First draft: 2021 11 25
This draft: 2023 07 17
Newest draft: http://urisohn.com/42

Interacting With Curves: How to Validly Test and Probe Interactions in the Real (Non-linear) World

Uri Simonsohn
ESADE Business School
urisohn@gmail.com

Abstract

Hypotheses involving interactions, where one variable modifies the association between another two, are very common. They are typically tested relying on models that assume effects are linear, e.g., with a regression like y=**a**+**b**x+**c**z+**d**x·z. In the real world, however, few effects are linear, invalidating inferences about interactions. For instance, in realistic situations, the false-positive rate can be 100% for detecting an interaction, and a probed interaction can reliably produce estimated effects of the wrong sign. This paper proposes a revised toolbox for studying interactions in a curvilinear-robust manner, giving correct answers 'even' when effects aren't linear. It's applicable to most study designs, and produces results that are analogous to those of current -often invalid- practices. The presentation combines statistical intuition, demonstrations with published results, and simulations.

Data and code to reproduce all results are available from:
https://researchbox.org/1569 (use code **TYBYZK**)

Studying interactions, where a variable moderates the relationship between two other variables, is common in psychology. For instance, 71% of articles in the March 2020 issues of JPSP, JEP:G, and *Psychological Science*, test for interactions. The general approach to studying interactions is the same for the majority of statistical models commonly used by social scientists (e.g., linear & logit regression, multilevel models, structural equation modelling), and it consists of three steps. For concreteness I discuss them relying on a stylized scenario where we wish to examine the interaction between the effects age and gender on weight. In the first step one estimates a model that imposes the assumption that all predictors (including possibly non-linear terms), and including the interaction, have linear associations with the (latent if applicable) dependent variable.[1] For example, in the first step one estimates this regression: $\text{weight} = \mathbf{a} + \mathbf{b}\ \text{female} + \mathbf{c}\ \text{age} + \mathbf{d}\ \text{age} \cdot \text{female} + \varepsilon$. In the second step one *tests* the interaction, evaluating whether the estimate of d, is significantly different from zero. In the third step one *probes* the interaction, assessing how much the effect of gender changes as a function of age, combining the point estimates of b and d. In psychology, the most common procedure for probing interactions consist of computing "simple slopes" (Aiken & West, 1991; Preacher, Curran, & Bauer, 2006), reporting the effect ('slope') of gender at specific values of the moderator age, e.g., 1 SD away from the mean.[2]

This article is concerned with the consequences of violating (the often implausible) linearity assumption in the first step, on the validity of the results in the 2nd and 3rd steps. These consequences depend on the nature of the variables in the interaction: whether at least one of them is randomly

[1] All these models do of course *allow* for non-linear terms, e.g., $y=\mathbf{a}x+\mathbf{b}x^2$, but they are very rarely included when testing interactions in practice. Moreover, prominent textbooks warn readers about likely downsides of doing so (Cohen, Cohen, West, & Aiken, 2003, p. 300)

[2] Other approaches and names for probing interactions include the Johnson-Neyman procedure (Johnson & Neyman, 1936), regions of significance (Preacher et al., 2006), pick-a-point, spotlight & floodlight analysis (Spiller, Fitzsimons, Lynch, & McClelland, 2013). In economics and political science it is common to use the umbrella term 'marginal effect' for these same computations (Ai & Norton, 2003; Greene, 2010).

assigned (e.g., in an experiment) vs not (e.g., in observational data).[3] For reasons to be discussed later, when at least one factors in the *x·z* interaction is randomly assigned, *testing* remains valid in the presence of non-linearities, but *probing*, with any of the methods listed above, such as simple slopes, the Johnson-Neyman procedure, and also the more recently proposed binning estimator (Hainmueller, Mummolo, & Xu, 2019), is invalid. When neither x nor z in *x·z* are randomly assigned, then, non-linearities invalidate both the testing and probing of interactions with all commonly used methods to study interactions in the social sciences.

As a motivating example, Figure 1 illustrates one way in which non-linear relationships invalidate the probing of interactions using the previous age and weight example. Panel B highlights that there is a complex interaction with gender and age, such that males get heavier than females starting at age 14 or so. A linear model cannot represent such a non-linear interaction, and we see in C how when we rely on the linear model, we end up projecting an incorrect sign reversal for babies, such that baby girls are predicted (incorrectly) to be heavier than baby boys.

Panel C shows, then, that probing the interaction by plotting the effect of gender for all ages, i.e., relying on the Johnson and Neyman (1936) procedure, one falsely but confidently concludes that baby girls are substantially heavier than baby boys.

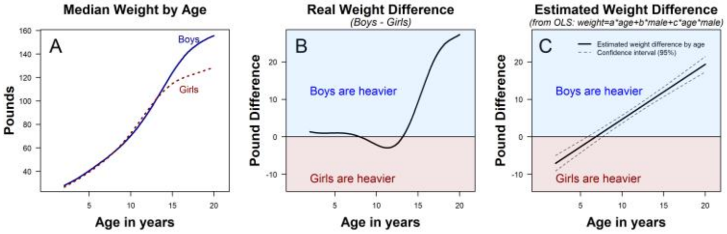
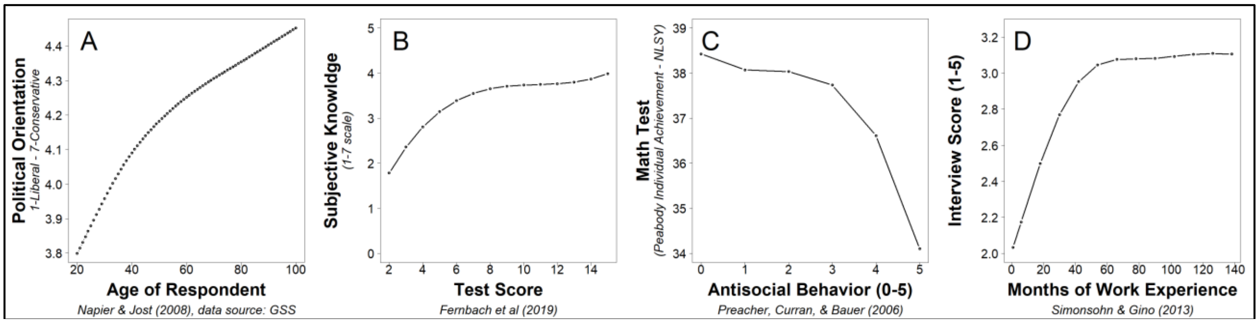


Fig. 1. *Non-linear Effects Lead to Misleading Interaction Results*
R Code to reproduce figure: https://researchbox.org/1569.2 (use code **TYBYZK**).

[3] Technically speaking the consequences depend on whether the two predictors in the interaction are statistically independent. In an experimental design, where treatment is randomly assigned, usually we expect predictors to be statistically independent. In observational data, they are almost never statistically independent. For simplicity I thus speak of experiment vs observational, instead of independent vs dependent predictors.

There are good reasons to expect that non-linearities, like those invalidating the linear regression from Figure 1, are common in data collected by social scientists. From psychology we know that perception of change in physical and numerical stimuli exhibits diminishing rather than constant sensitivity (Fechner, 1860; Kahneman & Tversky, 1979). From economics we know marginal benefit is diminishing rather than constant, and marginal cost is increasing rather than constant. Indeed, the foundation of supply and demand being upward and downward sloping respectively, arises from this non-linear relationship between inputs and outputs (see e.g., the principles of economics textbooks by Greenlaw & Shapiro, 2017). In addition, many of the variables collected by social scientists consist of bounded scales which inevitably show diminishing rather than constant effects, as some participants hit the ceiling or floor of the scale and can no longer be impacted by further changes of the predictor of interest.

Any study that involves the perception of physical or numerical stimuli, the presence of costs or benefits, or measurement through scales, then, is likely to involve non-linear relationships. Figure 2 provides some concrete examples of the kinds of non-linear relationships we observe in real data.



Napier & Jost (2008), data source: GSS

Fernbach et al (2019)

Preacher, Curran, & Bauer (2006)

Simonsohn & Gino (2013)

Fig. 2. *Examples of non-linear associations in real data.*

In all panels, the lines are formed by fitting the data with flexible models (Generalized Additive 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 on attitudes towards Genetically Modified (GM) foods. **C,** N=956 survey respondents from the NLSY. **D,** N=12427 interviews of applicants to an MBA program.
R Code to reproduce figure: https://researchbox.org/1569.5 (use code **TYBYZK**).

Prior work on non-linearities 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 et al., 2003; Preacher et al., 2006; Spiller et al., 2013). Aiken and West (1991) alone accumulate, as of May 2023, over 54000 Google citations, and Preacher et al., another 5500. 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 non-linearities 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, on establishing whether there is a statistically significant interaction, at least three papers (Cortina, 1993; Ganzach, 1997; Lubinski & Humphreys, 1990) have warned that if the effect of *x* or z on *y* are not linear, and *x* is correlated with *z,* the estimate of the interaction is biased, and its false-positive rate is inflated (the intuition for this bias is presented later in this paper). Throughout this article I refer to it as the problem of "**correlated non-linear predictors**." The authors of these 1990s articles assumed that all non-linearities are essentially quadratic, and thus considered true models only of this form: $y=a+bx+cz+d\ x\cdot z+ex^2+fz^2+\ \varepsilon$ (see footnote 4 for quotes).[4]

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, to address the problem of

[4] The following quotes, with emphasis added, document equating nonlinear 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**."

correlated non-linear predictors. In relation to this work, I relax the assumption that all non-linear relationships are quadratic, and relax the assumption that when *x* and *z* are correlated, *their* association is linear. Moreover, I empirically evaluate how well the proposal of including x^2 and z^2 as covariates performs for testing the *x·z* interaction with non-linear correlated predictors, finding it performs surprisingly well, but not best; having higher false-positive rates, and lower power, than the Generalized Additive Model (GAM) based alternative proposed here.

