First draft: December, 2016
This draft: May 17th, 2017

Two-Lines: The First Valid Test of U-Shaped Relationships

Uri Simonsohn
The Wharton School
University of Pennsylvania
uws@wharton.upenn.edu

Abstract:
Many psychological theories predict u-shaped relationships: x is good in low quantities, but bad in high quantities, or vice-versa. These predictions are tested primarily via quadratic regressions; here I show such approach is essentially never valid. I introduce a new test: estimating a regression with two separate lines, one for 'low' and one for 'high' values of x. A u-shape is present if the two slopes have opposite sign and are individually significant. A procedure to set the breakpoint for 'low' vs 'high' is proposed, the Robin Hood algorithm, and shown to be superior to all alternatives considered, including maximizing fit. The problems with the quadratic and advantages of the two-lines are demonstrated via simulations and examples from published psychological research.

Is there such thing as too many options, virtues, or examples in an opening sentence? Researchers are often interested in these types of questions, in assessing if the effect of x on y is positive for low-values of x, but negative for high values of x. For ease of exposition, I refer to all such relationships as 'u-shaped,' whether they are symmetric or not (i.e., U or J shaped), and whether the effect of x on y goes from negative to positive or vice versa (i.e., U or inverted-U).

U-shaped hypotheses are abundant in psychology. Just in 2016, for instance, papers published online in *JEP:G* examined possibly u-shaped relationships between strength of stimuli and priming effects (Payne, Brown-Iannuzzi, & Loersch, 2016) and between proficiency and cognitive performance (von Bastian, Souza, & Gade, 2016), in *Psychological Science* between trial-number and performance (Choi & Kirkorian, 2016), and between precision of opening bids and counter-offers (Loschelder, Friese, Schaerer, & Galinsky, 2016); in *JPSP* between age and materialism (Jaspers & Pieters, 2016) and between age and risk-taking (Josef et al., 2016); and in the *Journal of Applied Psychology* between team tenure and psychological safety (Koopmann, Lanaj, Wang, Zhou, & Shi, 2016), and between personality similarity and emotional display (Wilson, DeRue, Matta, Howe, & Conlon, 2016).

Psychologists, and other social scientists, have relied primarily on quadratic regressions, $y = ax+bx^2$, to test if the effect of x on y is u-shaped. In this article I show quadratic regression are a statistically invalid test of u-shapedness. Under realistic circumstances they can obtain extremely high false-positive rates; for instance, erroneously concluding with near certainty that $y=log(x)$ is u-shaped.

I provide concrete demonstrations of the problem by re-analyzing data from a few published psychology papers that appear to arrive at false-positive u-shaped relationships, interpreting what appear to be monotonic effects as u-shaped because they relied on quadratic regressions.

I propose an alternative test for u-shapedness: fitting two regression lines to the data, one for 'low' values of x, and another for 'high' values of x, and concluding that a u-shape is present if the slopes are of opposite sign and are both statistically significant. This two-lines approach follows, as statistical tests should, directly from the research question of interest. When testing for u-shapes we are asking this question: *Is the average effect of x on y of one sign up to a given value of x, and of the opposite sign, on average, from that point onwards*?

Linear regressions estimate the (weighted) average effect of x on y within a given range of x values; thus, to estimate two average effects of x on y for two different ranges of x values, estimating two regression lines is arguably the single most straightforward solution.[1] The two-lines approach requires determining the breakpoint, where one line ends and the other begins, or equivalently, the point at which x values are deemed 'low' vs. 'high.' (Once the breakpoint is set, both lines can be estimated within a single 'interrupted' regression for greater statistical efficiency). For hypothesis testing purposes, a natural objective is to set the breakpoint in a way that maximizes statistical power while preserving the nominal false-positive rate, e.g., α=5%. In other words, maximizing the probability of obtaining two regression lines of opposite sign, each $p<.05$, if the

[1] In particular, regression estimates are the weighted average of the slope of every pair of data points, weighting each pair by the square of the distance between predictor values. For instance, in the binary case: $\hat{b} = \sum_{i,j} \frac{y_i - y_j}{x_i - x_j} (x_i - x_j)^2 \,/\, \sum_{i,j} (x_i - x_j)^2$. See e.g., Gelman and Park (2008)

relationship truly is u-shaped, without exceeding 5% if it is not. I propose a novel procedure, the 'Robin Hood algorithm,' to set that breakpoint. It outperforms all alternatives considered, including, among others, setting the breakpoint that maximizes overall fit, and where a quadratic regression diagnoses that the effect changes sign.

It is important to make clear this article is not about nonlinear relationship in general, nor about polynomial regressions in general. It is about testing u-shaped relationships in particular. There is nothing wrong or invalid about including quadratic terms as covariates in regressions to allow for non-linear effects (e.g., controlling for age and age^2). There is something wrong, however, with interpreting the quadratic term as diagnostic of u-shapedness, as an answer to the question "does age at first increase y but then decrease it?" (Because it generally does not provide a valid answer to that question).

It is also worth distinguishing the testing of u-shapedness with the testing of monotonicity. A few tests have been proposed to test if the effect of x on y is monotonic (see e.g., Bowman, Jones, & Gijbels, 1998; Hall & Heckman, 2000). Monotonicity tests are not well suited to test for u-shapedness because, while all u-shaped effects are non-monotonic, not all non-monotonic effects are u-shaped.

To make this distinction concrete, consider two alternative hypotheses about the relationship between age and happiness. One hypothesis, H1, posits that age has a u-shaped relationship with happiness, such that people are happier the older they are up to a point, say 70 years old, and then happiness slowly starts decreasing with age. Another hypothesis, H2, posits that happiness increases with age up to death, but that around age 45 there is a short-lived mid-life crisis that leads to a short-lived drop in happiness. H1 and H2 involve a non-monotonic relationship, but H2 is not a u-shaped relationship.

Quadratic regressions don't test u-shapedness

The sophistication with which results from quadratic regressions are interpreted, for u-shape testing, can be classified into three levels; they differ in how many additional calculations are conducted upon estimation.

Level 1: is the quadratic term significant?

The most basic approach involves checking if the estimates of a and b, in $y=ax+bx^2$, imply a u-shaped function and $\hat{b}$ is statistically significant. This approach is advocated in some prominent textbooks. For example, Cohen, Cohen, West, and Aiken (2003) write, "The [quadratic] coefficient is negative [and significant]. . ., reflect[ing] the hypothesizes initial rise *followed by decline*" (p.198; emphasis added). The significant coefficient need not, in fact, imply a u-shape relationship.[2]

A JPSP paper by Simonton (1976), with about 150 citations, illustrates. He established correlates of the eminence of 'geniuses'. One key inference was that "*ranked eminence [is a] curvilinear inverted-U function of education*" (abstract, *p*. 218). The point estimates of interest, within a larger specification, were $y=4.872\ x\ -\ 11.96\ x^2$ where is the measure of eminence, and x of education (p. 223).

