

As mentioned earlier, a few methodological articles have proposed handling the impact of nonlinearities of the effect of x and z, on the estimated effect of $x \cdot z$, by simply adding x^2 and z^2 to the regression (Cortina, 1993; Ganzach, 1997; Lubinski & Humphreys, 1990). That proposal follows from the authors' premise that any nonlinearity in *x* and *z* can be adequately captured by said quadratic terms (see footnote 1). I considered two alternatives for adding flexibility to the effects of *x* and *z.* First, estimating an *interrupted* quadratic regression, where there is one quadratic relationship estimated for the effect of x on y for low and another for high values of x, and analogously for z (this reduces the impact of specification error in one portion of the data onto the estimates in another remote portion of the data). This approach performs very well. I also explore giving fuller flexibility to the functional form of the independent effects of x and z by estimating a general additive model, GAM (Hastie & Tibshirani, 1987; Wood, 2006), where the independent effects are flexibly estimated (not assuming linearity), but the interaction remains a linear effect. This combination is done seeking interpretable results, as it is not clear how to interpret the statistical significance of a fully flexible interaction smooth. This mixed-GAM solution

[5] Some articles do consider nonlinearities when discussing median splits, but not in the context of providing curvilinear robust answers to Question 3. Specifically, Rucker, McShane, and Preacher (2015, p. 674) point out that nonlinear relationships may not be studied via median splits, writing "When the relationship is nonlinear, we note that dichotomization via the median split procedure or otherwise is simply not appropriate as dichotomization necessarily conceals nonlinear relationships" and DeCoster et al. (2009, p. 359) focus on answering Question 1 &2, writing "Dichotomization does not produce an advantage when trying to *detect nonlinear relations* because any nonlinear relation that can be detected as a difference between dichotomized group means can also be detected by linear regression" (emphasis added).

[6] The aforementioned article by Hainmueller et al. (2019), which proposes the binning estimator, belongs to a separate literature (e.g., it does not cite any articles discussing discretization in the context of studying interactions).

(mixing flexible x & z with linear $x \cdot z$) performs, surprisingly, worse than either quadratic approach, and nearly as poorly as the fully linear model, perhaps due to an error in how the procedure computes standard errors in these kinds of specifications (bootstrapped standard errors lead to much better results).

Approach 3. Estimating a flexibly nonlinear model

We can further relax functional form assumptions by allowing also the interaction term to have a nonlinear effect. A fully flexible model. I rely on GAM for this as well. As a preview, this approach makes the *probing* of interactions particularly robust and just as easy to interpret as linear probing. I thus propose conducting what I am calling GAM simple slopes and GAM Johnson Neyman procedures, which are analogous to the current practices of simple slopes and Johnson Neyman procedure, but relying on a GAM rather than a linear model.

In the sections that follow I demonstrate with examples, and evaluate with simulations, how these three approaches to relaxing linearity assumptions perform for testing and probing. First with data from experiments (where r(x,z)=0), and then with data from non-experiments (where r(x,z)≠0). Table 1 previews the recommendations I arrive at based on the forthcoming results.

Table 1. Proposed Toolbox for Curvilinear-Robust Analysis of Interactions

	Testing interactions *(does z modify the effect of x on y?)*	**Probing interactions** *(what is the effect of x for a given value of z?)*
Case 1. Experiments (randomized x or z) *r(x,z)=0*	- Linear model is OK	- Discretize z, report effect of x for high vs low z - GAM Johnson Neyman
Case 2. Non-experiments (x & z measured) *r(x,z)≠0*	- Good: Quadratic regression (control for x^2 and z^2) - Better: If feasible, *interrupted* quadratic regression	- GAM Simple Slopes - GAM Johnson Neyman

Case 1: Experiments (the predictors in the interaction, x & *z,* are not correlated)

Testing interactions with experimental data

It is common in behavioral science experiments to evaluate the extent to which the effect of a manipulation that is randomly assigned to participants, *x*, on the dependent variable, *y*, is moderated by

a third variable, *z,* which may be observed (e.g., age, experience, body-weight), or manipulated (e.g., framing of the information, incentive levels, group size). Because *x* is randomly assigned, we expect r(*x*,*z*)=0 whether *z* is also randomly assigned or not. The lack of an expected association between x and *z* substantially reduces the consequences of nonlinearities on the validity of the linear model. Indeed, the linear model provides valid answers for testing interactions, and only the probing of the interaction is invalidated by nonlinearities.[7]

The intuition for why the *testing* of interactions from experimental data, with a linear regression model, is curvilinear-robust, is that while regression results are often presented as the estimates of a true linear relationship of interest, they can also be conceptualized as the estimate of the *average* slope of a nonlinear relationship (see e.g., Gelman & Park, 2008). That is to say, if the effect of *x* on *y* is not linear, where, say, $y_i=a+\mathbf{b}_i x_i$, and $\mathbf{b}_i$ possibly varies across observations indexed by *i*, then $\hat{b}$, the linear regression coefficient estimate, is the estimated *average* $\mathbf{b}_i$.

Nonlinearities, when seen from this perspective, produce regression estimates which are analogous to sample averages. There is nothing intrinsically incorrect about computing a single value (e.g., average income) to summarize a set of different values in the data (e.g., the different incomes received by different people); analogously, there is nothing intrinsically incorrect about computing the average slope to summarize a set of different slopes. "I lost 2 pounds per week during my diet" is a reasonable summary, even if during the first week one lost 3 pounds and during the last week only 1; that average-slope summary, 2 pounds per week, is the one provided by a linear regression.

Let's look at a concrete example I have used before, where the nonlinear relationship is $y = x^2$ and we estimate a linear regression $y=\mathbf{a}+\mathbf{b}x$ (Simonsohn, 2018b, p. 540): "Say . . . the data consist of three

[7] When the dependent variable is a proxy for a theoretical construct (something that is very common in social science, e.g., using performance in a math test as a proxy for understanding of mathematical concepts), an interesting conceptual challenge remains, which is that an observed interaction may reflect changes in how the proxy variable maps onto the latent one, rather than an effect of on the latent variable itself (Loftus, 1978; Wagenmakers, Krypotos, Criss, & Iverson, 2012) . See also the "conceptual" vs "mechanical" interaction distinction I have made (Simonsohn, 2018a)

observations, *x* = 1, 2, 3 and thus *y* =1, 4, 9. The slope between the first two points is (4 – 1)/(2 – 1) = 3, the slope between the last two points is (9 – 4)/(3 – 2) = 5, and the slope between the first and last points is (9 – 1)/(3 – 1) = 4. So, the average slope is (3 + 5 + 4)/3 = 4, and a linear regression will recover $\hat{b}$ = 4". Note that the average slope obtained from a regression gives different weights to each slope, where the weights depends on the distance between datapoints (Gelman & Park, 2008), in this stylized example the data points happen to be equidistant, simplifying the intuition.[8]

When *probing* an interaction, however, when estimating the effect of *x* on *y* for specific values of *z,* the answer provided by the linear model is no longer analogous to averaging; instead, it is analogous to fallaciously assuming that all observations are equal to the average observation. It is analogous to saying "if on average you lost 2 pounds per week during your diet, then between the 6th and 8th weeks, you lost 4 pounds".

Figure 1 showed an example of just this problem, a probed interaction between gender and age leading to incorrectly, but with statistical confidence, inferring that baby girls are substantially heavier than baby boys. The regression interaction meaningfully estimates the average difference on the effect of age on weight by gender, but that interaction estimate is *not* a valid value to use for establishing the gender difference *at specific ages*.

Having hopefully provided the intuition for the problem, I now turn to simulations to compare the performance of alternatives approaches that social scientists may switch to, to probe interactions in experiments, in a curvilinear-robust fashion. To present results that are easy to interpret, I rely on a hypothesis testing framework, leading to simulation results that are easy to evaluate as "good" (e.g., a

[8] Specifically, the weight given to each slope is the squared distance between predictor values. For example, if the x values were x=1,2,10, and thus the y-values were y=1,4,100, while the simple average slope is 8.6, the regression coefficient, the *weighted* average slope, is $\hat{b}$=11.38; giving (much) more weight to the slope between x=1 & x=100 (weight=81) than between x=1 & x=2 (weight=1).

25% increase of power) and "bad" (e.g., a 19% false positive rate). I start with false-positive rates, and then discuss power.

Probing Interactions with Experimental Data

False-Positive Rates.

To examine how nonlinearities impact the probing of interactions in data from experiments, I focus on attenuating interactions: instances where the effect of a randomly assigned treatment, *x,* on the dependent variable, *y,* is reduced but never reversed by a moderator variable, *z.* For example, the true model $y=x/z$, with z>0, meets this description, bigger z values reduce the effect of x, but never reverse it.

I focus on attenuating interactions for two reasons. First, they are commonly observed and commonly hypothesized in social science. Second, analyzing attenuating interactions through linear models *inevitably,* but incorrectly, predicts an eventual reversal of the effect (the reversal, however, could be estimated to occur for an impossible value of z). Attenuating interactions, therefore, constitute a frequent situation where social science's current toolbox for studying interactions is likely to be invalid. I set out to identify alternatives that perform sufficiently better.