In terms of *probing* interactions, on estimating the effect of *x* on *y* for different values of *z,* Hainmueller et al. (2019) discussed two problems with current practice (in political science). First, researchers sometimes probe interactions at moderator values for which no data exist, and second, sometimes *interaction effects* are non-linear.[5] They introduce the 'binning estimator' as an alternative to the linear probing of interaction, to addresses these two shortcomings. The binning estimator splits the data into segments based on the moderator (e.g., low, medium and high values of the moderator), and estimates separate linear models within segments.[6] The most important contrast between their work and the present paper, is that dichotomizing in general and the 'binning estimator' in particular *does not address* the threat posed by ***correlated non-linear predictors***. The results are often as invalid with this approach to probing interactions, as they are with the approach based on the simple linear model. See Supplement 3 for the intuition and an example. In addition, their paper focuses only on probing, and does not discuss the testing interactions. Lastly, their paper does not evaluate via simulations how well (or poorly) their proposed tools work.

Finally, in terms of interpreting interactions, in terms of assessing whether a validly tested and probed interaction has the implications for the theory that motivated the study, Loftus (1978) made the

[5] The Hainmueller et al (2019) abstract reads "Current empirical practice tends to overlook two important problems. First, these models assume a linear interaction effect that changes at a constant rate with the moderator. Second, estimates of the conditional effects of the independent variable can be misleading if there is a lack of common support of the moderator."

[6] The three independent segments are estimated jointly, thus, in the presence of covariates, the results may be more efficient that literally estimating three separate regressinos. See their equation (4).

important observation that when the variables we study are proxies for the latent variables of interest, an observed interaction with measured variables need not imply an interaction among the latent variables (see also Krantz & Tversky, 1971). This observation is important and unfortunately has been largely ignored by researchers (Wagenmakers, Krypotos, Criss, & Iverson, 2012), but it is separate from the issues that concern this current article. Loftus's observation is about the interpretation of statistically valid interactions, this article is about the statistical (in)validity of interaction terms in linear models.

Alternatives to assuming linear effects

In this paper I consider three main approaches for relaxing linearity assumptions: dichotomization, adding quadratic terms to a linear regression, and relying on generalized additive models (GAM). I provide overviews of the three approaches next, with more attention paid to the more novel approach: GAMs.

Approach 1: Dichotomization.

The first and simplest approach for handling non-linearities is to force linearity by dichotomizing the moderator. Rather than taking age as a continuous variable, we classify boys and girls into, say, above and below median age, and carry out a simple 2x2 comparison of the four means. With dichotomization, the intuition goes, the linearity assumption (for the moderator) cannot be violated, because two points always form a straight line. Dichotomization has long been relied on by psychology researchers, usually on grounds of its "analytical ease and communication clarity" (Iacobucci, Posavac, Kardes, Schneider, & Popovich, 2015, p. 652). Dichotomization has also long been objected to by psychology *methodologists* on grounds that it has lower statistical power (Cohen, 1983).[7] Considerations of lower statistical power aside, dichotomization has a subtler but more serious problem. As will be demonstrated in a later section, when the two predictors in the interaction are correlated (e.g., what we generally expect when neither x nor z were randomly assigned), underlying non-linearities in x can invalidate interactions with median splits of z *as much* as they invalidate interactions with continuous z. Thus, for testing interactions between correlated non-linear factors, median splits suffer from elevated Type 2 and also Type 1 errors.

Approach 2: Adding quadratic controls (x^2 and z^2)

As mentioned above, a few authors have advocated including quadratic terms of x and z, when estimating regressions with the purpose of testing an $x{\cdot}z$ interaction (Cortina, 1993; Ganzach, 1997; Lubinski & Humphreys, 1990). They conjectured quadratic controls were sufficient to deal with any non-linearity (see footnote 4), but they did not evaluate such conjecture. They did not carry out simulations assessing how well quadratic controls work if real functional forms are neither exactly linear nor exactly quadratic. I carried out that evaluation here. As will be shown later, I find that quadratic controls perform surprisingly well under a broad range of (non-quadratic) functional forms, but that they are on occasion insufficiently flexible to ensure valid inferences in the presence of realistic non-linearities. Quadratic controls, moreover, can lead to substantially lower statistical power than do the procedures proposed in this article. Lastly, when it comes to probing interactions, to assessing how big the effect of x on y is for different values of z, quadratic controls only lead to correct estimates if the true functional form is exactly quadratic, an arbitrary and (practically speaking) untestable assumption.

[7] Many articles have echoed Cohen's arguments against dichotomization based on statistical power considerations (DeCoster, Iselin, & Gallucci, 2009; Humphreys & Fleishman, 1974; Lubinski & Humphreys, 1990; Maxwell & Delaney, 1993; McClelland, Lynch, Irwin, Spiller, & Fitzsimons, 2015).

Approach 3: Generalized Additive Models (GAM)

A third approach moves us away from arbitrary assumptions about functional from (quadratic) and arbitrary dichotomizations (median split), and towards estimating the functional form of interest. A few procedures allow flexible functional form estimation (e.g., LOESS, kernel regression), but the one that seems most applicable to a broad range of data structures, in terms both of analytical flexibility and computational efficiency, while providing interpretable enough estimates is "Generalized Additive Models", GAM for short (Hastie & Tibshirani, 1987; Wood, 2017). For example, LOESS estimation does not produce results for individual predictors, making statistical inference (e.g.. "is the interaction significant?") more difficult to conduct.

While GAMs were developed decades ago, they have not been used much in psychological research yet. [8] I hope this paper will change that. GAMs are conceptually similar to linear regressions, in that they estimate the relationship between predictor variables and a dependent variable. The key difference is that regressions assume all entered effects are linearly associated with the (sometimes latent) dependent variable, while GAMs *estimate* the functional forms of each effect.[9]

From a user perspective, relying on GAMs can be quite similar to relying on linear regressions. I will next illustrate with a simple example, written in R, where data are analyzed with a linear regression and with a GAM side by side. We start simulating data, making variables x, z, and e, be (standard) normal, and making y depend linearly on them,

```
set.seed(123)
n = 500
x = rnorm(n)
z = rnorm(n)
```

[8] A Google Scholar search for "generalized additive model" on May 2022, found only 3 *Psychological Science* articles (FitzGibbon, Komiya, & Murayama, 2021; E. L. James et al., 2015; Ramscar, Sun, Hendrix, & Baayen, 2017).

[9] Note that a regression is linear even if it includes non-linear terms. For example, the regression $y=x+x^2$ +e is still 'linear', in the sense that the effect of x^2 on y is assumed to be constant. When x^2 increases by 1, y increases by the same amount, no matter how big or small x^2 is. In addition, one can force the linearity assumption on some GAM terms, estimating, for example gam(y~s(x)+z), where z would enter linearly, as in a linear regression.

```
e = rnorm(n)
y = x + z + x*z + e
```

We can estimate a linear regression with:

```
lm1 = lm(y~x+z+x:z)
```

And we can estimate a GAM with:

```
gam1 = gam(y~s(x)+s(z)+ti(x,z))
```

In the GAM, s() indicates a "smooth" (flexible functional form) main effect, and ti() a flexible interaction term. The output for lm1 is one we are familiar with: four point estimates. The GAM, like the linear model, produces 4 *p*-values, for the intercept, the two predictors, and their interaction. But when it comes to the coefficient estimate, interpretation is more difficult. For the example above, for instance, GAM produces 35 instead of 4 coefficients. I argue here we should tolerate GAM's harder to interpret output for two key reasons. First, GAM's output is more likely to be statistically valid and descriptively accurate. As we shall see later in this paper, the linear model's easy to interpret results can be extremely misleading and statistically invalid, especially for interactions. We can more easily read the output from a regression, then, but that easy to read output from the regression is also too easily wrong. The second reason I think we should tolerate GAM's harder to interpret output, is that, we can make it interpretable without much effort, relying on the same techniques we rely on to make regression coefficient estimates for interactions interpretable, the aforementioned 'simple slopes' (Aiken & West, 1991).

As interpretable as the output for a linear regression seems, when interactions are involved, it is actually not that easy to interpret. As researchers we want to know 'the' effect of x on y, but when an interaction is involved, there may be infinite effects of x on y, one for each possible value of z. This is why to interpret regressions we probe them (Aiken & West, 1991). With simple slopes in particular, we compute the expected value of y, for all values of one predictor, keeping the other predictor(s) fixed at a given value.

Returning to the simple example above, the estimated regression equation is $y = .02 + 1.06x + 1.06z + .975x \cdot z$. If we are interested in "the" effect of x on y, that regression equation gives us a different estimate for every possible value of z. With simple slopes we would fix z at some point, let's say at z=1, and plot the relationship between y and x for z=1. With GAM, we can do the same calculation, reducing the interpretability disadvantage that GAMs suffer from. Moreover, simple slopes for the linear model and for GAM can be estimated with the same (wrapper) function in R, 'predict'.[10] The code below, carries out simple slope calculations for the linear model and for GAM:

```
#Values of x to consider
   xs = seq(-2,2,.1)
#linear simple slopes when z=1
  yh1=predict(lm1,newdata=data.frame(xs,z=1))
#GAM simple slopes when z=1
  yh2=predict(gam1,newdata=data.frame(xs,z=1))
```

The first line of code generates the values of x we are interested in probing, going from -2 to 2 in increments of .1. The second line of code produces the expected value of y, given the linear model results, for x values between -2 and 2, when z=1. The third line does the same given the GAM results.

An (older) alternative to simple slopes is the Johnson and Neyman (1936) procedure (also known as floodlight analysis and 'regions-of-significance'). It estimates the *marginal effect* of one variable, for every possible value of the other. In this article I focus on, and advocate for, GAM simple slopes rather than GAM Johnson Neyman because depicting the levels of the dependent variable, rather

[10] As was pointed out by a reviewer, while the R function is the same, stats::*predict()*, that function is just a wrapper which invokes different underlying functions when applied to a linear regression object (predict.lm()), vs a GAM object (predict.gam()).

than the marginal effects, makes it easy to give results a proper contextualized interpretation. But GAM Johnson Neyman is easy to implement as well. Supplement 5 has an example with R Code.