[2] In a later section Cohen et al do warn against blindly relying on the quadratic terms writing "it is always important to examine the actual data against both the polynomial regression and some nonparametric curve such as lowess." This advice is wise but seldom followed, probably because it is insufficiently actionable as comparisons of raw data with fitted lines is too qualitative. The two-lines test provides a practical solution when the question of interest is whether a relationship is u-shaped.

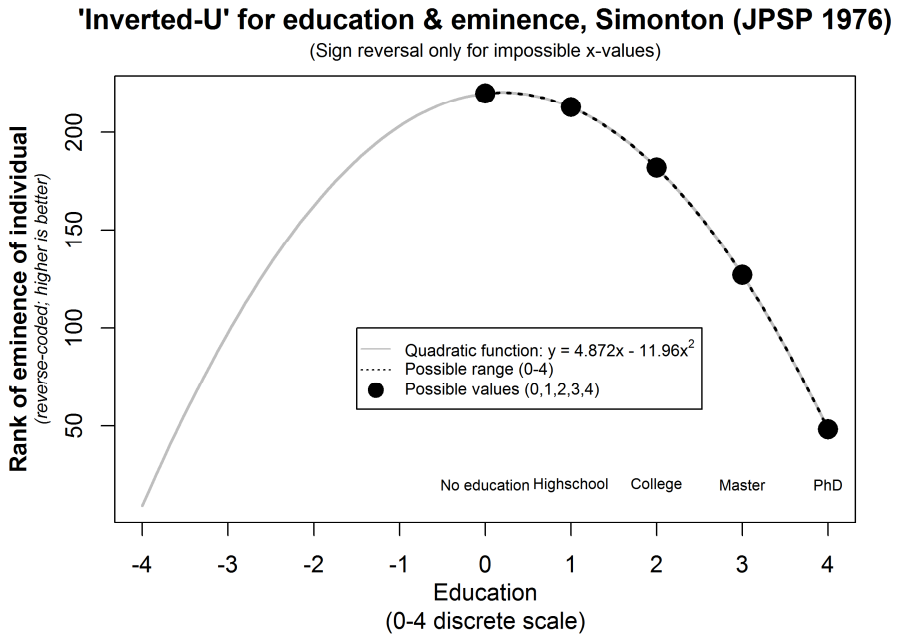

Fig. 1. *A significant quadratic term of opposite sign doesn't imply* u-shape.
R Code to reproduce figure: https://osf.io/kqjbn/

Figure 1 shows that, within the range of *possible* values the regression results do not imply a u-shape. For every possible value of x, higher x is associated with lower y. Only for negative (impossible) values of x is the sign positive and hence the overall pattern u-shaped.

Level 2: is the sign-flip within the range of values?

The discussion of Figure 1 above is an example of this additional step and some published papers carry it out as well. For example, Berman, Down, and Hill (2002) write "*the value [at which the sign flips] is actually above any value observed in the data, suggesting that, although negative returns are a theoretical possibility, they are not encountered*." (*p*. 23).

We need to take into account sampling error. The true relationship is, we assume, $y=ax+bx^2$, but we don't observe a and b, we observe estimates, $\hat{a}$ and $\hat{b}$, and as estimates they contain error, and so does, therefore, our estimate of the point at which the effect of x on y flips sign. We may thus obtain an estimated u-shape within the observed x-values because of sampling error.

Level 3: is the sign-flip "statistically-significantly" within the range of values?

The state-of-the-art procedure to test for u-shapes has been advocated in various methodological papers (Lind & Mehlum, 2010; Miller, Stromeyer, & Schwieterman, 2013; Preacher, Curran, & Bauer, 2006; Spiller, Fitzsimons, Lynch Jr, & McClelland, 2013). It builds on a technique dating back to Johnson and Neyman (1936).

This Johnson-Neyman technique, when applied to u-shape testing, estimates a quadratic regression and then builds a confidence interval for the estimated effects of x on y for different values of x. This approach more generally, beyond u-shape testing, is sometimes referred to as 'pick-a-point' or 'spotlight analysis' when focusing on few values of x, say ±1SD from the mean, and 'floodlight analysis' when applied to all values of x (Preacher et al., 2006; Spiller et al., 2013). With this floodlight analysis one concludes the effect of x on y is u-shaped if there are observed values of x where the estimated effect of x on y is *significantl*y positive, and also observed values of x where it is *significantl*y negative.

The problem is that the regression results hinge, as do the confidence interval calculations, on the assumption that the true relationship between x and y is quadratic and the results are not at all robust to deviations from such assumption, made purely for

mathematical convenience. Figure 2 provides realistic examples where the assumption is not met and the conclusions are erroneous.

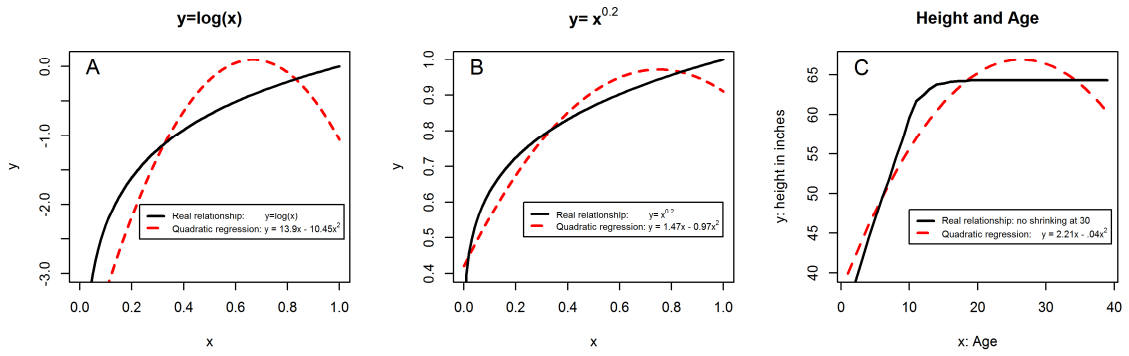

Fig. 2. *Examples of quadratic regressions misdiagnosing u-shapes.*
Panels **A** & **B** are obtained from a single simulated dataset with N=100000 where *x* is obtaining by squaring random draws from the U(0,1) distribution. Large samples without noise were used to convey the point that quadratic regressions get it wrong even asymptotically. Panel **C** uses data come from the Center for Diease Control.
R Code to reproduce figure https://osf.io/ywe79

For instance, panel A shows a scenario where the true relationship is y=log(x) and where a quadratic regression would result in $\hat{y}$ =13.96x-10.45x^2. In this equation, the effect of x on y is, dy/dx=13.96- 2*10.45x. When x=.25, the effect of x is positive, 8.735, but when x=.75, in contrast, negative, -1.71. Now, of course that result is wrong, the effect of x is never negative when y=log(x), but it is *estimated* as negative because we are incorrectly assuming the relationship is quadratic. Specification error is behind the erroneous conclusion. Panels B & C provide additional examples.