Simulated true models. In the simulations, I consider variations of two main true models that give rise to attenuating interactions. The first consists of a linear model, but with a ceiling effect. Specifically, I consider the true model $y=x+z$, the dependent variable is simply the sum of the two predictors, but there is a ceiling at some value y_k, so that $y=min(x+z , y_k)$. The ceiling causes the effect of *x* on *y* to get smaller for higher values of *z*, eventually becoming 0. But the effect of *x* is never negative. Second, I consider a similar association between *x*, *z* and *y*, through a more realistic (nonlinear) function: $y = -e^{-(x+z)}$ (to visualize this function see the scatterplot inside the right panel of Figure 3).

To avoid stumbling on simulation results that hinge on arbitrary operationalizations, I consider 36 variations of each of these two models. The variations involve different sample sizes, distributions of the

z values (e.g., normal vs skewed distributions), and precise functional form (e.g., how quickly the effect of *x* approaches zero). For details on these variations see caption for Figure 3.

Probing the interaction. After generating the data, in each simulation I run three alternative analyses that probe the *x·z* interaction. First, using the estimated coefficients from the linear regression, $\hat{y} = \hat{a} + \hat{b}\,x + \hat{c}\,z + \hat{d}\,x \cdot z$, to compute the effect of the *x* for different z-values (Johnson & Neyman, 1936), known as the Johnson-Neyman procedure, also as 'floodlight' (Spiller et al., 2013) and is sometimes also referred to as 'finding the regions of significance'. The false-positive rate is computed based on the marginal effect of *x,* when *z* is 2 SD above its mean, computing what is known as a simple slope (also known as spotlight and pick-a-point analysis). The result is coded as false-positive if the marginal effect of *x* is estimated as *negative*, $p < .05$.

The second analysis dichotomizes *z* into high vs low values, and estimates the average effect of *x* on *y* for high values of *z* (e.g., those above the median). This approach requires setting a cutoff demarcating high vs low *z* values. To avoid concerns that the results I obtain are specific to a particular cutoff that happens to match the assumed true functional form, I examine the performance of two alternative cutoffs: (1) the traditional median split, where half the *z* values are high and half low, and (2) tertile splits, where the top 1/3 of observations are high and the bottom 1/3 low (tertile splits are proposed by Gelman & Park (2008) for regression summaries in general, and by Hainmueller et al. (2019) for interactions in particular; their 'binning estimator'). The result of these dichotomized analyzes are coded as false-positive if the estimated average effect of *x* on *y* is negative, $p<.05$, for high z-values.

The third alternative analysis consists of estimating a general additive model (GAM), allowing a flexible functional form for the effect of the continuous predictor, z, on y. This function, moreover, is estimated separately for x=1 and x=0. Subtracting these two estimated functions, one can compute the effect of *x* for every value of *z*. This effectively generalizes the Johnson and Neyman (1936) procedure from linear to a flexible GAM model. I thus refer to it as the GAM Johnson Neyman procedure. The results

are coded as false-positive if the estimated effect of *x* on *y* is negative, $p < .05$, for *z* values 2 SD above the mean. This specific comparison at a fixed point is analogous to traditional simple slopes (Aiken & West, 1991), so I refer to it as GAM simple slopes.

Figure 3 reports false-positive rates for each of the procedures across 72 scenarios. Valid models should have false-positive rates near the nominal 5%. The poor performance of the linear model is striking. For many scenarios, the approach that is the current gold standard for much of social science for probing interactions, achieves a 100% false-positive rate; it *always* arrives at statistically significant evidence of something that is not true.

The two alternative approaches, in contrast, are slightly conservative for most scenarios and close to the 5% nominal rate even in the most extreme ones (note that we are testing a directional hypothesis with a two-sided test, and that often the true effect is positive rather than zero, so the false-positive rate for a perfectly calibrated test would be ≤2.5%).

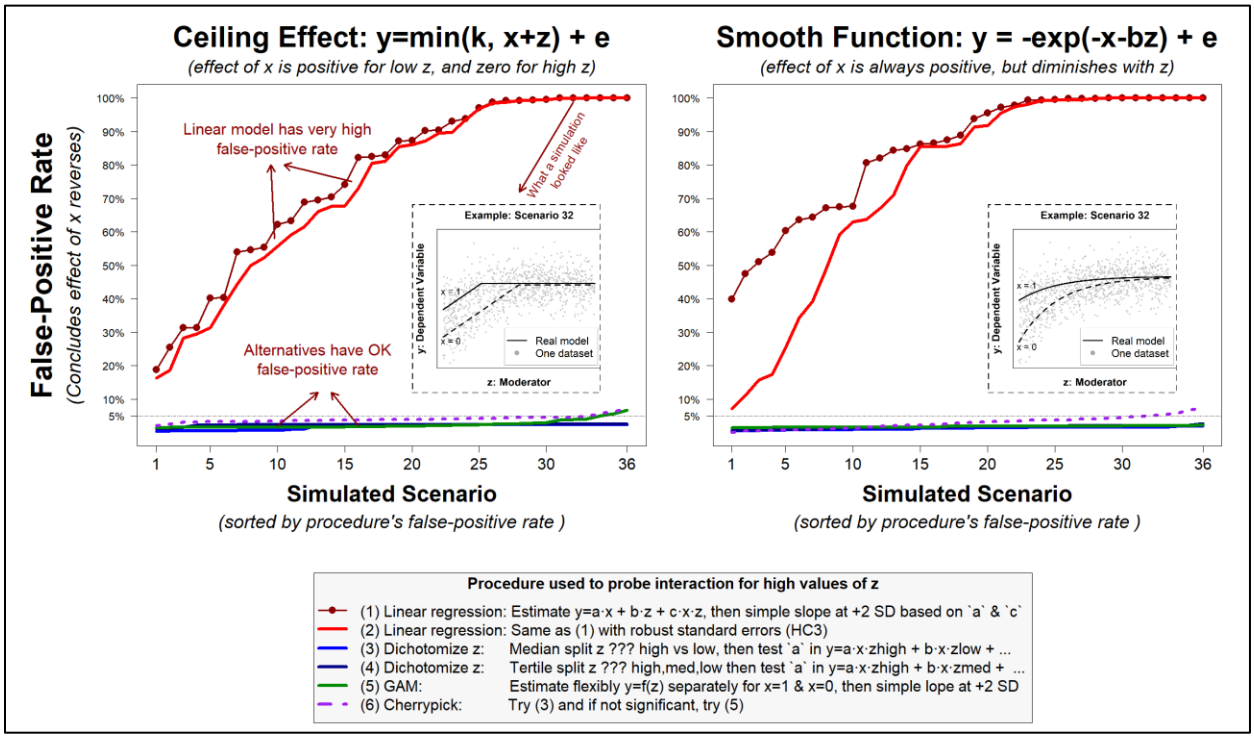

Figure 3. False-Positive rates for effect of *x* on y being negative for high *z*; true effect is never negative.
The y-axes depict the percentage of 5000 simulations, run for each scenario, where the probing of an interaction between x and *z* lead to an estimated negative effect of x on y, with $p < .05$, for a high *z* value, despite the true effect never being negative. In all simulations x is binary, and scenarios are generated by varying the distribution of *z* (standard normal, left-skewed, right-skewed, uniform; the latter three range between z > -2 and z < 2), and sample size per condition n=100, 200, or 500. Scenarios in the left chart vary the ceiling for the effect of x and *z* on y to -.5, 0, or +.5 (a ceiling effect prior to adding random noise), while the scenarios on the right vary the coefficient of b to be 1, 2, 3. These operationalizations lead to the 4x3x3=36 scenarios in each panel. The plotted false-positive rates are adjusted for simulation error, see footnote.[9]
R Code to reproduce figure: http://researchbox.org/313.39 (use code **SNSGCU**).

To help further understand why the GAM Johnson Neyman procedure outperforms the traditional (linear) Johnson Neyman procedure, Figure 4 depicts results for one of 5000 simulations, for one of the 72 scenarios depicted in Figure 3.

[9] Simulations *estimate* false-positive rates, and do so with error. If one were to re-run the same simulation again, a slightly different value would be obtained, what's called simulation error. The larger the number of simulations, the smaller simulation error gets. The figure is based on 500 simulations fore each of the 36 scenarios. Even if the true false-positive rate for all scenarios were exactly 5%, the smallest of the 36 is not expected to be 5%, but actually, only 3% (in R, one can computer this with:: qbinom(p=.5/36, size=500,prob=.05)/500). Similarly, the largest of 36 is not expected to be 5% either, bur rather, 7.2% qbinom(p=35.5/36, size=500,prob=.05)/500). To avoid simulation error from distorting the perceived efficacy of the procedures, the expected deviation from 5% is added to the obtained false-positive rates, making the reported false-positive. So to the smallest post-hoc estimated false-positive rate I added 2%, and to the largest I subtracted 2.2% (as long as it was not at the ceiling of 100%).