GAMs can be applied to a wide range of data structures. One can estimate a GAM when the dependent variable is binary or categorical (in lieu of probit or logit regressions). One can also include random effects into GAMs (Simpson, 2021; Wood, 2013, 2017), and estimate multilevel models see e.g. the `gamm4` package in R.

GAMs' estimation procedure includes a penalty for overfitting; in practice this means that if an association is best described by a linear model, GAMs will tend to deliver a linear model, but if it is best represented by a cubic, or sine function, or combination of the two, it will tend to deliver that instead. The performance of GAMs in recovering functional form is impressive for researchers who still rely on the 19th century technology of fitting data with straight lines. See Figure 3.

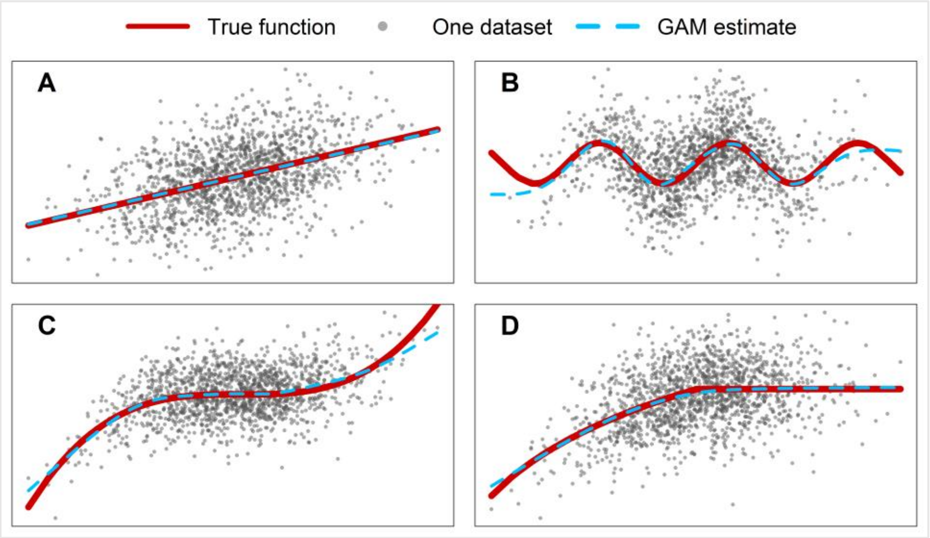


Fig. 3. *Examples of GAMs correctly recovering underling functional forms*

The four panels are based on the same draw of N=1500 x-values dawn from N(0,3), with the y-values corresponding to A:y=x, B:y=sin(x), C:y=x^3-x^2, and D:min(log(x+14),log(14)). Random noise N(0,SD) was added, with SD equal to twice the standard deviation of Y caused by x. R Code: https://researchbox.org/1569.15 (use code **TYBYZK**).

In the next section I demonstrate the application of GAM simple slopes to data from psychology papers, but first, in a short subsection, I provide a quick summary of the proposed toolbox I put forward in this paper for testing and probing interactions in social science.

Toolbox preview

The goal of this paper is to deliver a curvilinear-robust toolbox for studying interactions in social science. In the remaining sections, I motivate and demonstrate the tools in the proposed toolbox by re-analyzing published data. I then evaluate the validity of the tools, for a broad range of scenarios, via simulations. Anticipating the conclusions of these analyses, Table 1 summarizes the proposed toolbox.

	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. Randomized x or z (e.g., experiments) *r(x,z) = 0*	- Linear model is OK	- Good: Discretize z (e.g., median split) - Better: GAM Simple Slopes
Case 2. Neither x nor z randomized (e.g., observational data) *r(x,z) ≠ 0*	- Often adequate: linear model controlling for x^2 and z^2 - More reliable: GAM model	- GAM Simple Slopes

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

Examples of GAM simple slopes with data from psychology papers

In this section I report simple slopes, both linear and GAM, constructed based on data from two published paper. Let's first return to the MBA interview data from Figure 2D. The dataset includes applicant's work experience and country of origin. A linear regression using these variables to predict interview score, results in: *score = 2.2 + .013·experience - .166·American + .005·American·experience.*

Figure 4 shows the linear simple slopes implied by that equation. We learn that Americans benefit more from experience, and that the gap grows, linearly of course, with experience. The narrow confidence bands imply statistically significant differences for almost every level of experience.

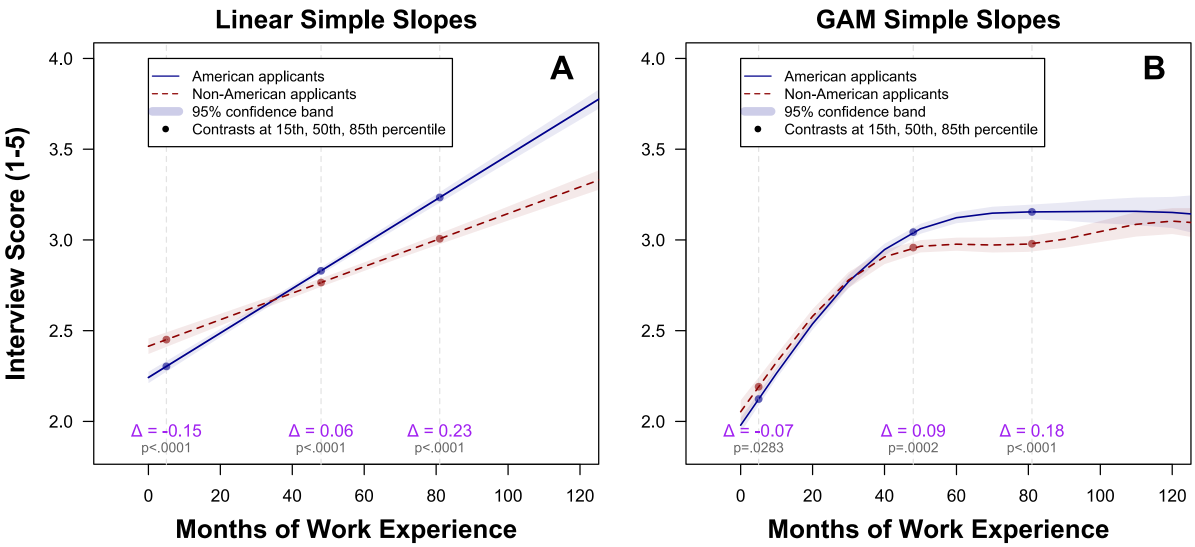


Fig. 4 *Linear vs GAM simple slopes reanalyzing data from Simonsohn & Gino (2013)*
N=11,740 interviews of applicants to an MBA program, rated on a 1-5 scale, predicted by work experience of applicant (M=44.9 months) and whether they are American (M=61%). Due to outliers in terms of experience, corresponding to possible coding errors, data were truncated at 200 months of experience. GAMs were estimated separately for American and non-American applicants: gam(y~s(x,k=5)). A Johnson-Neyman version of this figure, and robustness simple slopes plots with different cutoffs and k values is presented in Supplements 1 and 2 respectively.
R Code to reproduce figure: https://researchbox.org/1569.18 (use code **TYBYZK**).

To accompany the visual display of simple slopes, I report also statistical contrasts comparing the predicted y-value of the two plotted curves, for a few values in the x-axis. I use as defaults the median, and the 15th and 85th percentiles; the latter two correspond roughly to the mean ± 1SD for a normally distributed variable. We can think of those three contrasts as points in the Johnson and Neyman (1936) curve.[11] For instance, the first such contrast in Figure 4A indicates that the estimated effect of being an American applicant, among those with very little experience, is negative and highly significant in the linear model, and much smaller and barely significant in the GAM model.

While the three contrasts are of the same *sign* across the linear and GAM models, quantitatively the linear model's contrast are substantially larger (between 50% and 100% larger). Much more interesting in this case, however, is the qualitative comparison of the overall shape of the simple slopes.

[11] It is common to compute contrasts at the mean +-1SD (or 2SD). This can be misleading when a variable is not symmetrically distributed, for we may be focusing on very infrequent, on occasion even impossible values of x. Choosing contrasts based on quantiles alleviates the concern.

Most notably, the GAM model suggests that (1) scores do not seem to actually diverge across groups for applicants with more than 5 years of experience, and (2) for both groups, the benefit of experience plateaus rather than continues to increase at a constant rate constant throughout (plateauing earlier for non-American applicants, around 3 years of experience compared to 5 years for Americans). The picture that arises from the GAM simple slope is richer, and it is based on the data, rather than on our arbitrary assumption about the data. If we assume the impact of experience on interview scores is linear, our model will tell us the effect of experience on interview score is linear, but we are not really learning from the data (in this regard).

Figure 5 below provides a second example, reanalyzing data collected by Lawson and Kakkar (2021). Their core hypothesis was that "the sharing of fake news is largely driven by low conscientiousness *conservatives*" (abstract, italics added). Note that here the dependent variable is binary and thus the "linear" simple slopes is linear in the "generalized linear models" sense, where a logit regression is considered a linear model.

Interestingly, the GAM simple slopes allow for a more focused, and supportive in this case, evaluation of the prediction that the effect of interest is driven by conservatives. Looking again first at the contrast at the 15th quantile of the focal predictor, we see that while the linear model indicates a reversal for very liberal participants, the GAM model suggests such reversal may be spurious, arising from imposing linearity on all effects (like the heavier baby girls in Figure 1).