Assuming a quadratic relationship may also lead to false-negatives, failing to diagnose a u-shape relationship that is present, even with an *infinite* sample size. This will occur when the true relationship is u-shaped but deviates sufficiently from the quadratic. See Figure 3.

The intuition for the poor performance of the quadratic regression in Figures 2 & 3 is that it minimizes the sum of squared errors, ($\hat{y}$ –y)^2, without taking into

account overall-shape. If obtaining a better fit requires outputting a quadratic function that generates a non-existent u-shape, or missing a real u-shape, there is no 'penalty.'

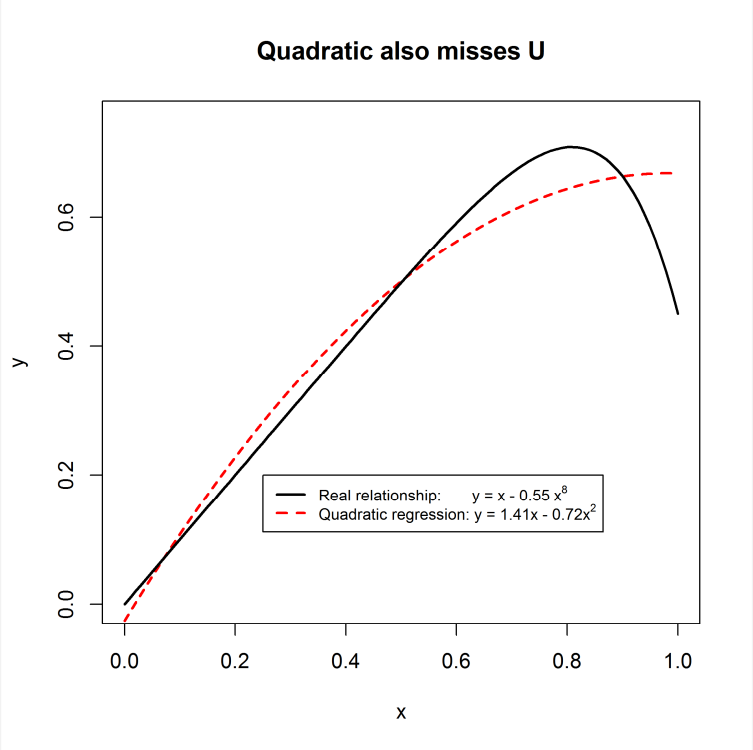

Fig. 3. *Example of quadratic regression false-negatively concluding u-shape is absent.* A single dataset with N=100000 observations was created, x was generated by drawing at random from U(0,1) distribution and squaring the result. R Code to reproduce figure https://osf.io/ywe79

What about diagnostic tests?

Textbooks indicate researchers should conduct diagnostic tests before interpreting their regression results, are those enough to protect us from wrong inferences about u-shapes based on quadratic regressions? I argue below the answer is no.

First, in practice, researchers do not run or at least report diagnostics on their regressions. There currently are 1021 *Journal of Experimental Psychology: General* papers containing the word "regression", only 2 the words "qq plot" or "q-q plot",

perhaps the most well-known diagnostic plot.[3] None of the 8 u-shape papers published in 2016 listed in the opening paragraphs reported diagnostic tests or plots.

Second, regression diagnostics qualitatively assess the general adequacy of the model, but we want to quantitatively assess the adequacy of the conclusion *about u-shapedness*. Figure 4 illustrates the difference. It depicts two true relationships that are not exactly quadratic. One is u-shaped, the other not. Applying the quadratic to either dataset we conclude the relationship is u-shaped, wrongly in the left column, and correctly in the right one. Not only are the observed quadratic regression results indistinguishable, so are the diagnostic plots. Diagnostic plots are thus not diagnostic; they look the same when the u-shape conclusion is false- vs true-positive.

Third, even if we assumed that researchers were to start reporting diagnostics and assumed that diagnostic tests were diagnostic, it is not clear what researchers should do when they diagnose their quadratic regression as misspecified. If not a quadratic model, then what model should they estimate? There is no default alternative, researchers would need to try multiple functional forms until one seems to -subjectively- fit well enough.

Say running higher order polynomials, interrupted log regressions, various interactions, etc. This leads to two problems. First, when those more complicated models are estimated, how would the researcher go about testing for u-shapedness? Second, the modelling ambiguity opens the door to over-fitting in general and *p*-hacking in particular.

[3] The qq-plot is a graph that has the quantiles of a normal distribution in the x-axis and of the residuals in the y-axis. The fact that I felt it was useful to include this footnote reflects my beliefs about how common is the use of these tools.