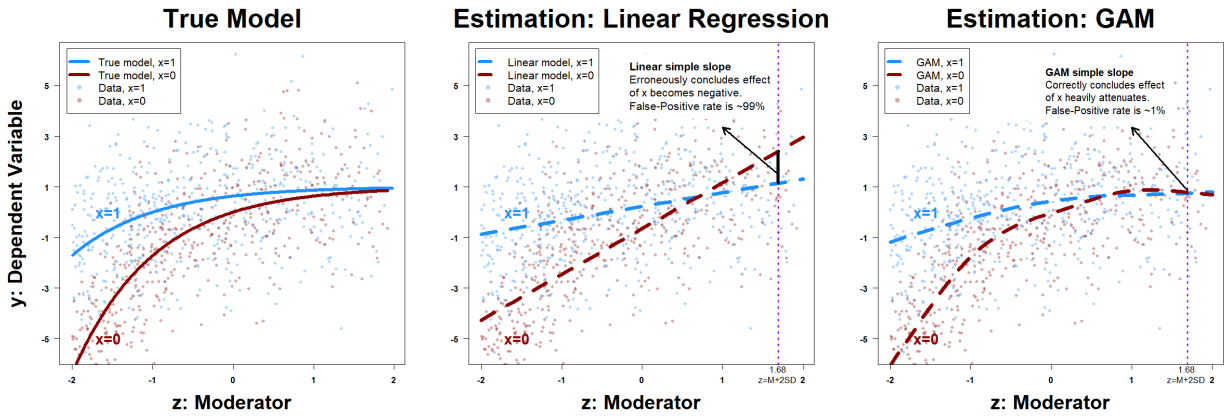

Figure 4. Example of simulated scenario in Figure 3, with high false-positive rate for linear model

The three panels depict the same simulated dataset, encompassing 500 observations per condition (x=1 vs x=0), where the true functional form, $y=1-e^{-(x+b\cdot z)}$. Due to specification error, a linear regression that includes an interaction leads to an elevated false-positive rate (99%) for detecting a sign-reversal of the effect of *x* for high values of *z*.

R Code to reproduce figure: http://researchbox.org/313.40 (use code **SNSGCU**).

Statistical power for probing interactions in experiment.

The false-positive rates of the linear model, when probing interactions, seem sufficiently high to justify abandoning this approach, even if it provided higher statistical power than its alternatives. But, interestingly, for *probing* interactions, the linear model can easily have *lower* statistical power than the median split does. To understand this, we do not need new simulations. Look back at the right panel of Figure 3. For high values of *z*, recall, the true effect of *x* on *y* is positive, but the linear model estimates it as negative, often with false-positive rates of 100%. That means the model has a 0% chance of detecting the true effect, which is positive. So it has 0% power.[10]

In other words, while median splits have been justifiably criticized for decades for having lower power than the linear model to *detect* an interaction (Cohen, 1983), they can have greater power for

[10] An odd aspect how of statistical power is often defined, is that it adds the probability of obtaining a statistically significant of the correct and incorrect sign (e.g., it includes the probability of a statistically significant negative estimate, when the true value is positive). A purist may thus be troubled by my claim that power is low when the true effect is positive and the estimated effect is significant and negative. To appease such concerns, consider that in many cases the specification error will be insufficient to bias a probed interaction all the way to a statistically significant effect of the wrong sign, and will merely underestimate the magnitude of the effect with the correct sign, and in those cases, power will be reduced even in its more literal definition, potentially all the way down to 0% power.

probing those interactions. *So it is defensible, and sensible, to estimate a linear model to test for the interaction (*$y=\mathbf{a}+\mathbf{b}x+\mathbf{c}z+\mathbf{d}x\cdot z$*), and then estimate a median split to probe the interaction.* Nevertheless, it is generally the case that the median split will have lower power for probing interactions than the GAM Johnson Neyman procedure will.

It is interesting to consider, as a boundary case, the unlikely scenario where the true model is actually linear. How much less power does the GAM Johnson Neyman procedure have compared to the traditional (linear) Johnson Neyman procedure? In simulations reported in supplement 2, I find small differences. Across 8 scenarios calibrated to give the linear probing 50% power, GAM Johnson Neyman obtained essentially identical power (about a 1.5% drop), and in those same scenarios calibrated for 80% power, GAM obtained about 4% less power (76% vs 80%).

In sum, switching from the current standard of probing interactions in experiments with a (linear) Johnson Neyman, to instead relying on the GAM Johnson Neyman procedure, would (1) eliminate the unacceptably high false-positive rates when the effects are not linear, (2) substantially increase power in many situations, and (3) be *un*likely to meaningfully reduce power.

Case 2: Curvilinear-robust interactions for non-experiments (where r(x,z)≠0)

Testing interactions in non-experimental data

When the two predictors that go into an interaction, *x* and *z,* are correlated, nonlinearities in the effect of *x* or *z* on *y,* impact the linear model's ability not only for probing interactions (as is the case with experimental data) but also to test them. In other words, estimating the regression $y=\mathbf{a}+\mathbf{b}x+\mathbf{c}z+\mathbf{d}x\cdot z$ when the effect of *x* or *z* is nonlinear, and *x* and *z* are correlated, produces bias, elevated false-positive rates, and possibly lower power, for the estimate of **d** (Cortina, 1993; Ganzach, 1997; Lubinski & Humphreys, 1990).

Examples with a binary moderator.

To build an intuition for why bias arises in the presence of correlated nonlinear predictors, I begin considering a situation where the moderator, *z,* is binary, allowing illustrating with simple figures the source and nature of the problem. Let's say we compare men (*z*=1) and women (*z*=0) in situations where gender, *z,* is correlated with the predictor of interest, x. That is, men and women have different *x* values.

The top row in Figure 5 below depicts a stylized example, which facilitates understanding the nature of the problem, the bottom row of Figure 5 does so with data used for a published article (from Simonsohn & Gino, 2013).

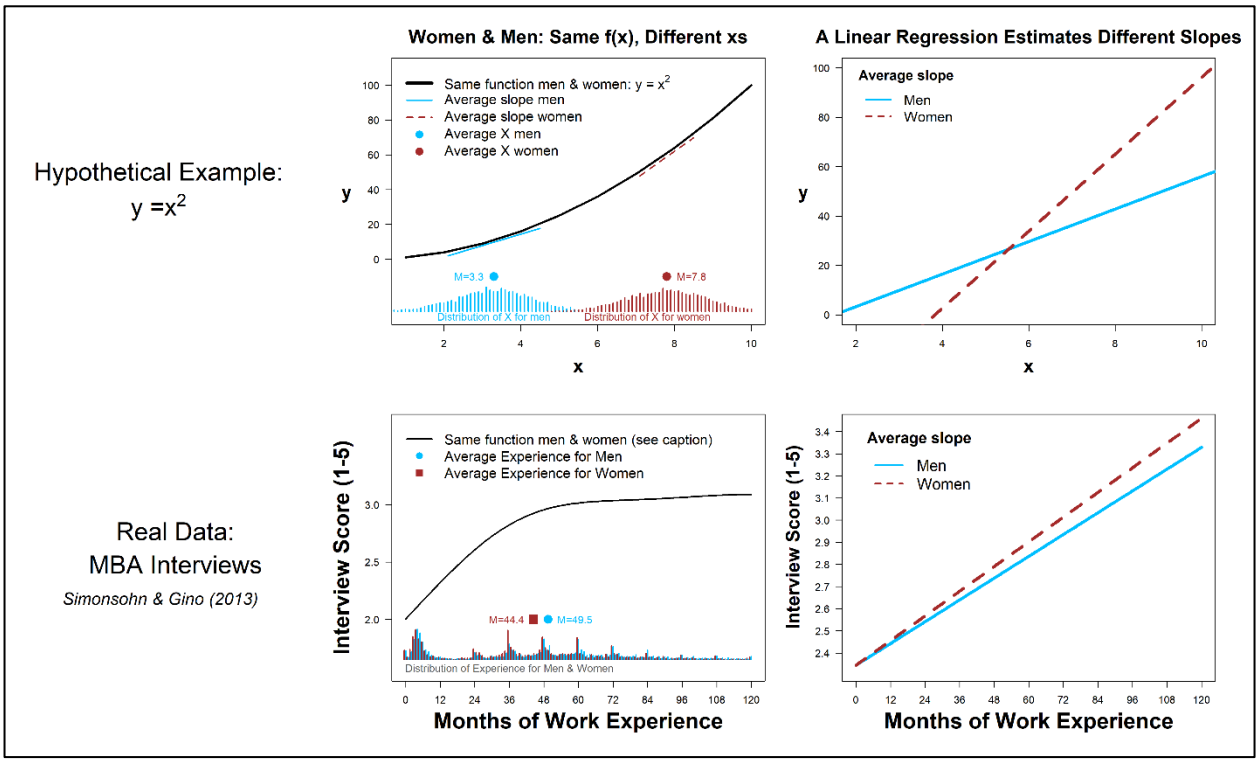

Figure 5. Examples where correlated nonlinear predictors invalidate interaction
The panels on the left depict a true nonlinear relationship that is identical for men and women (no gender interaction), and the panels on the right depict the results from a linear regression which, due to specification error, generate (statistically significant) interactions implying different slopes for men and women.
R Code to reproduce figure: https://researchbox.org/313.31 (use code **SNSGCU**).