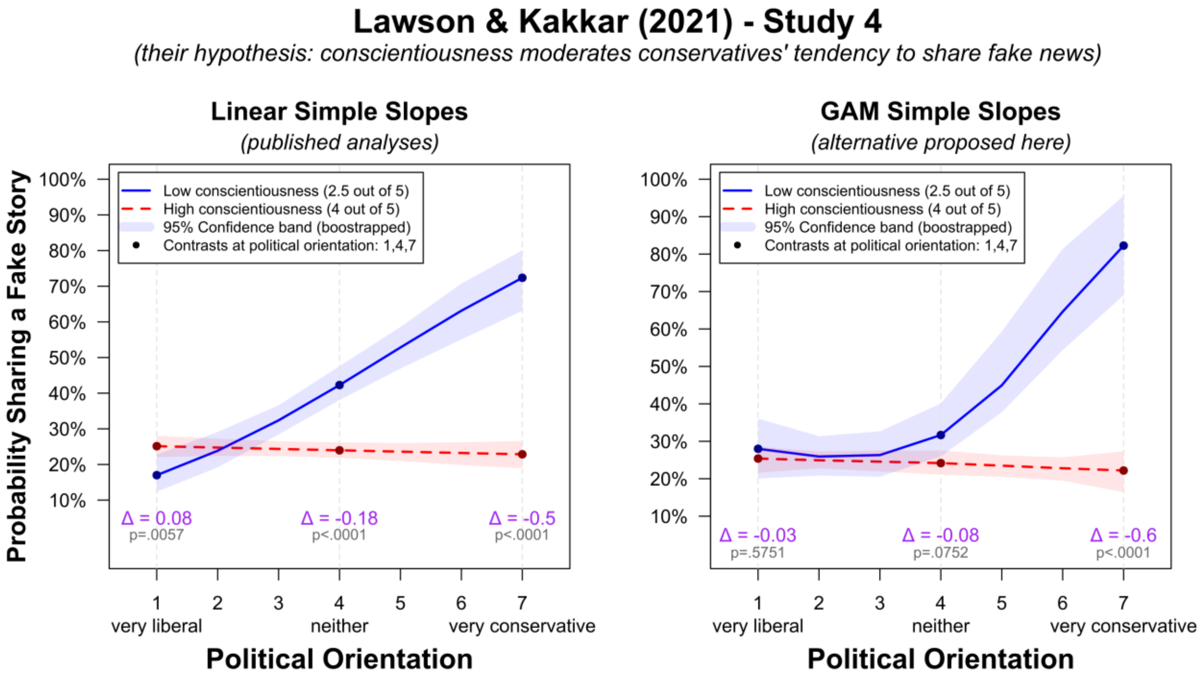


Fig. 5. *A hypothesis receiving stronger support with GAM simple slopes.*

Panel A reproduces the probed interaction reported by Lawson & Kakkar (see their Figure 5, but see footnote 12 here). It is generated by estimating a logistic regression on 967 participants, with 24 observations each (N=23208), and then plotting predicted probabilities to share news, with simple slopes computed at 2.5 and 4 of conscientiousness (see footnote for justification).[12] Panel B depicts equivalent calculations relying on a GAM model (gam(y~s(x,k=4), method='REML', family='binomial). Contrasts comparing the two simple slopes are reported at 1,4,7 in the political orientation scale (there are N=127, N=176 and N=74 respondents at each of those three buckets respectively). The confidence bands were computed via bootstrapping, by randomly resampling participants, rather than rows, to maintain the dependence across observations in the resamples. The confidence bands depict the values obtained in the 95% less extreme resamples.

R Code to reproduce figure: https://researchbox.org/1569.35 (use code **TYBYZK**).

Lin, Rand, and Pennycook (in press) have noted that Lawson & Kakkar's result should not be interpreted as supporting an effect on the sharing of *fake news* because low conscientiousness conservatives were more prone to sharing *in general*, not only fake news. They also report five conceptual replications of the studies by Lawson & Kakkar, which did not replicate the interaction. I have chosen to keep this example here, despite its apparent lack of robustness, because only upon constructing GAM Simple Slopes on the data collected by Lin et al, did I become convinced of their failure to replicate. It does therefore illustrate the practical use of GAM Simple Slopes: providing more informative descriptions of interaction results.

[12] The 15th and 85th percentile of conscientiousness in these data are 3 and 4.6 respectively. Using 3, the midpoint of the scale, for low conscientiousness seemed undesirable so I used 2.5, strictly below the midpoint. The original article by Lawson and Kakkar (2021) computed simple slopes at +-1SD of the moderator.

Having illustrated the similarities and differences between linear and GAM simple slopes, in the next subsection I rely on simulations to contrast their performance for probing interactions when at least one of the two variables in the interactions was randomly assigned (i.e., in experiments).

Probing Interactions in experiments (when at least one factor in *x·z* is randomly assigned)

False-Positive Rates for Probed Interactions

In this section I rely on simulations to evaluate the performance of three alternative tools that can be used to probe interactions: (1) linear simple slopes, (2) dichotomization (median split), and (3) GAM simple slopes.

The simulations in this section consider experiments, where treatment x is randomly assigned (x=1 or x=0), and the true association between x, z, and y, involves an attenuating interaction: the effect of *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. Focusing on attenuating interactions simplifies the reporting of results: I report how often each tool used to probe interactions falsely concludes the effect of *x* on *y* flips in sign with high enough zs, when in fact it never does.

In terms of the true associations, as is detailed in the caption for Figure 6, I considered two baseline scenarios: a linear effect with a ceiling, and concave smooth function. For each I created 36 variations with different sample sizes, functional forms, and distributions of the moderator variable. Each of the resulting 72 variations were used to run 5000 simulations, keeping track of how often a false-positive sign-reversal was detected.

For linear and GAM simple slopes, I consider a result to be false-positive when the effect of x is estimated as significantly negative, p<.05, when the moderator is at 85th percentile of its observed values (again, the true effect of x is never negative, no matter what value z takes). For the median split I consider a result to be false-positive when the interaction term is negative and p<.05.

Valid procedures have a false-positive rate, for $p \leq .05$, no greater than the nominal 5%. The poor performance of linear simple slopes depicted in Figure 6 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 does not in fact exist. 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.[13]

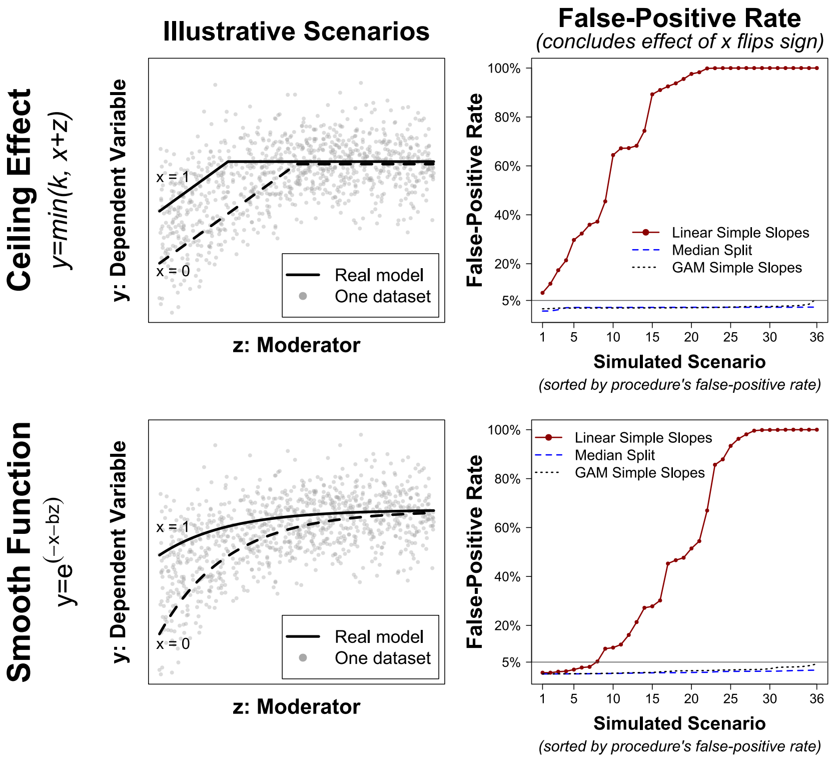


Fig. 6. *False-Positive rates for effect of x on y being negative for high z; true effect is never negative.*
The y-axes depict the percentage of 5000 simulations, run for each of 72 scenarios, where the probing of an interaction between x and *z* lead to an estimated negative effect of x on y, with $p < .05$, 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 by expected simulation error (deviation from a true rate of 5%) for 1s, 2nd… 36th most extreme post-hoc value of simulation error.
R Code to reproduce figure: https://researchbox.org/1569.25 (use code **TYBYZK**).

Going beyond *p*-values and false-positive rates (FPR) for comparing statistical approaches, note that if a procedure has an inflated FPR then we know its confidence interval will not have proper

[13] 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%.

coverage, thus deciding which model to estimate based on confidence interval performance, rather than false-positive rate, would also discourage relying on linear simple slopes. In terms of model-fit, the mean-squared-error for the linear model is, across the 72 scenarios, between 3 times and 13 times larger than for GAM. Because these are simulated data, we can compare fit in terms of the ground truth. That is, instead of asking how well a model fits the data, we can ask how close to the truth is each model. In addition to its intuitive appeal, this approach to assessing model fitness builds-in protection against overfitting. GAMs could have lower MSE by over-fitting the data, but they cannot get closer to ground truth by over-fitting the data. If a model is reading random error as signal, then it is going to get *further* from the ground truth. I thus also computed a "truth-MSE", the average squared error between each observations true y-value, and the fitted value based on the models. It is not close. Across the 72 scenarios, 'true-MSE' is between 10 times, and 382 times higher with the linear model.

Why do GAM Simple Slopes outperform?

To illustrate why GAM simple slopes outperform linear simple slopes, Figure 7 depicts results for one of 5000 simulations, for one of the 72 scenarios depicted in Figure 6. It illustrates how the arbitrary linearity assumption forces a spurious sign reversal for the estimated effect of x on y. This is again analogous to the heavier baby girls from Figure 1, and the apparently spurious reversal of the impact of conscientiousness for very liberal respondents in Figure 5.