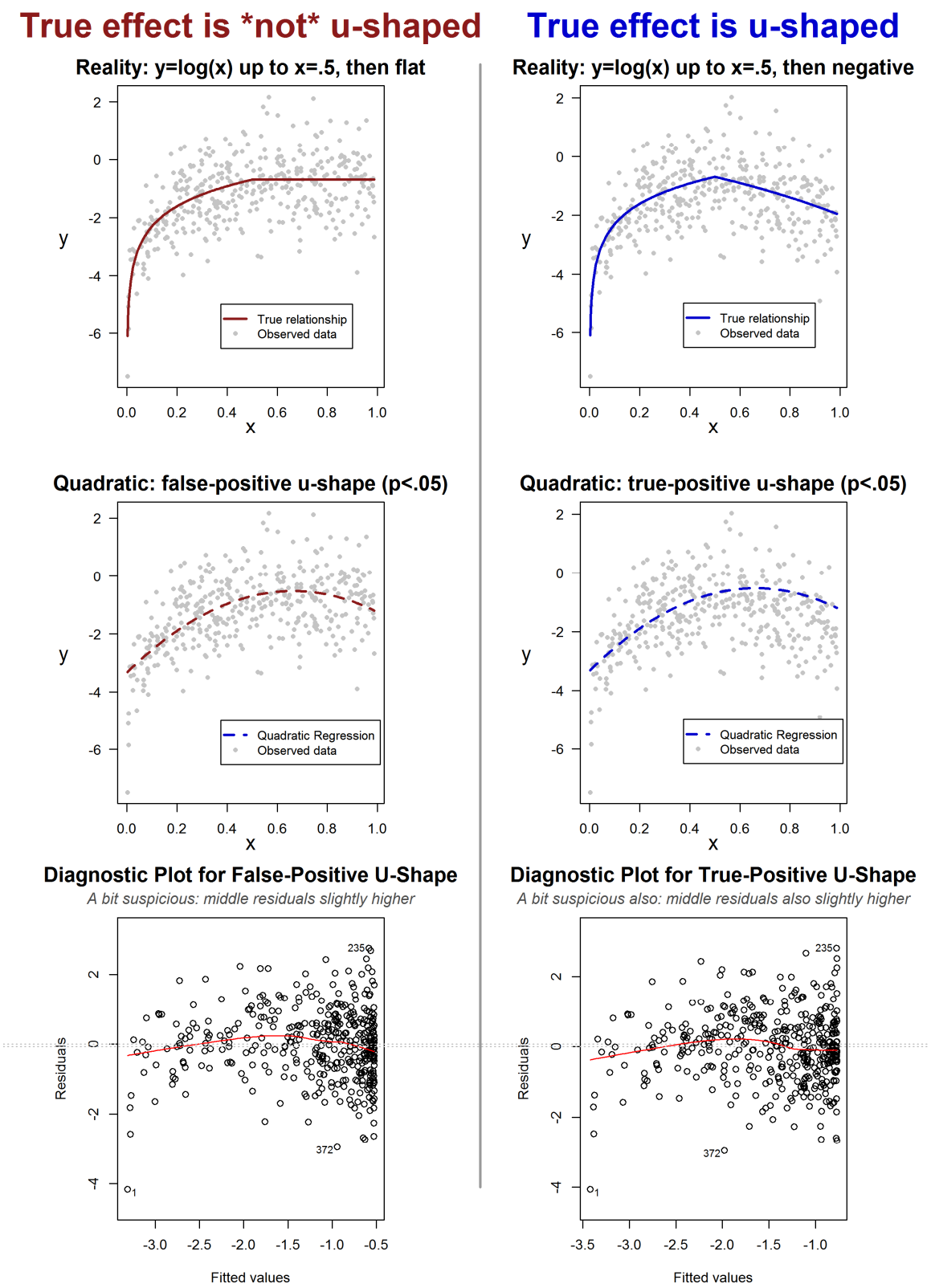

Fig. 4. *Diagnostic plots are not diagnostic about u-shapedness*

Notes: The data were generated by drawing 400 observations from U(0,1) for x and adding noise N(0,1) to the true y-value (see top-row for true model). Each column has the same dataset for the three charts, but they differ across columns.

R Code to reproduce figure: https://osf.io/cb7he

The two-line solution

Let's define an observed relationship between x and y, y=f(x), as u-shaped, if there exist an x value within the observed data, x_c, such that the average effect of x on y is of opposite sign for $x \leq x_c$ vs $x \geq x_c$. To test for the presence of a u-shaped relationship, therefore, we need to compute two average effects of x on y, one for $x \leq x_c$ and one for $x \geq x_c$. Considering that linear regressions provide unbiased estimators of the weighted average effect in the given range of x values, whether the underlying relationship is linear or not (see footnote 2), a natural approach to test for a u-shaped relationship consists of estimating a regression for x values below a breakpoint, $x \leq x_c$, and another for values above it, $x \geq x_c$, obtaining the two desired average effect estimates. One may increase statistical efficiency by simultaneously estimating both lines in a single regression, relying on what is often referred to as an *interrupted* regression (see e.g., Marsh & Cormier, 2001, p. 7). Specifically, interrupted regressions conform to the following general formulation: [4]

$$y = a + b\, x_{low} + c\, x_{high} + d\, high + Z\, B_z \qquad (1)$$

Where:

x_{low}=x-x_c if $x \leq x_c$, and 0 otherwise

x_{high}=x-x_c if $x \geq x_c$, and 0 otherwise

high=1 if $x \geq x_c$ 0, otherwise.

Z is the (optional) matrix with covariates, and B_z its vector of coefficients.

[4] If d is forced to be 0, thus not allowing a discontinuity at x_c, the regression is called *segmented* instead of *interrupted* (see e.g., Muggeo, 2003). As shown below, forcing d to zero can introduce bias onto both $\hat{b}$ and $\hat{c}$, therefore, for u-shape testing purposes, one must rely on interrupted rather than segmented regressions. One must include *high* as a predictor.

Discrete x values

Equation 1 involves weak inequalities, such that the breakpoint, x_c, is in both the high and low segment. For continuous x this is not a problem since there is zero mass an any given value, but for discrete x it is a problem, as one cannot estimate a regression where one observation belongs to two segments (and most observed data is, in practice, discrete, due to rounding). One solution is to arbitrarily assign x_c to one of the segments. A much better solution is to run two separate interrupted regressions, one were we include x_c in the low segment and one in the high segment.

It helps to consider the most extreme case, where x can only take three values (if it can take only two values, there is no room for estimating non-linear effects). So say $x=\{1,2,3\}$. We would test if a u-shape is present by assessing if the slope between x=1 and x=2 is of opposite sign as between x=2 and x=3. So $x_c=2$ would be included in both lines. To allow x_c to belong to both segment we create six instead of just three variables: First, including x_c in the low segment,

1) $x_{low,1}=x-x_c$ if $x \leq x_c$ and 0 otherwise (inequality)
2) $x_{high,1}=x-x_c$ if $x > x_c$ and 0 otherwise (strict inequality)
3) $high_1=1$ if $x > x_c$ 0 otherwise (strict inequality)

Then including x_c in the high segment

4) $x_{low,2}=x-x_c$ if $x < x_c$ and 0 otherwise (strict inequality)
5) $x_{high,2}=x-x_c$ if $x \geq x_c$ and 0 otherwise (inequality)
6) $high_2=1$ if $x \geq x_c$ 0 otherwise (inequality)

With these six new variables one then estimates two separate interrupted regression

The first is of the form $y= a_1 + b_1\, x_{low,1} + c_1\, x_{high,1} + d_1\, high_1 +$ covariates.

The second of the form $y= a_2 + b_2\, x_{low,2} + c_2\, x_{high,2} + d_2\, high_2 +$ covariates.