Starting with the top row in Figure 5. The relationship considered is $y=x^2$. Because *y* does not depend on *z* in any way, we *know* there is no true interaction. Nevertheless, as the figure shows, *x* and *z* are highly correlated (*r* = .91), as 'women' have systematically higher values of *x* than do 'men' (in this

extreme case, virtually non-overlapping distributions). The problem with estimating the linear regression that includes an interaction term $x \cdot z$, is that it will leverage this association between x and z to improve overall fit, at the expense of a spurious interaction. Specifically, when x is higher, on average $x \cdot z$ is higher too, and when x (and thus x·z) is higher, the effect of x on y is larger (since the slope, dy/dx, is 2x), so the point estimate for the interaction $x \cdot z$ will be biased upwards, to capture some of that slope change in the effect of x. We thus spuriously conclude that women have a steeper function for the effect of x on y than do men. Even though for both men and women $y=x^2$, same f(x), our estimated *linear* function is different for men and women.

This example is extreme. But we do not need an extreme example for the problem to arise. We don't need a very high correlation between x and z, nor non-overlapping distributions of z for spurious interactions to arise. All we need is (1) a non-linear association of x with the dependent variable, and (2) a correlation between x and z. This problem of correlated nonlinear predictors, seems likely to apply to many, perhaps the vast majority of datasets where neither x nor z were randomly assigned.

Let's now consider one such real-world dataset. The bottom row of Figure 5, which uses the same dataset from Figure 2D. The dependent variable is the 1-5 interview scores received by MBA applicants, and the predictors are the applicant's gender, z, and how many months of work experience they have, x. Here the distributions of x for male and female applicants have substantial overlap, and the correlation of gender and experience is miniscule, r = -.06 (but $p < .001$). But it is not exactly zero, and it is visually obvious that the effect of experience on interview score is nonlinear, which means that a nonlinear predictor, x, is correlated with the moderator, and thus we expect a biased interaction term. Indeed, the false-positive rate for the interaction is 63%, rather than the nominal 5% (see footnote for calculation details).[11]

[11] To generate data under the null I estimate a flexible model predicting interview score with months of experience. In R, mgcv::gam(score~s(experience)). The predicted values from this model is treated as the true model. I generate simulated samples

In some sense the "false-positive" interaction results *are* correct; men do have a flatter average slope than do women, but this difference, which in an empirical paper would typically be attributed to men and women having a different f(*x*), that is, with men and women having their work experience rewarded differently, should instead be attributed to men and women having different *x* values, different experience levels. It is analogous to Simpson's paradox. For every single *x*-value men and women have the same f'(*x*), but, men have more observations with smaller f', thus their *average* f' is smaller. This problem of correlated nonlinear predictors also invalidates a calculation that in economics is known as the Blinder-Oaxaca decomposition. [12]

Examples with a continuous moderator.

Having provided the intuition with stylized examples and dichotomous predictors, I now move on to examples from published research which are neither stylized nor comprised of dichotomous predictors, but where nevertheless interactions are invalid because they combine correlated nonlinear predictors. Strikingly, both examples come from methodological articles providing tutorials on the interpretation of interactions in linear regression. The authors of those tutorials did not realize, however, or at least did not mention, that the linear models they used for teaching about linear models were invalid.

The first example comes from the tutorial by Preacher et al. (2006), probably the most cited peer-reviewed methodological article giving researchers guidance on how to probe regression interactions. I re-analyze here the only example in that paper (see their section "An Example", p.444-446). The dataset,

by shuffling the residuals and summing them to the predicted values. Each bootstrapped dataset is then analyzed with a linear regression that includes a gender*experience interaction. In 63% of simulations the interaction term was p<.05, even though by construction there is no interaction. R Code to reproduce simulations: http://researchbox.org/xxxx

[12] The Blinder-Oaxaca decomposition in economics (Blinder, 1973; Jann, 2008; Oaxaca, 1973), seeks to decompose an observed difference in outcomes across groups (e.g., salaries for men and women) to differences in x-values (e.g., different education levels for men vs women) and differences in functions (e.g., different returns to education for men vs women). The Blinder-Oaxaca decomposition also assumes linearity and is also invalidated in the presence of nonlinearities. Specifically, differences in outcomes could be incorrectly attributed to groups having different linear functions, when in fact they have identical nonlinear functions.

from the National Longitudinal Survey of Youth (NLSY), involves N=956 children as the unit of analysis, performance on a math test as the dependent variable, and measures of children's antisocial tendencies, *x*, and hyperactivity, *z,* as the key predictors. The example is used to illustrate how to test and interpret an $x \cdot z$ interaction in a linear regression model. Preacher et al. do not discuss whether *x* and *z* have nonlinear effects on *y*, or whether *x* & *z* are correlated. I obtained the same dataset and started by successfully reproducing the regression results reported in their Table 1. I obtained the same point estimate and *p*-value for the antisocial*hyperactive interaction (b = -.3977, $p = .0055$).[13]

Figure 6 below reports results exploring the issue of correlated nonlinear predictors. At least one of the predictors, *hyperactivity,* has as an apparent nonlinear relationship with the dependent variable, and that the two predictors in the $x \cdot z$ interaction are highly correlated (r=.504). For reasons outlined above, we expect this combination to produce elevated false-positive rates for the interaction, and the fourth panel shows that this is indeed the case. Robust standard errors do not meaningfully reduce the elevated false-positive rate, but the three alternatives for testing interactions examined in this article, do obtain acceptable results. [14]

[13] I received the dataset from Kristopher Preacher via email on July 8th, 2016. I had requested it when working on a different, ultimately abandoned, project.

[14] The calculations are analogous to those reported earlier in footnote 11. I estimated a model predicting math scores with antisociality and hyperactivity as factors (i.e., with fixed effects for each possible values they take) and all the predictors used in the original regression as covariates (see their Table 1, the covariates are: age, grade, female (1/0), and minority status (1/0)). I use the predicted values from that regression as the true model in my simulations. It is a null model because the dependent variable is known to not depend on the interaction between the two predictors. To generate random samples from this true model I randomly shuffle the residuals from that specification across observations, generating bootstrapped datasets. For each bootstrapped dataset I compute the specification from Preacher et al (2006), keeping track of the share of simulations in which the *p*-value for the interaction is $p < .05$.

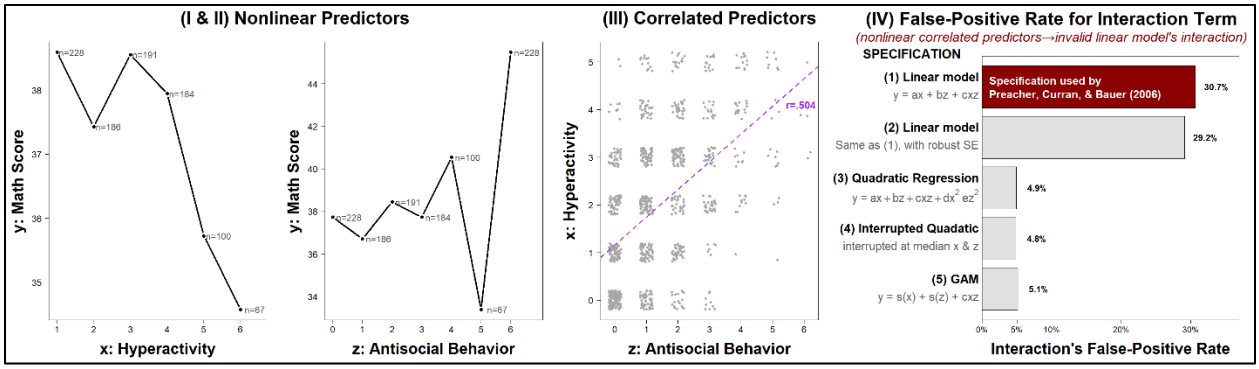

Figure 6. Correlated nonlinear predictors in Preacher et al. (2006) invalidate the interaction results.
The tutorial by Preacher et al (2006) on how to probe interactions includes this single example with real data (originally from the NLSY, with children, N=956, as the unit of observations), their analyses neglect the problem of correlated nonlinear predictors, which invalidates the interaction they probe. Panels I & II show mean values of the dependent variable for each possible value of the predictors. Panel III a scatterplot and best fitting regression line for the predictors. Panel IV shows the false-positive rates for testing the interaction of interest to Preacher et al. See footnote 14 for calculation details.
R Code to reproduce figure: https://researchbox.org/313.40 (use code **SNSGCU**).