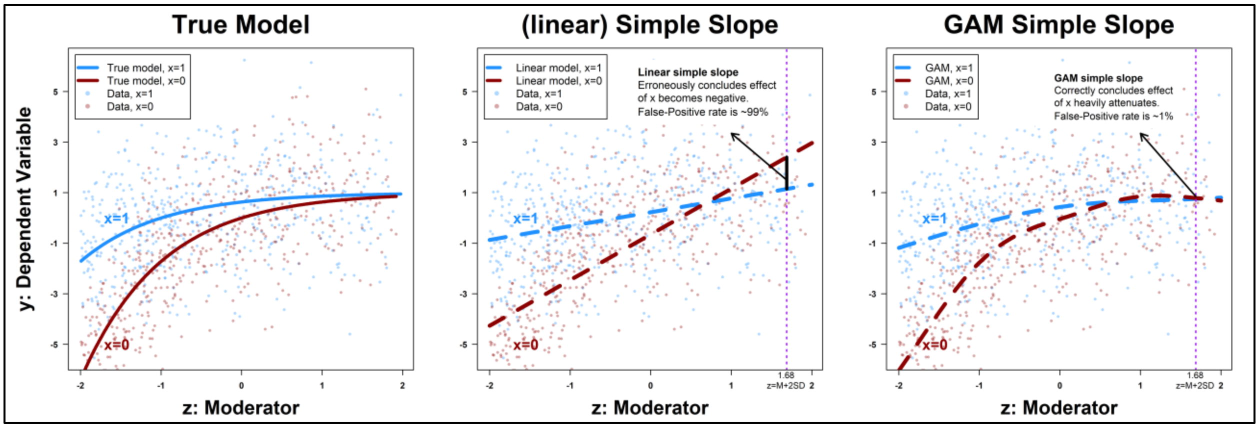


Fig. 7. *Example of simulated scenario in Figure 6, with high false-positive rate for linear model*
The three panels depict the same simulated 500 observations per condition (x=1 vs x=0). The true functional form is y=1-e$^{-(x+b\cdot z)}$.
R Code to reproduce figure: https://researchbox.org/1569.26 (use code **TYBYZK**).

Statistical power for probing interactions in experiments

Here I consider the power properties of GAM simple slopes for experimental data (where either x or z in *x·z* were randomly assigned); statistical power for non-experimental data is discussed in a later section. The false-positive rates of the linear simple slopes shown in the previous subsection seem sufficiently high to justify abandoning the approach even if it provided higher statistical power than the alternatives. But, interestingly, linear simple slopes can easily have *lower* statistical power than even the median split. To appreciate this, we do not need new simulations. Look back at the right panel of Figure 6. For high values of *z,* simple slopes estimate the effect of x as negative, in the most extreme cases with false-positive rates of 100%, thus the model has a 0% chance of detecting the actually positive effect.

In other words, while median splits have been justifiably criticized for decades for having lower power than the linear model to *test* interactions (Cohen, 1983), they can have greater power for *probing* those interactions. Just to be safe let me state this again. The long standing claim that dichotomization lowers power for testing an interaction is correct. What does not follow from this true fact, however, is that a linear simple slope is a more powerful approach for *probing* an interaction. By avoiding specification error, the median split can outperform linear models when probing interactions (again, in experiments, when one factor in *x·z* is randomly assigned). Nevertheless, it is generally the case that the median split will have lower power for probing interactions than will GAM simple slopes. Unless one is unable to implement a GAM simple slope, then, a median split is not the best option for probing interactions.

It is interesting to consider, as a boundary case, the unlikely scenario where the true model is perfectly linear. How much less power do GAM simple slopes have, to probe such interactions in experiments, compared to linear simple slopes? Figure 8 reports encouraging results. Across 16 scenarios, varying the distribution of the moderator, and the slopes involved in the linear interaction, GAM simple slopes achieves effectively the same level of precision/power as does the linear simple slopes. The intuition is that GAMs build in protection against over-fitting, and thus they report linear effects when true effects are linear. After paying a small penalty in power for the functional form flexibility*, we get linear model estimates with GAM when the true model is linear*.

In sum, switching from linear to GAM simple slopes to evaluate experimental results, would (1) eliminate the threat of possibly high false-positive rates when effects are not linear, (2) substantially increase power in some cases when the true effect is not linear, (3) improve the qualitative overall characterization of the relationship of interest, and (4) not meaningfully reduce power in the unlikely scenario of an actual linear relationship. There do not seem to be any reasons to continue relying on linear rather than GAM simple slopes.

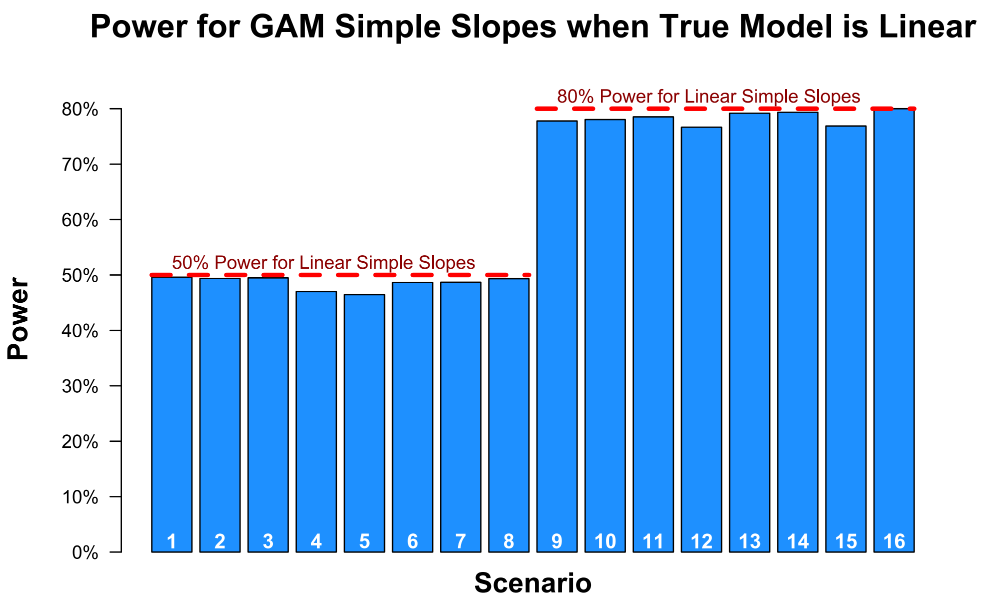


Fig. 8. *Relative power for linear- vs GAM Simple slopes.*
The barplots depict the statistical power, obtained by GAM simple slopes, testing the null that the effect of x is zero when the moderator z is at its 85th percentile. The simulated functional form is perfectly linear, and calibrated across scenarios to have 50% or 80% power for linear simple slopes. Across the 16 scenarios, the simulations vary the distribution of the moderator z (to be uniform, beta, or normal), the size of the interaction (true coefficient for *x·z*) and the sample size.
R Code to reproduce figure: https://researchbox.org/1569.30 (use code **TYBYZK**).

Testing Interactions with observational data (with measured rather manipulated variables)

The previous section covered the *probing* of interactions composed of predictors expected to be uncorrelated (e.g., where x or z in *x·z* were randomly assigned in an experiment). This section moves on to (i) *testing* rather than probing, and to (ii) interactions where x and z could be correlated (e.g., when both x and z are measured rather than manipulated).

The challenge curvilinear relationships pose to testing interactions, is a special case of the challenge omitted variables pose to testing regression coefficients more generally. When we omit a relevant variable from a regression, coefficients for included variables that correlate with the omitted variable can be biased.

Testing interactions through a model that incorrectly assumes the effects of *x* and *z* on *y* are linear, is equivalent to *omitting the non-linear portions* of the effects of x and z from the regression. This highlights the key role that the association between x and z plays on the validity of the interaction term. In experiments, where x or z are randomly assigned, any omitted non-linear effects of x or z are expected to be *un*correlated with the interaction x·z, and thus incorrectly assuming linearity does not actually invalidate the testing of the interactions in experiments. This is why the proposed toolbox (see Table 1) indicates that it is OK to test interactions in experiments with linear models. It's the same intuition for why it is OK to analyze experiment without controlling for covariates; the variables we are omitting will not bias our estimates of the effect of a *randomly assigned* treatment.[14]

In observational data in contrast, we typically expect (most) variables to be correlated (Meehl, 1990), especially pairs of variables that are not chosen arbitrarily, but rather, that are chosen because it is believed that both are associated with the same dependent variable. With observational data, then, we expect the omitted non-linearities of x and z to correlate with *x·z.* In other words, when variables are measured rather than manipulated, incorrectly assuming linearity introduces bias in the interaction term (see e.g., Ganzach, 1997). The next subsection provides an example with data from a published paper.

Example of an invalid interaction test in data from a published paper.

Preacher et al. (2006) probably constitutes the most cited peer-reviewed 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, from the National Longitudinal

[14] In terms of interactions being uncorrelated with omitted nonlinear terms with experimental data, the issue is subtler than it may seem at first. If x is a random and independent 0,1 variable, and z is a continuous variable, the interaction *x·z* will typically be actually highly correlated with any omitted nonlinear terms of z. But what matters for bias is the *partial* correlation, accounting for other predictors in the regression. Because the correlation between *x·z* and omitted nonlinear z terms is mediated by the z in *x·z*, controlling for z eliminates such correlation. For example, imagine the true model is $y=x+z^3$, but we estimate $y=\mathbf{a}x+\mathbf{b}z+\mathbf{c}x\cdot z$. While the omitted term z^3 can be highly correlated with *x·z*, the correlation is through the linear term z in *x·z*, and because z is also in the regression, the partial correlation of *x·z* with z^3, *controlling for z*, is expected to be zero, and thus *x·z* is expected to be unbiased.

Survey of Youth (NLSY), involves N=956 children as the unit of analysis, performance on a math test as the dependent variable, *y*, and measures of children's antisocial tendencies, *x*, and hyperactivity, *z,* as the key predictors. Preacher et al used this dataset to illustrate the use of (linear) simple slopes. Here, in contrast, I am not interested in probing the interaction, but on testing it. Despite being a tutorial on the interpretation of interactions, Preacher et al. do not discuss the issue of interest here, the invalidating impact of correlated non-linear predictors.