And we proceed to test for a u-shaped relationship by verifying that $\hat{b}_1$ and $\hat{c}_2$ have opposite sign and are both individually significant.

Setting the breakpoint

In situations where theory is rich enough to make predictions about where the effect should switch sign, it would be sensible to set x_c, the breakpoint, there (see e.g., Seidman, 2012; Ungemach, Stewart, & Reimers, 2011). This will be a rare occurrence in psychological research where theories tend to be insufficiently quantitative to make point predictions (none of the examples from the opening page, for instance, generate point predictions for the sign switch). Absent a theoretical a-priori breakpoint, it must be set based on the observed data. This, in turn, can be done seeking to maximize fit, *or,* seeking to maximize statistical power. That is, seeking to arrive at a model that fits the data best, *or* at one that has the highest probability of diagnosing that f(x) as u-shaped if indeed it is.

Maximizing fit. Setting the breakpoint to maximize fit involves answering this question: "Given that we will fit the data with two lines, which breakpoint leads to two lines that best fit the data overall?" In this case, however, we want to answer a different question: "If the true relationship where u-shaped, which breakpoint maximizes the chances we will detect such shape?" The distinction between these two questions is the distinction between fit and statistical power. Figure 5 provides two illustrative examples.

Panel A depicts a relationship that is formed by three straight lines. The first change in slope affects more observations and involves a bigger change, thus, a procedure that seeks to minimize squared errors will set the breakpoint there, but to detect the u-shape we need a breakpoint around x=.8. Panel B depicts a similar scenario with a less stylized function. To maximize power for the u-shape test we do not need to maximize fit. [5]

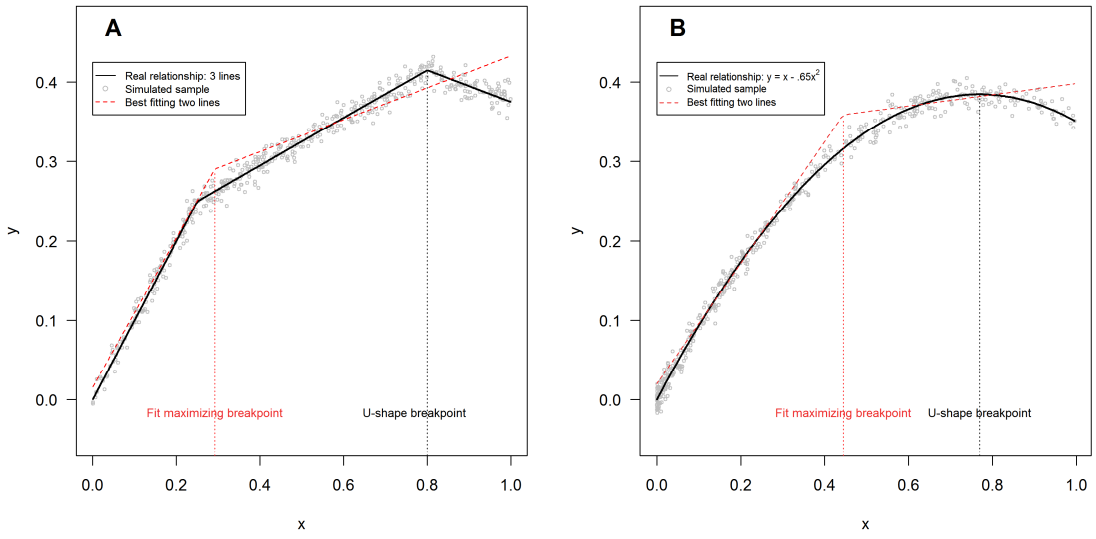

Fig 5. *The breakpoint that maximize overall fit does not necessarily maximize power to detect a u-shaped relationship.*
Note: the figure depicts the breakpoint for two regression lines that maximizes overall fit, using Muggeo (2003)'s procedure, and contrasts it with the breakpoint actually associated with the x-value at which the sign of the effect of x on y changes. R Code to reproduce figure https://osf.io/p2myj

Maximizing power. Without making strong assumptions about (a) the functional form of the relationship between x and y, f(x), (b) the distribution of x, and (c) the distribution of the error term, it does not seem possible to arrive at a theoretically optimal

[5] A tempting solution to the specific examples chosen in the figure is to allow *two* breakpoints, and thus three segments. There are two problems with this solution. First, three lines can also be misspecified, of course, and thus four, or more lines be needed to recover the u-shape when setting breakpoints based on fit. Second, even absent residual specification error, e.g., when the true functional form does indeed consists of three linear segments, the three-line estimation has almost always *lower* power to detect the u-shaped relationship than the two-line one. See Figure 9.

breakpoint that maximizes statistical power for u-shape testing. The approach I propose here, instead, is algorithmic, designed to have high power, rather than demonstrably maximal power, for a very broad range of situations. I developed the algorithm keeping in mind three key ideas: (i) because the two-lines test requires both slopes to be significant, to increase its power requires increasing the power of the statistically weaker of the two lines. Segments of an interrupted regression, in turn, have more power when (ii) they are steeper (bigger effect), and (iii) they include more observations within their segment (smaller standard error). Thus, conceptually, the algorithm seeks to set a breakpoint that will increase the statistical strength of the weaker of the two lines, by placing more observations in that segment, without overly attenuating its slope. I refer to it as the Robin Hood algorithm, for it takes away observations from the more powerful line and assigns them to the less powerful one.

I rely on Figure 6 to describe the Robin Hood algorithm. Every panel involves the same true underlying relationship between x and y, depicted by the solid line in Panel A, and the same single random sample, depicted with the same scatterplot in every panel. The top row illustrates increasingly more sophisticated approaches for setting the breakpoint, culminating in the proposed Robin Hood algorithm in the right-most column. The bottom row the resulting two-line regression estimates.

For illustrative purposes, consider attempting to obtain two steep slopes by setting x_c, the breakpoint, at the x value associated with the most extreme observed y value (first column in Figure 6). An obvious problem is that individual observations, especially the most extreme one, can be greatly influenced by random error. Panel A, for example

shows that the x value associated with the most extreme observation, x=.78, falls outside the range with maximum true y values, .5<x<.7.

We can cancel much of the aforementioned random error by estimating a flexible model of f(x), e.g., a polynomial, local, kernel, or spline, regression, and use the model's fitted values instead of the observed values to identify the most extreme observations.