The second example comes from Hainmueller et al. (2019), who provide a tutorial for political scientists on how to probe interactions in a way that is robust to nonlinear effects of the interaction itself. Hainmueller et al, however, do not take into account in their analyses, or recommendations, the impact of nonlinearities in the *independent* effects (i.e., nonlinearities in the effects of x and z, rather than in the effect of $x \cdot z$).

The relevant example for us, comes from their section "Case 4: Nonlinearity" (p.181-182), where they re-analyze data first examined by Clark and Golder (2006). In that dataset, elections are the unit of analysis, the number of political parties competing are the dependent variable, *y*, and the two focal predictors are the number of presidential candidates running in the election, *z* (where 0 presidential candidates is a common value), and temporal proximity of the election, *x*. Hainmueller et al. (2019) posted their code and data, enabling the reanalysis presented here. [15]

Figure 7 below shows that also in this dataset, correlated nonlinear predictors invalidate the interaction from the linear model. Its false-positive rate is maximal: 100%. In other words, if there were no interaction, a linear regression like the one estimated in the paper would, with certainty, conclude

[15] Hainmueller et al. (2019)'s code and data are available from https://doi.org/10.7910/DVN/Q1V0OG

erroneously that there is an interaction. The fourth panel shows that in this particular dataset, the simple solution of adding quadratic controls, x^2 and z^2, is not close to sufficient, and relying on a mixed GAM (with flexible effects of x and z but a linear interaction), also has an elevated false-positive rate (more on the limitation of this GAM based solution in the next subsection). Only the interrupted quadratic obtains the nominal false-positive rate. For calculation details see figure caption.

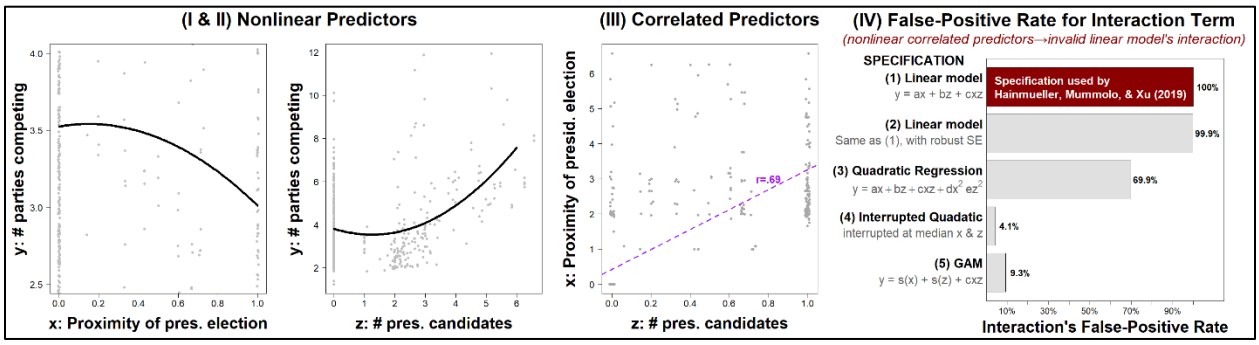

Figure 7. Correlated nonlinear predictors in Hainmueller et al. (2019) invalidate the interaction estimate
The dataset was originally analyzed by Clark and Golder (2006) and re-analyzed by Hainmueller et al. (2019), for the purpose of demonstrating how to *probe* an interaction nonlinearly. Hainmueller et al., however, did not take into account the problem of *correlated* nonlinear predictors. The unit of observation is an election taking place between 1947 and 2000 within a large set of countries (N=487). Panels I & II show associations between each predictor in the *x·z* interaction individually, and the dependent variable (with quadratic fit). For ease of exposition the y-axis is truncated in the first panel. Panel III shows a scatterplot, and best fitting regression line, for the association between predictors. Panel IV shows the false-positive rates for testing the interaction of interest to Hainmueller et al. For details on computations of false-positive rates see footnote.[16]
R Code to reproduce figure: https://researchbox.org/313.33 (use code **SNSGCU**).

Having provided an intuition for why correlated nonlinear predictors invalidate interaction estimates in linear regression, and having provided examples of undiagnosed instances of this problem in datasets used by influential tutorials on how to interpret interactions, in the next section I systematically evaluate, through simulations, the alternative approaches for curvilinear-robust testing of interactions.

Simulations: False-Positive Rates Testing Interactions of Correlated Nonlinear Predictors

[16] To compute the false-positive rate, I first created a null model where there is no interaction. In this case that corresponded to a GAM run on the posted data, but including only the independent effects of x and z as predictors, and not their interaction (i.e., *gam(y~s(x)+s(z)+covariates).* With the estimated GAM I computed predicted values, and treated it as the true null model. To create random datasets in each of 10,000 simulations, I shuffled the residuals from the GAM estimation, and added them to the null values. I then estimated the 5 specifications from Panel IV, keeping track of whether the interaction was $p < .05$

The goal of the simulations is to identify procedures for testing interactions that are curvilinear robust for a broad range of functional forms. I thus run simulations for a broad range of possible scenarios by combining the distribution used to generate x values, the functional form for the association between *x* and *z*, and the functional form of each of the predictors, x and z, and the dependent variable, y. Figure 8 depicts the set of operationalizations considered.

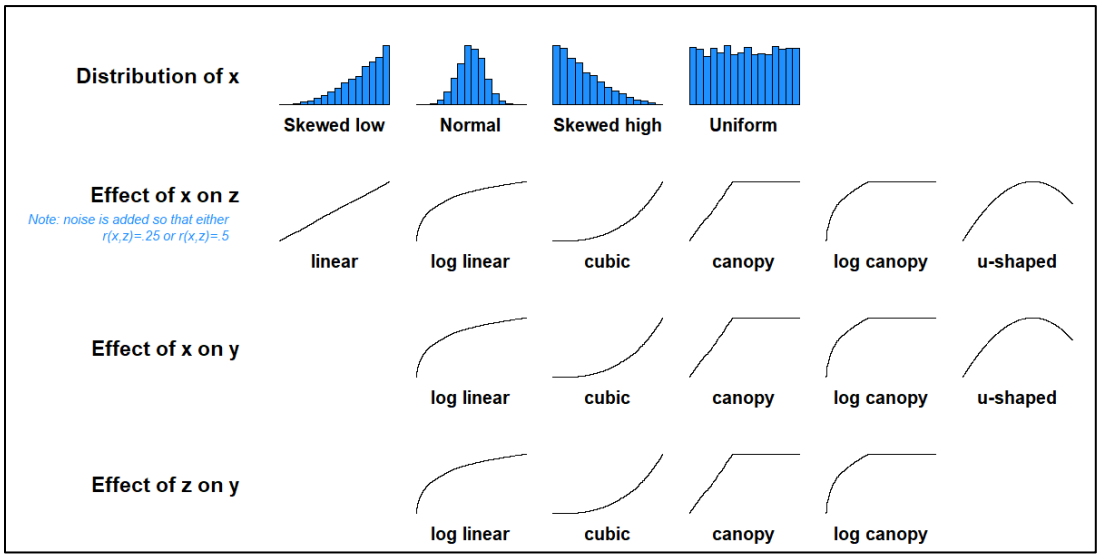

Figure 8. Operationalizations for computing false-positive rate of interaction with correlated predictors
Simulations to evaluate the performance of alternative testing procedures for examining the presence of an *x·z* interaction in the presence of correlated nonlinear predictors considered a wide range of scenarios crossing the operationalizations depicted here. For example, in one scenario, *x* had a skewed high distribution (top-row, 3rd histogram), was correlated r=.25 with z through a log-linear association (2nd row, 2nd plot), x had a cubic effect on y (3rd row, 2nd plot), while z had a log-canopy effect on y (4th row, 4th plot). The full set of combinations of these operationalizations, plus sample size (n=750 vs 1500) and amount of noise in the sample (200% or 400% of the variance caused by x and z) leads to 3840 scenarios. Figure 9 depicts results for simulations of a random sample of 200 of them.
R Code to reproduce figure: https://researchbox.org/313.45 (use code **SNSGCU**).

Crossing the operationalizations depicted in Figure 8, 4x6x2x5x4=960, with whether sample size was N=750 or N=1500 observations, and with whether random noise was twice or four times the variation of *y* caused by *x* and *z*, one arrives at 960x2x2=3840 possible scenarios to simulate. I randomly selected 200 of them, and simulated each 5000 times.

Each simulated dataset was analyzed through seven alternative procedures to test for the presence of an interaction, including the current standard of a simple linear model, and then also adding quadratic terms, adding interrupted quadratic terms, and GAM models where the independent effects

are flexibly estimated and the interaction is still linear. Because in all scenarios there is no actual interaction between x and z, any statistically significant estimate for an interaction is false-positive. The false-positive rates for the alternatives procedures are depicted in Figure 9.

The results show a strikingly inflated false-positive rates for the linear model, the false-positive rate even reaches 100% for some scenarios including the simple dichotomization of the moderator.