Estimating the linear model in their paper, I perfectly reproduced the reported point estimate and *p*-value for the focal interaction (b = - .3977, *p* = .0055).[15] Figure 9, however, shows that for this dataset, we should expect, and actually observe, the linear model having an inflated false-positive rate for the interaction.

Panel I shows that at least one predictor in the interaction has a likely non-linear effect, and Panel III that both predictors are correlated. As explained earlier, correlated non-linear predictors bias the interaction term, which will typically raise the false-positive rate (FPR). Panel IV reports that the estimated false-positive rates are indeed well above the nominal 5%. I computed FPR by simulating data under the null, forcing the absence of an interaction, and assessing how often the linear model obtained a statistically significant interaction. See Supplement 4 for details.

[15] 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.

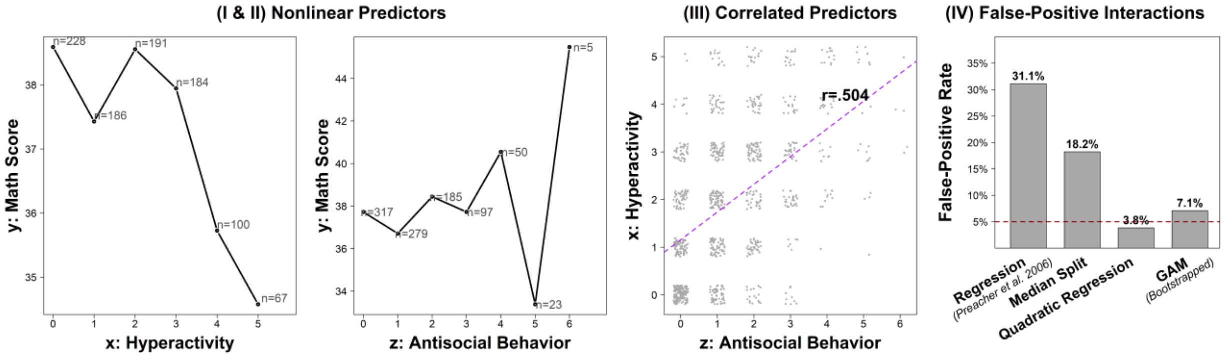


Fig 9. *Correlated non-linear predictors in Preacher et al. (2006) invalidate their interaction results*
Panels I & II show mean values of the dependent variable for each possible value of the predictors. Panel III a scatterplot and best fitting regression line for the predictors.
R Code to reproduce figure: https://researchbox.org/1569.33 (use code **TYBYZK**).

It is important to make clear that these elevated false-positive rates do not imply that Preacher et al.'s conclusions from the data are wrong; what they show is that their statistical analysis is not valid a-priori. Sometimes, a-posteriori, an invalid tool arrives at the same answer as a valid one. My goal for this example is not to challenge their conclusions, but to demonstrate that (1) real datasets do exhibit the problem, and (2) competent (even the most expert) researchers have generally ignored the problem.

Having provided the intuition for the problem correlated non-linear predictors pose for testing *x·z* interactions, and demonstrated the problem with data from a published paper, in the next subsection I explore the performance of alternative tools for testing for interactions in the presence of correlated non-linear predictors.

Simulations and the False-Positive Rates Testing Interactions Between Correlated Non-linear Predictors

The goal of the simulations reported in this subsection is to assess the performance of the alternative tools for testing interactions in a very broad range of scenarios involving non-linear effects. The alternative tools considered, which I contrast to the traditional approach of testing the coefficient of *x·z* in the linear regression model, involve (i) dichotomizing the moderator, (ii) adding quadratic x^2 and z^2

as covariates, and (iii) estimating a GAM instead of a linear model. Because I find that the GAM model *p*-values are incorrect (I rely on the R package 'mgcv' that comes bundled with R), I also report results for GAM models with bootstrapped *p*-values.[16,17] As a preview of the results, this latter approach, *bootstrapped* GAM, is the only one that performs adequately in all scenarios considered.

To avoid stumbling on a simulated scenario that by chance happens to make one tool work better than the other, I created 3840 scenarios through the exhaustive combination of several of the key operationalizations behind the simulated data. Figure 10 below shows stylized depictions of those operationalizations.

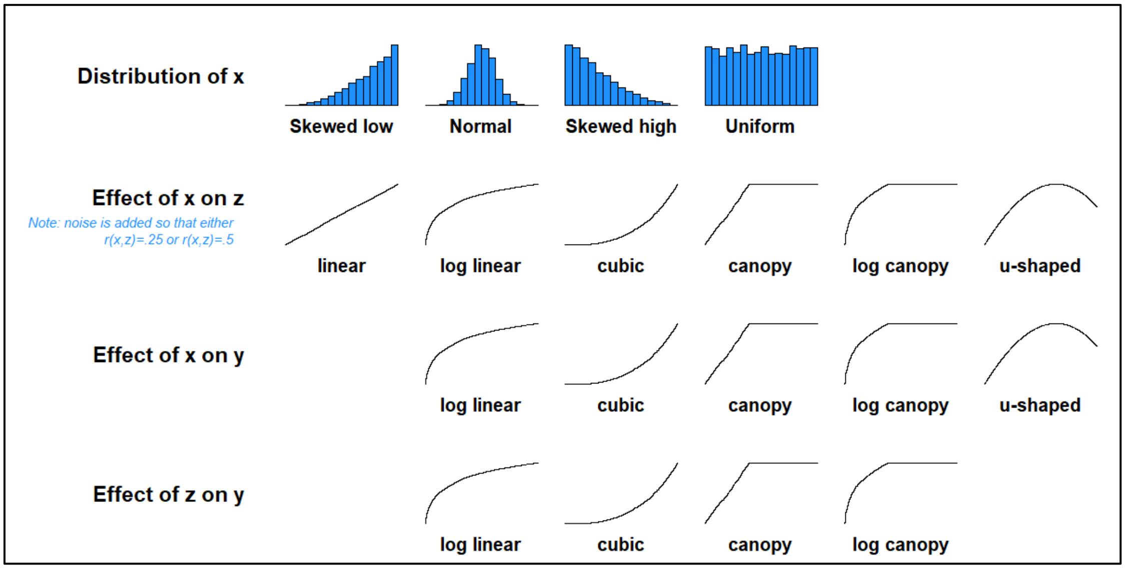


Fig. 10. *Operationalizations for computing false-positive rate of interaction with correlated predictors.* R Code to reproduce figure: https://researchbox.org/1569.37 (use code **TYBYZK**).

[16] The package 'mgcv' is a 'recommended' package, which means that the official distribution of R includes it (you can run 'library(mgcv)' without installing the package it). It is not a 'base' package, so it has its own version number, separate from R's, and can be updated within an R version; base packages like 'stas' or 'utils' may not be updated.

[17] Specifically I rely on what is known as the "Wild Bootstrap" (Davidson & Flachaire, 2008; Liu, 1988; Roodman, Nielsen, MacKinnon, & Webb, 2019), because it requires only assuming that residuals are symmetrically distributed around any one observation (i.e., that an observation with an observed residual of, say, +3.5, was just as likely ex-ante to have instead a residual of -3.5). When relying on the wild-bootstrap, bootstrapped samples are created by taking fitted values from the model, and then adding the observed residuals multiplied by a random variable. A few options for such random variable have been considered, the most intuitive of which is to multiply each residual by +1 or by -1 with a 50:50 probability.

For example, one of the 3840 scenarios involved x having a skewed-high distribution, while being correlated r=.5 with z through a log-linear relationship, where x has a cubic effect on y, and z has a log-canopy effect on y, studied with n=750 observation. In consideration of computing time, I randomly selected 300 of these 3840 scenarios and simulated 5000 datasets for each scenario. For each I tested for an interaction using the 5 aforementioned analytical tools. Because there is no interaction in any of the scenarios, all obtained p<.05 results are false-positive. A curvilinear robust tool for testing interactions should thus obtain p<.05 in about 5% simulations, for each of the 300 scenarios considered. The actual proportions of p<.05 for each tool are depicted in Figure 11.