I rely on splines here, because they easily accommodate covariates, can be used to construct confidence intervals for f(x), and do not rely on functional form assumptions (see section 3.2.1 in Wood, 2006).[6] In particular, Panel B depicts the fitted values, $\hat{y}$s, obtained from a cubic spline regression, and showcases the consequences of moving the breakpoint from the x associated with the most extreme observed y, to the x associated with the most extreme *fitted* value: $\hat{y}_{max}$.

In the example from Figure 6, and presumably in many psychological phenomena, relationships are U rather than V shaped, having regions with a relatively flat maximum. It seems therefore sensible to identify the *set* of most extreme $\hat{y}$s rather than the single most extreme $\hat{y}_{max}$. Here I define $\hat{y}$s within one standard error of $\hat{y}_{max}$ as that set, and refer to it as $\hat{y}_{flat}$. Thus, every $\hat{y}$ in $\hat{y}_{flat}$ is within one standard error of $\hat{y}_{max}$. The solid line in Figure 6C depicts $\hat{y}_{flat}$.

We now have a set of candidate x_c values, those associated with $\hat{y}_{flat}$. The goal is to choose the one among them that we expect to give higher statistical power to detect a u-shape, and thus the one among them that we expect to give higher statistical power to the weaker of the two lines within the interrupted regression. The algorithm pursuits that

[6] In particular, using the R library *mgcv*, the command gam(y~s(x,bs="cr")) estimates a cubic spline predicting the dependent variable y with the predictor x. The option bs="cr" specifies a cubic spline be used, instead of the default which is a "plate regression spline" (Wood, 2006, p. 219). The entire R Code used to generate Figure 6 is available here: https://osf.io/3fst7.

goal by setting x_c so that it allocates a disproportionate share of the observations in $\hat{y}_{flat}$ to the weaker line; by increasing the number of observations in that segment, it reduces its standard error, increasing its statistical power.

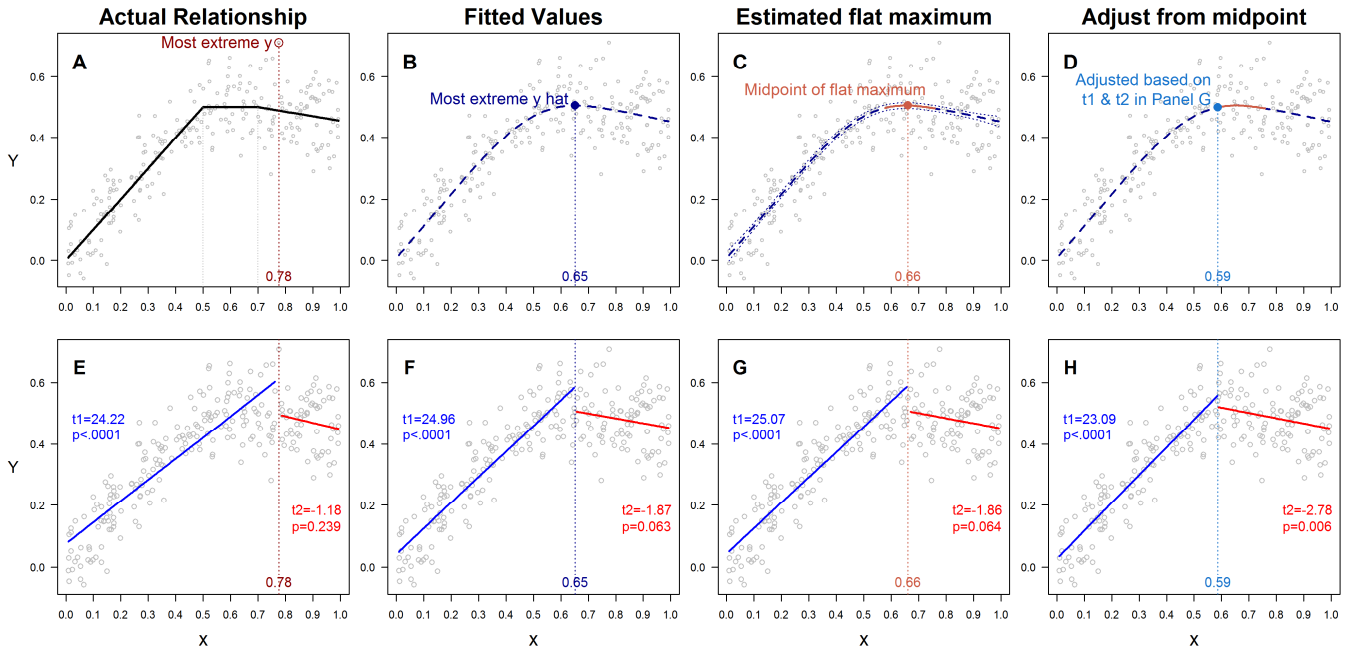

Fig 6. *Different procedures to identify the breakpoint, and their consequences.*
Notes: All panels are based on the same random sample (gray scatterplots) based on the true relationship between x and y, solid line in Panel A. The effect of x on y is positive up to x=.5, flat up to x=.7, and negative onwards. Top row shows 4 alternative ways to set the breakpoint, bottom row the resulting two-line regressions. Fitted values in panels B-D obtained by smoothing the scatterplot with a cubic spline. Flat region in C&D is where $\hat{y}$s are within 1 standard error of the max($\hat{y}$). R Code to reproduce figure: https://osf.io/3fst7

The algorithm proceeds in two steps. In the first step it identifies which of the two lines is statistically weaker. In the second step it sets the breakpoint by allocating observations in $\hat{y}_{flat}$ to the first vs second line in inverse proportion to their relative statistical strength. Specifically, in the first step the algorithm sets the x-value that is the midpoint of $\hat{y}_{flat}$ as an interim breakpoint. It estimates an interrupted regression and computes the (absolute value of the) test-statistics for both lines, t_1 and t_2, and then sets the breakpoint for the second step in inverse proportion to these ts. That is, the breakpoint becomes the $t_2/(t_1+t_2)$ percentile of the x-values within $\hat{y}_{flat}$.

To build an intuition: if both lines are about equally strong, statistically speaking, with roughly identical test statistics, the breakpoint will remain roughly at the midpoint of $\hat{y}_{flat}$. If the t-value of the first line in the first step were, say, 3 times that of the second line, then the breakpoint would be set at the 75th percentile of xs within $\hat{y}_{flat}$, so that the second (weaker) line has 75% of $\hat{y}_{flat}$ and the first line the remaining 25%. The intuition, again, is that the algorithm allocates additional observations from within the $\hat{y}_{flat}$ region to the weaker line so that its standard error gets smaller.