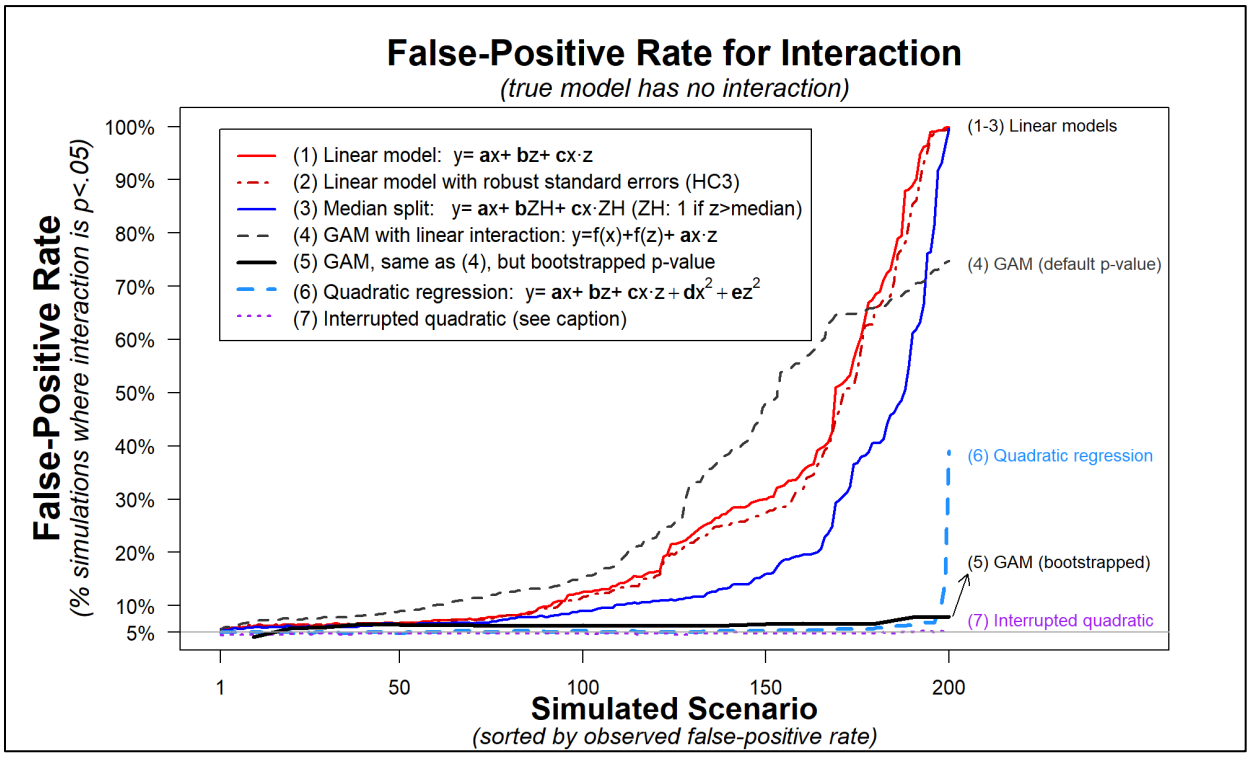

Figure 9. False-positive rates for interactions with nonlinear and correlated x & *z* predictors

The y-axes depict the percentage, out of 1000 simulations Oeach scenario, where the interaction term obtained a statistically significant result ($p \leq .05$), despite the true interaction being zero. The 200 simulated scenarios are generated combining the following operationalizations: x is distributed (i) normal, (ii) left skewed, (iii) right skewed or (iv) uniform. *z* is correlated with *x* with (i) r=.25 or (ii) r=.5, through a (i) linear, (ii) log-linear, (iii) cubic, (iv) canopy, (v) log-canopy, or (vi) u-shaped relationship. The independent effect of *x* on y can have also any of these functional forms, except for linear, and the independent effect of *z* on also can have any of those functional forms, except for linear or u-shaped. The level of noise in the data is either (i) 200% or (ii) 400% of the variation caused by x & *z*. Sample size is n=250 or n=750. This leads to 4x2x6x5x4x2x2 = 3840 scenarios. A random subset of 200 were run. The GAM model was estimated (in R) using package 'mgcv-1.8-34', gam(y~s(x)+x(z)+x:z). In light of its poor performance, a bootstrapped variant was run also, where via case-resampling 100 bootstrapped samples were computed, and the interaction was deemed significant if the 95% confidence interval excluded 0. The interrupted quadratic regression fits quadratic functions of x and *z* separately for values of x and z above/below their respective medians. The reported false-positive rates are adjusted for simulation error expected when the true rate is 5% (see footnote 17). [17]

R Code to reproduce figure: https://researchbox.org/313.46 (use code **SNSGCU**).

[17] When the true probability for a binary outcome is 5% (e.g., when drawing valid *p*-values under the null), if we draw 500 draws each of size 500, as done in the simulations, the expected lowest rate observed is lower than 5%. Indeed, it is just 2%, and the highest rate is not 5%, but 8.2%. The reported false-positive rates are thus adjusted by this expected difference arising from simulation error. E.g., adding 3% to the lowest of 500 false-positive rates, and subtracting 3.2% from the highest false-positive rate, and in between ranks receive in between adjustments.

More surprisingly, Figure 9 shows that the GAM approach also has a very high false-positive rate for the interaction. Wood (2006)'s textbook that accompanies the R package 'mgcv' for GAM estimation provides a possible explanation for this poor performance; he writes that reported *p*-values in GAM models "*are typically lower than they should be . . . because smoothing parameter uncertainty has been neglected in the reference distributions used for testing*" (p.191), but perhaps something is more fundamentally wrong here with the procedure used by the R package to estimate standard errors for a parametric interaction composed of variables that also enter as smooth terms. In any case, at least for now, one should not use precision estimates and *p*-values from GAM models that combine smoothed terms and a linear interaction. An alternative to using the output generated by the mgcv::gam procedure is to bootstrap the GAM model. For illustrative purposes I evaluated this approach relying on 'case resampling', randomly drawing rows of data with replacement, re-estimating the gam model for that resample, and using the distribution of point estimates across bootstraps to approximate the uncertainty of the estimated model (e.g., to build a 95% confidence interval). Figure 9 shows this approach substantially alleviates the problem, 95% confidence intervals for the interaction included 0 in about 90% to 95% of cases, implying still inflated false-positive rate, but much less so.

Turning to the simplest solution, merely adding x^2 and z^2 in the linear model, it achieves near nominal false-positive rates in the vast majority of cases considered, despite substantial specification error (in none of the models is any true relationship quadratic). As was the case with the re-analyzed dataset from Hainmueller et al. (2019), we see a few specifications where this approach suffers from markedly inflated false-positive rates and where only the addition of an interruption (at the median), achieves the necessary flexibility to ensure a nominal false-positive rate throughout.

Earlier proponents (Cortina, 1993; Ganzach, 1997; Lubinski & Humphreys, 1990) of adding quadratic controls to regressions that examine interactions proposed those controls on grounds that nonlinear relationships are likely to be approximately quadratic (see footnote 2). The simulations,

however, show that quadratic controls reduce inflated false-positive rates under much more general circumstances. Intuitively what is required for the quadratic terms to fix the problem of interest is that the nonlinearity in the effects of *x* and *z* on *y* is better approximated by their quadratic terms than by the product of *x·z*.

Perhaps the simplest way to get an intuition for this fact is through a counter-example.[18] Consider a stylized scenario where $y=x^3$ and $z=x^2$. Here estimating the regression model including quadratic controls, $y=\mathbf{a} + \mathbf{b}\, x + \mathbf{c}\, z + \mathbf{d}\, x{\cdot}z + \mathbf{e}\, x^2 + \mathbf{e}\, z^2$, would, despite those quadratic controls, have an elevated false-positive rate for the interaction, **d**. The reason is that the interaction, x·z, is *equal* to the omitted term x^3, so the interaction better captures that nonlinearity of the effect of x on y (which is exactly x^3) than does the control x^2.

It is ultimately an empirical and difficult to answer question whether functional forms in the real world tend to look like the majority of scenarios where the quadratic controls fix the problem at hand, or like the minority of scenarios where it does not and the interrupted quadratic regression is needed. It seems clear, however, that either of these approaches is vastly more likely to be valid than the current default of imposing a linear functional form on all effects.

Best practice might be to estimate both the quadratic and the interrupted quadratic regression, and if the results are qualitatively consistent, focus the discussion on the easier to interpret simpler model.

Statistical Power.

I conducted additional simulations where the true interaction effect consisted of a linear (x·z) or a nonlinear $(x{\cdot}z)^{.5}$ interaction, and compared the statistical power of these two solutions, quadratic vs interrupted quadratic regression. I calibrated the data generating process to give the quadratic model about 50% power, finding that the interrupted quadratic obtained about 5% less power on average (45%),

[18] This example is very similar to, and was inspired by, Simulation 6 in Table 3 by Matuschek and Kliegl (2018).

albeit for a substantial minority of scenarios power was higher for the interrupted regression. Without knowing functional form, therefore, it is not clear if one procedure is generally more powerful than the other.