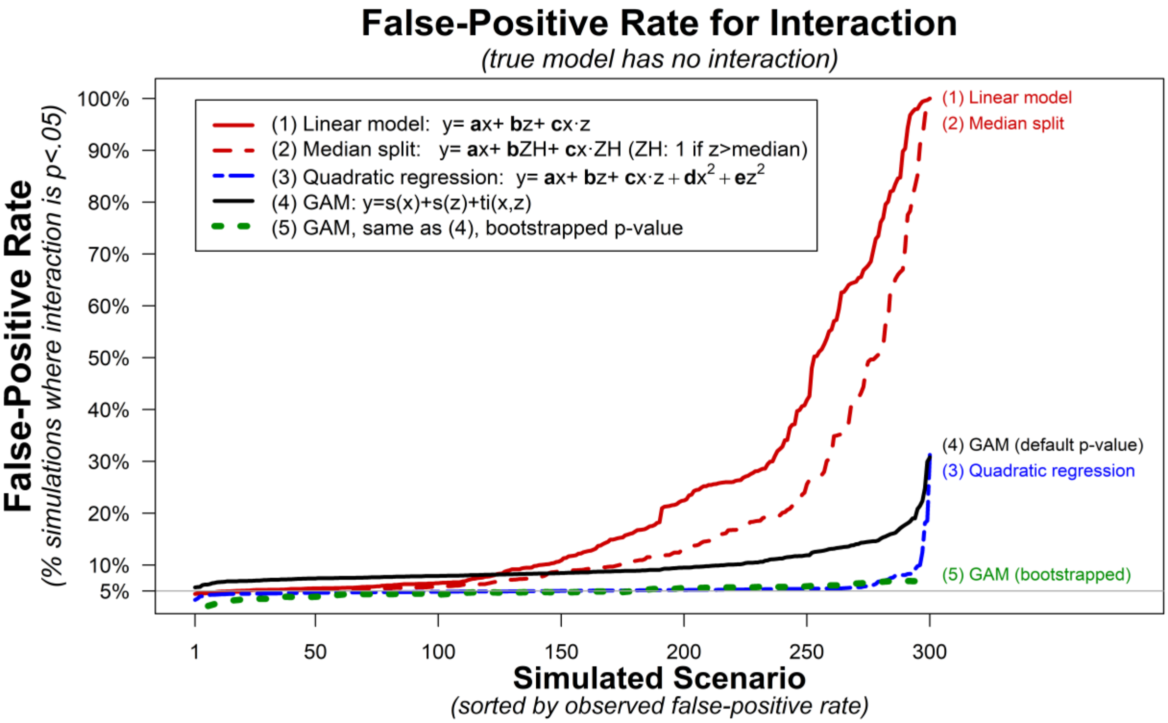


Fig. 11. *False-positive rates for interactions with non-linear and correlated x & z predictors*
The y-axes depict the percentage, out of 5000 simulations, for each scenario, where the interaction term obtained a statistically significant result ($p \leq .05$), despite the true interaction being zero. The 300 simulated scenarios are generated combining the operationalizations depicted in Figure 10 (a random subset of 300 out 3840 scenarios were run). The GAM model was estimated using R's 'recommended' package 'mgcv', with syntax: gam(y~s(x)+s(z)+ti(x,z)). The bootstrapped GAM *p*-value for the interaction smooth is obtained by first estimating a model without the interaction, gam(y~s(x)+s(z)) and, then using this 'null' model to generate 100 (wild) bootstrapped samples, adding to each predicted value, the observed residuals from that model, each multiplied with independently drawn at random 1 or -1. This is repeated 100 times, and the adjusted *p*-value is the share of these bootstrapped samples where the *p*-value for the interaction, in gam(y~s(x)+s(z)+ti(x,z)) is at least as low as that obtained in the observed data. Given computational costs, a sample of the 300 scenarios were re-ran with bootstrapping. Specifically, I re-ran the 20 scenarios with the highest FPR for the GAM model, and then every 10th scenario below the 280th, (so, the 270th highest FPR for the GAM interaction, the 250th highest, and so on). See Supplement 4 for more details.
R Code to reproduce figure: https://researchbox.org/1569.42 (use code **TYBYZK**).

First, we see that the linear model, and the median split, are strikingly invalid for the majority of scenarios considered. It is worth emphasizing that approximately all papers in social science that test interactions rely on one of these two tools. This does not mean, however, that all published interactions are false-positive. In fact, almost surely many are not.

Second, and surprisingly, the simple solution of merely adding x^2 and z^2 to the linear model (Cortina, 1993; Ganzach, 1997; Lubinski & Humphreys, 1990) achieves near nominal false-positive rates in the vast majority of scenarios considered, despite substantial specification error (in none of the models is *any* true relationship exactly quadratic). It is worth pointing out that while adding quadratic controls has been advocated for in several papers, this seems to be the first effort to systematically evaluate how such solution performs when the assumed functional form, quadratic effects of x and z, is not correct. While the idea of using quadratic controls is old, evidence that this is a good idea is new.

We do see a few specifications, however, where this approach suffers from markedly inflated false-positive rates, falsely rejecting the null over 25% of the time. It is ultimately an empirical and difficult to answer question whether functional forms in the real datasets analyzed by social scientists, 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. For what is worth, returning to the Preacher et al. example discussed in the previous section, adding quadratic controls leads to the nominal 5% false-positive rate for the linear model. Returning to the simulation results from Figure 12, we see that while the *p*-value from the GAM model over-rejects the null, the bootstrapped *p*-value performs well for all scenarios considered, though it does go over 1 to 2 percentage points above 5% for some scenarios.

Note that in the absence of the bootstrapping correction, the quadratic solution outperforms GAMs. I return to the relevance of this poor performance of *p*-values generated by the GAM procedure in the general discussion

Simulations and Statistical Power for Testing Interactions with Correlated Predictors

The previous subsection demonstrated the superior performance of GAM over linear models, in terms of false-positive rates, when the true functional form is not linear. A relevant question is the price paid in terms of power to achieve this lower Type 1 error rate. As before, I consider the corner case when the true model is linear, but I also consider the case when the true model is not. See Figure 12.

We see that when x and z in *x·z* are uncorrelated, there is essentially no difference in power across the four procedures. The more correlated x and z are, the higher the consequences of misspecification (as mentioned before, this is because the omitted non-linear portion of an effect is then correlated with the interaction and thus biases it) . Overall, we see that while there are scenarios where switching to GAM imposes power losses, these tend to be small losses. We also see that in more realistic scenarios, where functional form is not perfectly linear GAM offers higher power than the current tools in the social scientist toolbox do, possibly much higher power. It's important to note that in some of the scenarios in these figures, the linear model has very high "negative" power, a high probability of obtaining p<.05 for an effect of the wrong sign. It is difficult, and perhaps impossible, to be confident about functional form in social science, and when functional form is uncertain, GAM offers better expected performance in terms of both Type 1 and Type 2 errors.

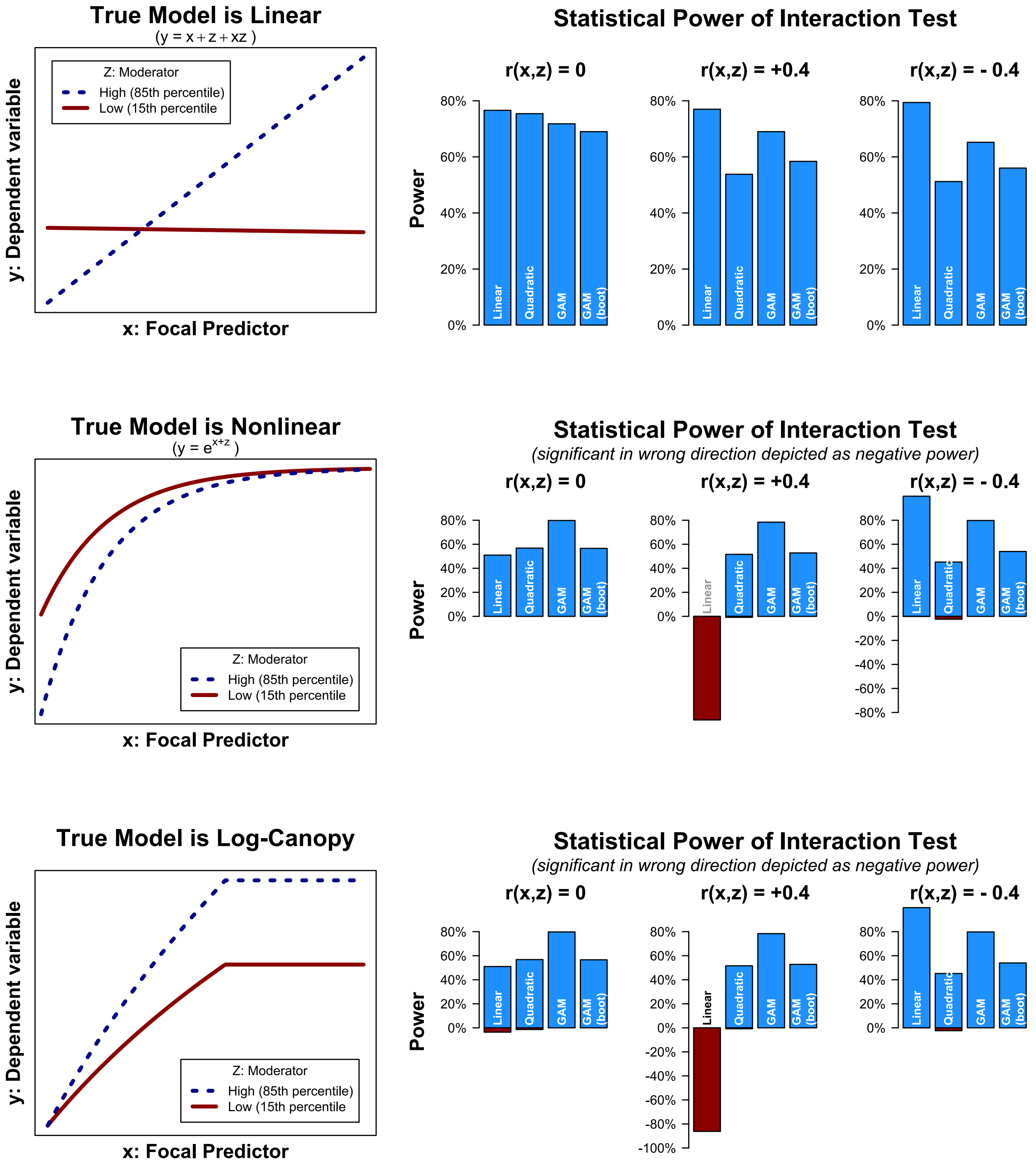


Fig 12. Statistical power to detect an interaction, when true functional form is and isn't linear

Each row reports estimated power for one scenario, depitcted in the first column, varying the correlation between the factors in *x·z*. Every simulation has n=500 observations, x and z are normally distributed. Negative power indicates the probability of obtaining a p<.05 results for the interaction with the wrong sign.

GAM's limitations

Having advocated for GAMs through most of this article, in this section I discuss what I consider to be its four main limitations.

Limitation #1: Interpretability

A commonly raised shortcoming for GAMs is that their black-box nature makes them "somewhat less interpretable than linear regression" (G. James, Witten, Hastie, & Tibshirani, 2021, p. 26). This shortcoming, however, is not too difficult to circumvent, by expressing GAM results in familiar forms, such as GAM simple slopes. Interactions even in linear models are actually difficult to interpret, which is precisely why we probe them, computing linear simple slopes. As I have shown throughout the article, GAM simple slopes are easy to produce, and just as easy to interpret as linear ones.

In addition, it seems odd to give any weight to interpretation ease, when the alternative, the easier to interpret result is simply wrong. Imagine a choice between two watches. A digital watch with large and easy to read numerals, but broken, being permanently stuck at an easy to read "3:45:00 PM", vs an analog watch without any numerals, but always showing the correct time of day. We would not give weight to interpretation-ease when choosing among these watches.