Returning to Figure 6. Panel G shows that first step, where the midpoint of $\hat{y}_{flat}$ is the breakpoint. It leads to t_1= 25.07 and t_2=1.86. Computing the ratio we obtain $t_2/(t_1+t_2)$=6.9% so the breakpoint is set at the 6.9th percentile of the x values associated with $\hat{y}_{flat}$, which in that sample corresponds to x=.59. Using that breakpoint we obtain the final interrupted regression used to test the presence of a u-shape, and in this case we obtain a much stronger result for the second slope, p_2=.006.

In sum, the Robin Hood algorithm consists of the following 5 steps.

1) Estimate a cubic spline for the relationship between x and y
2) Identify $\hat{y}_{max}$, the most extreme internal fitted value.
3) Identify $\hat{y}_{flat,}$ the set of $\hat{y}$ values within a standard error of $\hat{y}_{max}$
4) Estimate an interrupted regression using as the breakpoint the median x value within $\hat{y}_{flat}$. The regression will result in two test statistics, one for each line. Let their absolute values be t_1 and t_2
5) Set as the breakpoint at $t_2/(t_1+t_2)$ percentile of the x values associated with $\hat{y}_{flat.}$

R Code needed to run the two-lines test is available from https://osf.io/psfwz/. I have also created an online app to run the test without requiring any programming. http://webstimate.org/twolines (note to review team: this is an alpha stage, I'll continue developing as you review this paper).

Why not just the spline?

The proposed algorithm relies on a cubic spline, see step 1, why do we even need the other steps? Why estimate an obviously inferior model, two straight lines, if we already have a much better approximation of the true model? The answer goes back to the distinction between fitting data and testing hypotheses. A spline provides a better summary of the overall pattern, but does not lend itself to asking a stylized fact question about the relationship of interest: is it u-shaped? We need to assess if the *apparent* shape a spline suggests, is or is not consistent with the null hypothesis (no u-shape). Splines do not lend themselves to Gestaltic tests of this nature. Moreover, the visual results a spline provides hinge on the arbitrary choice of smoothness, penalty term, number of knots, etc. For the same data, some arbitrary choices will lead to *apparent* u-shapes, others will not.

Perhaps the best way to see why splines do not answer the question of interest is by drawing an analogy to why we compute averages and run t-tests. We are able to see and plot the raw data, why run our statistical test on an impoverished representation, the mere average?

The reason is that we are not interested if one individual observations happens to be higher or lower than another, but rather, on whether, *in general,* observations in one condition are larger than in the other, and we operationalize *in general* by computing the average and asking if the average is larger. Here we want to know if *in general* the slope is positive at first and in general negative later on. We operationalize *in general* with the average here as well, the average *slope*.

The linearity and discontinuity assumptions are benign

The two-line test involves estimating an interrupted regression: two linear segments with a discontinuity in between. True causal models in psychology, the true underlying f(x), however, will almost never be properly captured by the combination of two straight lines with a discontinuity in between. This lack of correspondence between model and reality, however, is inconsequential for the purposes of conducting statistical inference for the presence of a u-shaped relationship (as opposed to for the purposes of fitting the data as well as possible).

This is in stark contrast to the assumptions involved in testing u-shapedness relying on a quadratic regression. There statistical inference hinges directly on the functional form assumption. Moreover, the two-lines test does not assume the true functional form is linear. The inference is just as valid if the data are vs are not linear, and do vs do not have a discontinuity. The regression lines are being used merely to compute an average slope, and averages are meaningful and valid construct "even if" the underlying data have variance. Figure 8 illustrates. The true relationship combines two different slopes, thus estimating a single slope introduces specification error, but that single slope nevertheless properly estimates the average slope in that range.

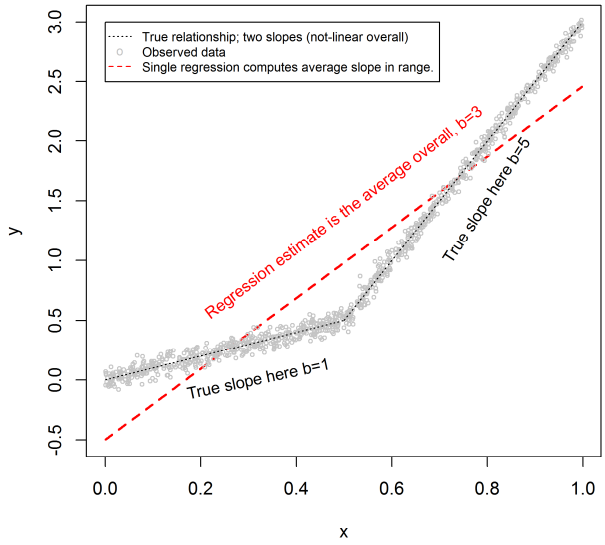

Fig. 7. *Linear regression recovers average slope in range even if effect is not linear.*
Note: R code to reproduce figure: https://osf.io/eswg6

Moreover, the two-lines procedure must include a discontinuity at x_c precisely because we simply want to rely on regression to estimate the average slope in the range, and if we do not allow a discontinuous change, if we estimate a *segmented* rather than an *interrupted* regression, then we introduce bias in the slope estimate. Figure 8 illustrates.

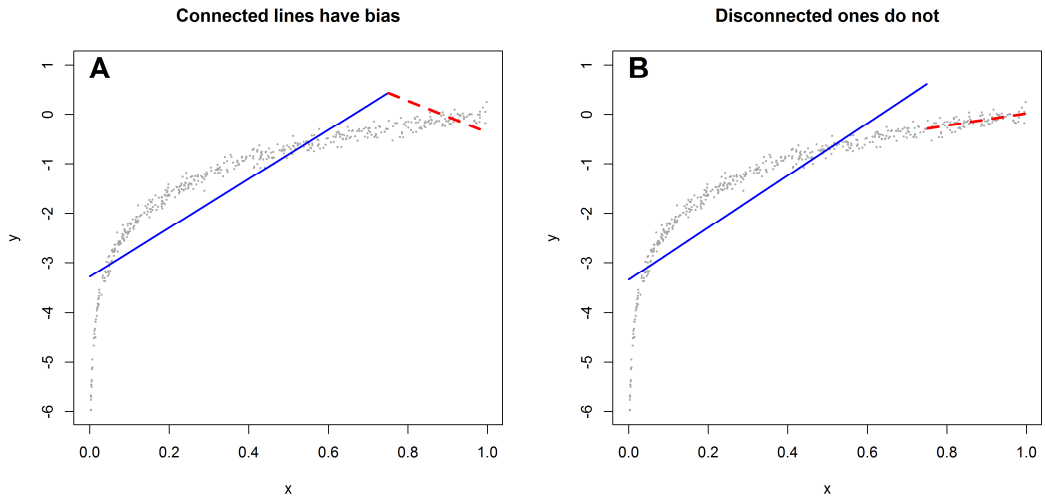

Fig. 8. *Forcing the two-lines to connect introduces bias*
Note: R Code to reproduce figure: https://osf.io/2w6xy