Probing interactions with non-experimental data

Earlier we looked at the consequences of nonlinear effects when probing *x·z* interactions from experimental data (r(x,z)=0), arriving at two key conclusions. First, the most common approach for probing interactions in social science, combining point estimates from a linear regression (again, known across different fields as Johnson Neyman procedure, simple slopes, spotlight, pick-a-point, and conditional marginal effect), is invalid "if" the impact of *x·z* on *y* is not linear. Recall for instance how in Figure 1 baby girls were incorrectly predicted to be heavier than baby boys. Second, curvilinear-robust alternatives for probing interactions with data from experiments included (i) dichotomizing the moderator to compare the effect of *x* for high vs low values of z, and (ii) the GAM Johnson-Neyman/GAM simple slopes approach.

Here we consider the probing of interactions from non-experimental data, where *x* and *z* may be correlated, making three main points: (1) the traditional approach of probing interactions by combining point estimates from linear regressions in this context can be even *more* misleading, producing results that are entirely unrepresentative of reality, (2) probing interactions by dichotomizing the moderator no longer produces curvilinear-robust results; indeed, results from median-splits or from the 'binning estimator' by Hainmueller et al. (2019), can be *as biased* as those arising from the regression that forces the effect of z and *x·z* to be linear throughout, and (3) that GAM Johnson-Neyman/GAM simple procedures, continue to provide curvilinear-robust results. I next provide a simple example that illustrates all three of these points.

A stylized example of probing interactions between correlated predictors

Let's consider a scenario where a dependent variable *(y)*, say happiness, has an inverted u-shaped relationship with a predictor, say age (*x*), and a monotonic relationship with another predictor, say wealth (z), and where the two predictors are correlated r(*x,z*) = .5 (people are happiest during midlife, wealthier people are happier, and wealth is correlated with age). For concreteness sake, I simulated one dataset with 1000 observations where the true model is, y = 0.5*x* - x^2 + z + e, with *x, z and e* distributed N(0,1). Note how the true model does not have an interaction, the effects of age and wealth are independent of one another.

I then estimated the linear regression y=**a**+**b**x+**c**z+**d***x·z,* obtaining y = - **.55 + .64** *x* **+ .96** *z* **- .78** *x·z*. The *p*-value for $\hat{d}$ was p<.0001, confidently but incorrectly, suggesting the presence of an interaction. The motivation to probe interactions, in general, is to facilitate the interpretation of point estimates. In this case in particular, to facilitate interpreting that $\hat{d}$= -.78. Let's now turn to Figure 10 which depicts side by side the true and estimated associations between happiness and age, for low, medium, and high values of the moderator, wealth.

The first panel depicts the true association, parallel u-shapes: age has a u-shaped effect, and wealth impacts happiness, but independently of age. There is no interaction. The effects are additive. The second panel shows the utterly misleading depiction a researcher relying on a linear model combined with simple slopes would produce. To be clear, the goal of the simple slopes in the second panel is to provide an easy to understand approximation of the first panel.

The (linear) simple slopes are wrong in a predictable and intuitive fashion. Understanding why these specific simple slopes are wrong provides an understanding for why they will be generally wrong. The simple slopes here imply that for wealthy individuals, the effect of age on happiness is negative, while for poor individuals it is positive. In fact, the relationship across groups is identical. The reason for this erroneous conclusion is the correlation between wealth and age. While the linear model cannot

accommodate the effect of age changing with different ages, it can accommodate it changing with different wealth levels, and since wealth and age are correlated, it does. Specifically, the regression makes wealthier people show negative effects of age, to approximate older people having a negative effect of age. Lastly, the third panel shows how the GAM model in general, and GAM simple slopes in particular, accommodate the nonlinearity and produce results that do resemble reality.

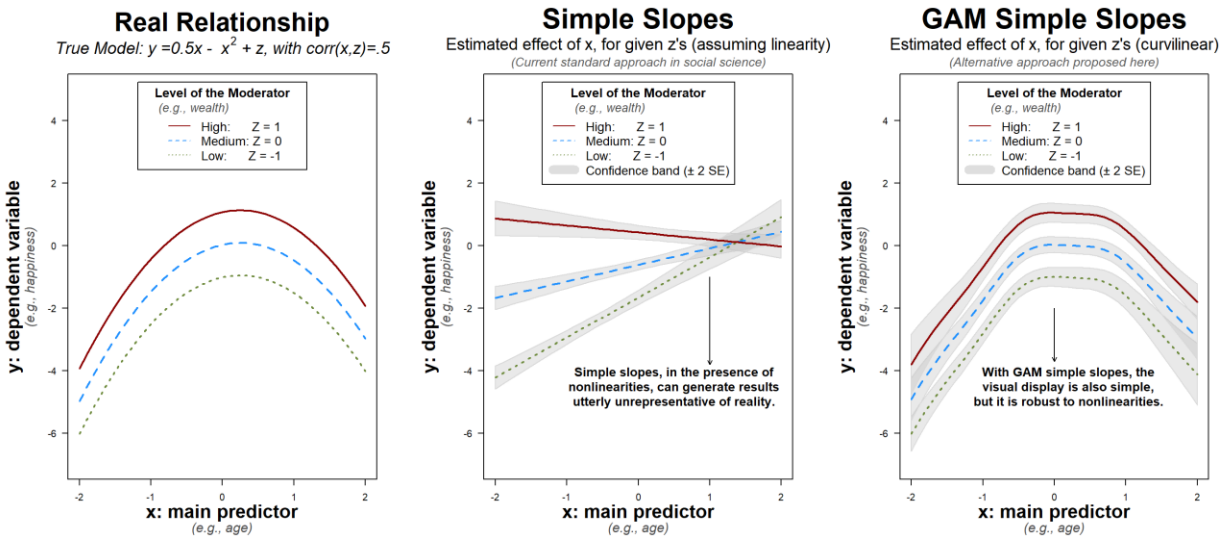

Figure 10. Simple slopes to probe interactions with correlated predictors

A single simulation was performed with 1000 observations where x & *z* are each distributed N(0,1) and are correlated r=0.5; random normally distributed noise is added to create the y-values. The second panel depicts *linear* simple slopes. It is produced following the current standard practice of using the estimated linear regression coefficients to project linearly how the effect of x changes for different value of *z.* The third panel depicts '*GAM* simple slopes', these are produced by first estimating a GAM ('general additive model', see Wood, 2006), and then obtaining fitted values for different values of one predictor, x, keeping the value of the other predictor(s) constant.

R Code to reproduce figure: https://researchbox.org/313.49 (use code **SNSGCU**).

Considering that linear probing of interactions leads to invalid inferences even with uncorrelated nonlinear predictors, it is perhaps not particularly surprising that it is also invalid when the predictors are correlated. But the fact that correlated nonlinear predictors invalidate dichotomized interactions is less intuitive. When we dichotomize *z,* nonlinearities in the underlying effect of z on y indeed do not introduce bias, the effect of high vs low z is necessarily linear; the connection of any two points is linear. But, nonlinearities in the effect of x on y, do still matter, because we have left x as a linear predictor. And therefore, the linear model will again bias the interaction; while it cannot change the effect of x for higher x values, it can change it for higher *x·z* values, and it does. Returning to the happiness example, if we dichotomize wealth and probe an age*dichotomized_wealth interaction, the wealth effect is indeed now necessarily linear, but the age effect is not, and thus effect of age among the wealthy will be biased towards the effect of age among the older.

Figure 11 illustrates. The left panel shows three alternative methods for computing marginal effects of x for different values of z, all leading to invalid inferences. Most relevant for our purposes, it shows the (linear) Johnson-Neyman procedure, and Hainmueller et al. (2019)'s binning estimator (a tertile split on z) being entirely consistent with one another's incorrect conclusions. The tertile split measures the average effect of age for low, medium, and high levels of wealth, assuming the effect of age is linear. The right panel shows how the GAM Johnson Neyman procedure does not suffer from this problem.