Limitation #2: Specification ambiguity.

Linear regressions have relatively few options in term of implementation, concerning primarily how standard errors are computed (e.g., relying on robust, clustered, or homoskedasticity-assuming standard errors). GAMs, in contrast, have many options, from the estimation procedure (e.g. "REML" vs "GCV"), to the penalty for over-fitting, to whether the number of knots is pre-set or estimated, to the flexibility (number of base functions) behind any particular smooth. Software that implements GAMs does make default choices for all of these, but those defaults may change over time and differ across statistical packages. Moreover, any user can opt-out from these defaults when analyzing any given dataset. This specification ambiguity poses two main challenges, one for authors and one for readers.

In terms of authors, they must somehow make all those decisions, and they may not have a priori basis or sufficient understanding to do so (in fact, probably most GAM users lack both, especially if GAMs become popular as quickly as I hope). In terms of readers, simply reading from a paper that a GAM is behind a particular result is not enough to know just what did the authors do. This challenge of specification ambiguity, which impacts authors and readers, does seem important, but manageable.

First, default values do tend to be sensible and are not often consequential, absent additional information, using the default is not a bad strategy for most authors in most situations. For example, a reviewer asked that I change the estimation procedure for all calculations in this paper, from the current default in the R package, "GCV", to an alternative that may become the default in the future, "REML". I did, and not a single figure or result was perceivably impacted by this choice. Second, for transparency sake, when a researcher deviates from default settings, it seems advisable to report results for a few alternative settings, ensuring results do not hinge on a specific and possibly arbitrary choice, or that if they do hinge, that this fact is shared with readers (e.g., "we find an interaction but only when using REML, not GCV, this may be because …"). Third, in terms of reproducibility, papers that rely on GAM estimation should include in the main text the exact specification ran. For example, a footnote that reads "Using the mgcv package v1.8-41, we estimated the model gam(y~s(x)+(z)+ti(x,z))".

Before moving to the next limitation, let's note that specification ambiguity is not unique to GAMs. While linear regressions don't have much ambiguity, many other methods already popular in social science do, including 'mixed models', structural equation modeling, factor analysis, meta-analyses, etc. Specification ambiguity is a good reason to show robustness or explicitly justify choices; it is not a good reason to avoid a tool altogether.

Limitation #3: Wrong p-values

The third GAM limitation I consider is that its *p*-values are often simply wrong, exhibiting much larger than nominal false-positive rates (see e.g., figure 11). This problem appears to be larger when predictors in the GAM model are correlated. This problem, moreover, does not seem to be widely known or appreciated by GAM advocates and users. In this paper, however I do provide a promising solution to the inferential problem with GAMs: bootstrapping-under-the-null. In the paper I applied the same bootstrapping approach across all examples and simulations. It would seem worthwhile for statisticians to make progress understanding what is causing the problem with GAM's *p*-values, and what other (possibly superior) solutions exist. In Supplement 4 I compare the implementation of bootstrapping I relied on in this paper, to 9 alternatives I considered.

Limitation 4: Accessibility

Despite being more than 40 years old, GAMs are not universally accessible. GAMs are quite accessible to R users; one of the packages that implements GAMs, 'mgcv', comes bundled with R (which is unusual, only about 20, out of over 15,000 CRAN packages do, highlighting that GAMs are not a niche tool). But GAMs are indeed less accessible in other software used by social scientists. STATA, does have a user-contributed GAM module, but it has not been updated in a couple of decades and does not seem to work with current operating systems.[18,19] GAMs are simply not available with more basic statistical tools like Excel, JASP, or SPSS (although SPSS users could rely on the R plugin to run mgcv).[20] I plan on creating an R package that will make creating GAM simple slopes a one-line-of-code job. But to rely on GAM for testing interactions in observational data, researchers will need to learn how to work with

[18] https://ideas.repec.org/c/boc/bocode/s428701.html
[19] https://www.statalist.org/forums/forum/general-stata-discussion/general/1507070-alternative-to-gam-module
[20] https://www.ibm.com/support/pages/does-ibm-spss-statistics-offer-generalized-additive-models-gams

GAMs, presumably relying on R or Python. This is in my mind GAMs' biggest shortcoming, I return to in the general discussion.

General Discussion

This article proposes that social scientists change how they test and probe interactions. The toolbox proposed in Table 1 constitutes an important departure from current practice. The strength of the case for each of the cells in Table 1, however, is not uniform.

The case is strongest for the top-right cell, for abandoning the linear probing of interactions in data from experiments. The case is strongest because going from linear to GAM simple slopes conveys virtually no cost. The statistical power loss of GAM simple slopes, when the true model is exactly linear, is negligible (see Figure 8), while the benefits in terms or reducing false-positive rates when the true model is not linear can be dramatic (see Figure 6); the benefits of a richer and more accurate understanding of the relationships we collect data to study is evident to the naked eye (see Figure 4 & 5). It is rare in statistics to have a tool that strictly dominates current practice, especially when current practice has existed for nearly a century, but that is the case, for examining interactions with a randomly assigned variable, when it comes to choosing GAM simple slopes over linear simple slopes.

With observational data, where neither x nor z were randomly assigned, and where we thus should not expect them to be uncorrelated, the tools proposed in Table 1 are also clear improvements over current practice, but they are not free of meaningful downsides. First, for testing interactions, no tool obtained strictly nominal false-positive rates for all scenarios (see Figure 12). The best performing tool, bootstrapped GAM, obtains false-positive rates around 7% in the worst scenarios considered. While 7% is higher than the nominal 5%, these excessive rejection rates, by 2 percentage points, pale in comparison to the >50% of false-positive rates obtained by the linear model, and median splits, for a large number of scenarios and to the 20-30% FPR in the data from Preacher et al. That is not, however, the only downside of bootstrapped GAM as a solution for testing interactions.

Estimating bootstrapped GAMs adds a few levels of complexity over current practice. First, we need to go from straightforward linear models to sophisticated GAMs. For simple designs, like those present in most experiments, this is a straightforward endeavor. But for more complex designs, say nested data, structural equation models, imputation of missing data, etc., the implementation may be harder, the documentation scarcer, and the evidence that those GAM estimates will work properly less well established. Moreover, the need to rely on bootstrapping poses a minimal challenge for simple designs (trivial, for example to build into a easy-to-use function in an R package), but implementing bootstrapping for more complex data structures requires thinking deeply about the data generating process and proceeding to create a custom bootstrap procedure, plus an at least intermediate level of comfort with programing. Realistically speaking, these technical steps are not accessible to all, perhaps not accessible to most, social scientists. Realistically speaking, then, bootstrapped GAM will hopefully be used for relatively simple data structures with observational data, but probably won't be used for more complex ones, at least not until further work documents, simplifies, and establishes the validity of bootstrapped GAMs for those types of situations.

For those situations a two-fold solution seems to be a promising way forward. First, as indicated in Table 1, simply adding quadratic terms for the factors in the $x{\cdot}z$ interaction, as was proposed by various authors around 30 years ago (Cortina, 1993; Ganzach, 1997; Lubinski & Humphreys, 1990), obtains a nominal false-positive rates in most, but not all cases considered. Second, the main challenge with relying on GAM for more complicated structures involves obtaining valid statistical inference, calibrated estimates of uncertainty (confidence intervals and *p*-values), but relying on GAM to obtain a qualitative description of the functional form should still work quite well in most cases, and surely better than the arbitrary linear model we have been using for over a century. In other words, when a data structure makes it difficult to implement the bootstrapped GAM solution, testing the interaction with the approach of choice to the analyst (SEM, mixed-model, regression with clustered errors, etc), while simply adding quadratic terms for x and z in the $x{\cdot}z$ interactions, and then probing a documented interaction in a descriptive fashion with GAM simple slopes without paying much attention to the probably too-tight confidence intervals, would seem to offer a vast improvement over current practice, while still leaving room for future work to further improve how we study interactions with observational data in social science.

References

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics letters, 80*(1), 123-129.

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

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

Cohen, J. (1983). The cost of dichotomization. *Applied psychological measurement, 7*(3), 249-253.

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.

Fechner, G. T. (1860). *Elemente der psychophysik* (Vol. 2): Breitkopf u. Härtel.

FitzGibbon, L., Komiya, A., & Murayama, K. (2021). The lure of counterfactual curiosity: people incur a cost to experience regret. *Psychological science, 32*(2), 241-255.

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

Greene, W. (2010). Testing hypotheses about interaction terms in nonlinear models. *Economics Letters, 107*(2), 291-296.

Greenlaw, S., & Shapiro, D. (2017). [eTextbook] Principles of Economics 2e. Retrieved from https://openstax.org/details/books/principles-economics-2e

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

James, E. L., Bonsall, M. B., Hoppitt, L., Tunbridge, E. M., Geddes, J. R., Milton, A. L., & Holmes, E. A. (2015). Computer game play reduces intrusive memories of experimental trauma via reconsolidation-update mechanisms. *Psychological science, 26*(8), 1201-1215.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R.

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

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica, 47*(2), 263-291.

Krantz, D. H., & Tversky, A. (1971). Conjoint-measurement analysis of composition rules in psychology. *Psychological review, 78*(2), 151.

Lawson, M. A., & Kakkar, H. (2021). Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. *Journal of Experimental Psychology: General*.

Lin, H., Rand, D., & Pennycook, G. (in press). Conscientiousness does not moderate the association between political ideology and susceptibility to fake news sharing. *Journal of Experimental Psychology: General*.

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.

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.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(1), 195-244.

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.

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the “cost” of learning, not cognitive decline. *Psychological science, 28*(8), 1171-1179.

Simpson, G. (2021). Using random effects in GAMs with mgcv. Retrieved from https://web.archive.org/web/20220428213452/https://fromthebottomoftheheap.net/2021/02/02/random-effects-in-gams/

Spiller, S. A., Fitzsimons, G. J., Lynch, 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.

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. N. (2013). A simple test for random effects in regression models. *Biometrika, 100*(4), 1005-1010.

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