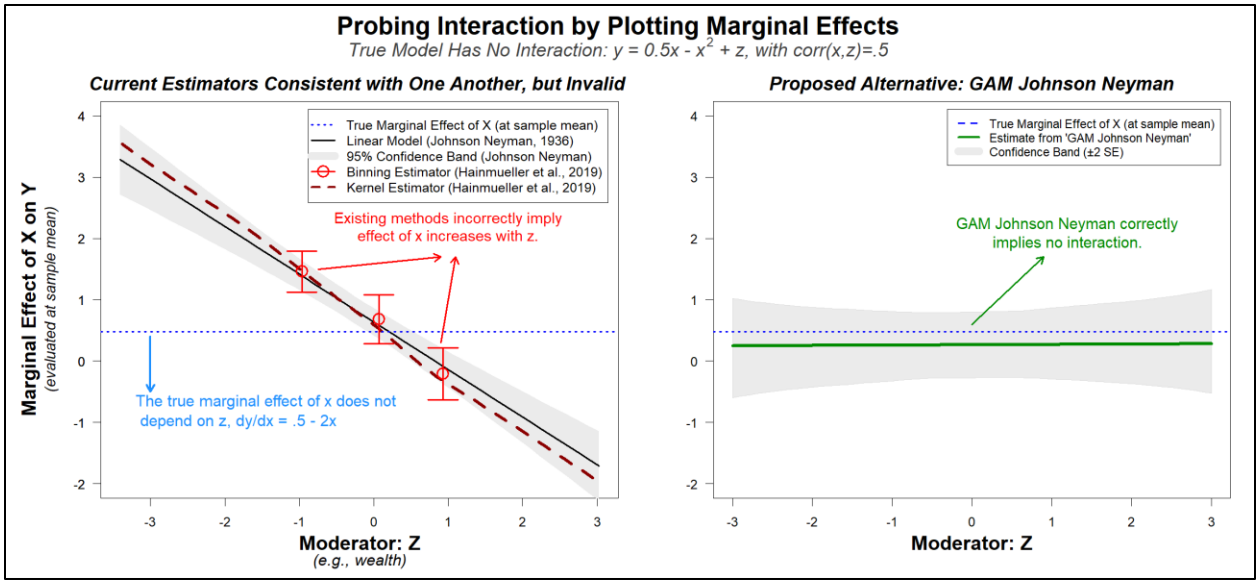

Figure 11. Plotting marginal effect of x across different z values, when x and z are correlated

A single simulation was performed with 1000 observations where x & *z* are distributed N(0,1) and are correlated r=0.5. The true model is $y = x^2 - 0.5x + z$, plus random (normal) noise. A linear model, y=ax+bz+cxz is estimated, and the resulting interaction then probed using the R Package 'interflex' created by Hainmueller et al., using their two proposed approaches, (i) "binning" and (ii) "kernel" estimator.[19] The figure shows that both robustness checks lead to misleading 'corroboration' of the spurious interaction between *z* and x obtained from the mis specified linear model. The probed interaction produces a false-positive sign reversal of the effect of x on y as a function of *z.* The linear model, the binning estimator, and the kernel method, are all invalid because of the simultaneous presence of nonlinearities and correlated predictors which require a more flexible approach than any of them provide.

R Code to reproduce figure: https://researchbox.org/313.49 (use code **SNSGCU**).

Caveats with using GAM to probe interactions with nonexperimental data

The results above suggest that the best tool for probing interactions in nonexperimental data, where predictors are correlated, is GAM Johnson-Neyman / GAM simple slopes. These GAM based solutions seem to be vast improvements over the status quo of relying on models that impose, but two caveats are worth keeping in mind.

First, as discussed above, the standard errors produced by GAM tend to be too small (though for experiments, the issue seems ignorable). While a general assessment of the performance of bootstrapping

[19] The figure was created using output from the 'interflex' package created by the same authors from the cited paper. The default figure produced by the package was not used, instead, the output was saved as an object and a new figure that combines both the binning and the kernel estimator in a single figure was created. R packages can be updated at any point, the results are based on "interflex" version 1.2.6, published on May 18th, 2021.

of GAM models is beyond the scope of this paper on interactions, it seems wise, at least until this issue is explored more systematically, to bootstrap GAM models when analyzing nonexperimental data and when estimates of the precision are at least somewhat relevant to the research question (i.e., to not take R's mgcv::gam *p*-values at face value). Second, while GAM models in experiments are simple enough that we should expect them to work with typical sample sizes, for nonexperiments, the growth in number of parameters to be estimated may lead to sample size requirement that exceed the sample size available. This is a common problem with flexible models, known as the 'curse of dimensionality', where sample size needs grow faster than the number of parameters estimated. This means that with some samples where we could probe al linear interaction, we would not have enough observations to probe a GAM interaction. In those cases, the researcher's choice is between an incorrect answer, and accepting the question of interest cannot be answered with the available data.

Conclusions

This article is organized around testing and probing interactions, for experimental and nonexperimental data. For probing interactions with experimental data, the tools discussed here provide a clear improvement over current practice, with little if any downside. Researchers can expect a substantial reduction in false-positive rates, and a likely *increase* in statistical power, if they abandon the eight-decades old idea of probing interactions with a linear model (Johnson & Neyman, 1936). To put it bluntly, the influential textbook by Aiken and West (1991), and the seminal tutorials by Preacher et al. (2006) and Spiller et al. (2013), should no longer be used to guide practice by professional researchers. The reliance on unjustified and extremely consequential linearity assumptions, renders the recommended techniques in these sources as inadequate. Moving from simple slopes, to GAM simple slopes, and from Johnson-Neyman to GAM Johnson-Neyman is as close to a free lunch as it gets in statistical analysis.

For nonexperimental data the message of this article is more nuanced. First, in terms of testing interactions, adding quadratic terms (x^2 and z^2) to the linear regression $y=\mathbf{a}+\mathbf{b}x+\mathbf{c}z+\mathbf{d}x{\cdot}z$ leads to a substantial reduction in false-positive rates, generally down to the nominal 5% level, but there are realistic circumstances (e.g., the data reanalyzed by Hainmueller et al. (2019), see Figure 7) where it is an insufficient solution and the false-positive rate is still well above 30%. Adding *interrupted* quadratic terms, allowing the slopes of x, x^2, z and z^2 to change at the median of x and z respectively, does in all scenarios considered generate the nominal false-positive rate of 5% for the interaction. Both solutions, adding quadratic and interrupted quadratic terms, can lead to power losses for testing the interaction in relation to the linear model, but among these two, one is not generally more powerful than the other. The loss in power seems unavoidable; when we decide to abandon a statistical tool that easily reaches a 100% false-positive rate, naturally it becomes harder to obtain a $p<.05$ result. It is nevertheless easier to obtain a *diagnostic* result.

Lastly, when it comes to probing interactions with nonexperimental data, relying on fully flexible models, e.g., with GAM simple slopes and GAM Johnson Neyman, seems to be the only viable option. In some circumstances the sample size needs for these approaches will be larger than the data available (i.e., the probed interaction will be too imprecise to be usefully interpreted). In such cases it seems better to accept the reality that we cannot meaningfully probe the interaction, that it is to pretend we can by assuming linearity throughout, obtaining an answer that is so easily utterly disconnected from reality.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*: Sage.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources*, 436-455.

Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis, 14*(1), 63-82.

Clark, W. R., & Golder, M. (2006). Rehabilitating Duverger's theory: Testing the mechanical and strategic modifying effects of electoral laws. *Comparative Political Studies, 39*(6), 679-708.

Cohen, J. (1983). The Cost of Dichotomization. *Applied psychological measurement, 7*(3), 249-253. doi:10.1177/014662168300700301

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, New Jersey 07430: Lawrence Erlbaum Associates, Inc. Publishers.

Cortina, J. M. (1993). Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. *Journal of Management, 19*(4), 915-922.

DeCoster, J., Iselin, A.-M. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological methods, 14*(4), 349.

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

Gelman, A., & Park, D. K. (2008). Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician, 62*(4), 1-8.

Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis, 27*(2), 163-192.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association, 82*(398), 371-386.

Humphreys, L. G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables.

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M. J., & Popovich, D. L. (2015). Toward a More Nuanced Understanding of the Statistical Properties of a Median Split. *Journal of Consumer Psychology, 25*(4), 652-665. doi:10.1016/j.jcps.2014.12.002

Jann, B. (2008). The Blinder–Oaxaca decomposition for linear regression models. *The Stata Journal, 8*(4), 453-479.

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

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6*(3), 312-319.

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

Matuschek, H., & Kliegl, R. (2018). On the ambiguity of interaction and nonlinear main effects in a regime of dependent covariates. *Behavior research methods, 50*(5), 1882-1894.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological bulletin, 113*(1), 181.

McClelland, G. H., Lynch, J. G., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false–positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology, 25*(4), 679-689.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review*, 693-709.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31*(4), 437-448.

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology, 25*(4), 666-678.

Simonsohn, U. (2018a). DataColada [57] - Interactions in Logit Regressions: Why Positive May Mean Negative. Retrieved from http://datacolada.org/57

Simonsohn, U. (2018b). Two lines: a valid alternative to the invalid testing of U-shaped relationships with quadratic regressions. *Advances in Methods and Practices in Psychological Science, 1*(4), 538-555.

Simonsohn, U., & Gino, F. (2013). Daily Horizons Evidence of Narrow Bracketing in Judgment From 10 Years of MBA Admissions Interviews. *Psychological science, 24*(2), 219-224.

Spiller, S. A., Fitzsimons, G. J., Lynch Jr, J. G., & McClelland, G. H. (2013). Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *Journal of Marketing Research, 50*(2), 277-288.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review, 64*(3), 153.

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40*(2), 145-160.

Wood, S. (2006). *Generalized additive models: an introduction with R* (1st ed.): Chapman & Hall/CRC Monographs on Statistics and Applied Probability.