Performance of two-line test

False-positive and False-Negative U-Shapes

I conducted simulations for a broad range of scenarios (see notes for Figure 9 for details) to assess type-1 and type-2 errors for examining u-shapes with a quadratic regression vs. the two-line test. For the two-lines approach I considered not just the breakpoint identified by the aforementioned algorithm, but also for various alternative approaches. Setting it at: the highest fitted value from a quadratic regression, highest fitted value from the cubic spline, and at the point that maximizes the overall fit of the interrupted regression. Per the suggestion of the editor, I also estimate regressions with two interruptions, three total segments, and set the two breakpoints to maximize overall fit.

Beginning with the false-positive rates calculations. I focused on true relationships that would be most likely to lead to a false-positive u-shape: an initial strong effect, followed by a long flat segment (if the true relationship has a strong effect throughout, it is virtually impossible to obtain a false-positive u-shape with the two-lines test).

Figure 9A shows the quadratic regression approach to testing u-shapes has an unacceptably high false-positive, well above the nominal 5%. In stark contrast, the two-lines test has an acceptable false-positive rate regardless of how the breakpoint is set.[7,8]

[7] The breakpoint that maximizes overall fit is set using the 'segmented' package in R (Muggeo, 2003).
[8] Supplement 1 shows similar results for simulations where the true relationships is created using the logistic function rather than a combination of two linear or log-linear segments.

Because the quadratic regression is an invalid test, the right panel does not report power results for it. For statistical inference, we should select the most powerful test, *among those that satisfy the nominal false-positive rate*. If we are willing to use an invalid test, one with an elevated false-positive rate, we should rely on a coin that reads "u-shape" on either side: flipping the coin leads to 100% power. Figure 9b shows the two lines test has the most power when the breakpoint is set using Robin Hood algorithm.

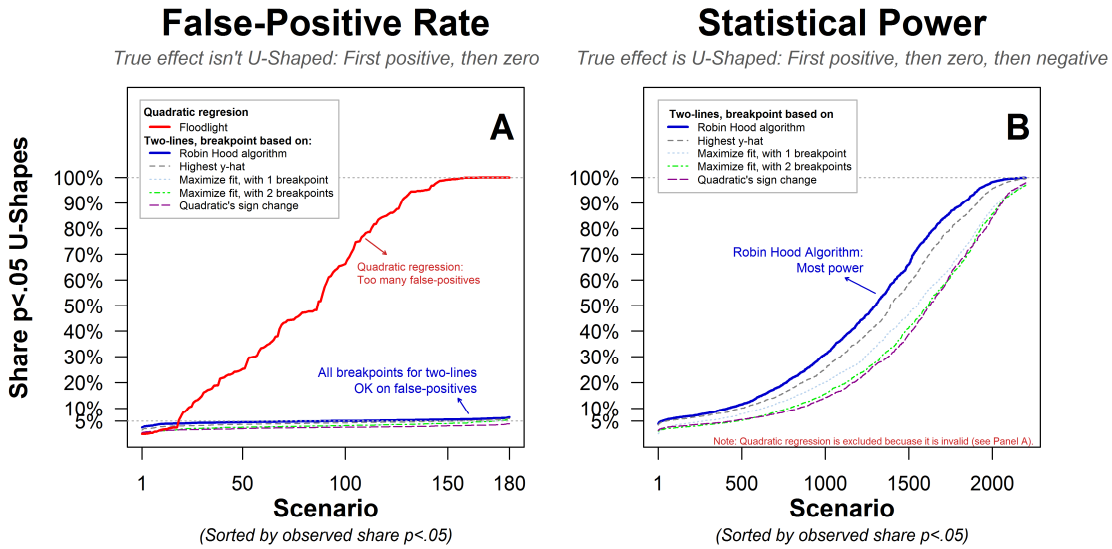

Fig. 9 *Detecting U-shapes*

Notes: Each simulated scenario involves a relationship between variables x and y where $y=x+e$ for $x<x_c$. For the false-positive simulations $y=x_c$ for $x \geq x_c$ (flat; no further effect of x). For the power simulations at some point x_d, with $x_d \geq x_c$ the effect becomes negative. For example, $y=x_c-(x-x_d)$ if $x \geq x_d$. The scenarios combine the following parametrizations: (i) the distributions of x (normal, uniform, beta with left, beta with right skew, optimized for the quadratic as in McClelland (1997))[9], (ii) the effect of x on y is y=x vs y=log(x), that is, linear vs log-linear, (iii) sample sizes of 100, 200, or 500, (iv) σ in $e \sim N(0,\sigma)$ with σ being 100%, 200% or 300% of the SD(y) before adding *e*, (v) the value of x_c: 30th, 50th percentile of x, (vi) the value of x_d : 30th, 50th, 70th, or 90th percentile of x, (vii) the slope of the negative effect of x on y when $x>x_d$ being 25%, 50%, 100% or 200% the magnitude of the slope when $x<x_c$. The full combination of options leads to 180 scenarios for false-positives and 2520 for power. The power panel shows only 2300, the remaining 220 have 99%+ power for all procedures. False-positive rates are based on 5000 simulations, power calculations on 500. R Code to reproduce figure: https://osf.io/9xwke

[9] In particular, 25% of observations are x<.2, 25% x>.8, and 50% are .4<x<.6 (McClelland, 1997; Table 1). This distribution is said to maximize power to detect a u-shape if the true relationship is quadratic and the maximum value is obtained at "intermediate values of X" (p.9).

Demonstrations

Figure 10 applies the two-line test to two examples in the published literature that appear to arrive at false-positive u-shape conclusions because they relied on quadratic regressions. Panels A&B revisit the analyses by Sterling, Jost, and Pennycook (2016) who wrote (in their *secondary* analyses section), that "those who were moderate in terms of their support for the free market appeared to be more susceptible to bullshit than extremists in either direction." (p.356). They arrive at this inverted-U conclusion because the quadratic term in the regression is significant (p=.026; Figure 10A).

I successfully reproduced their results analyzing their posted data. Figure 10B, however, shows that the second line, while negative, is far from significant (p=.41). Keep in mind that if x and y were uncorrelated for high values of x, that is, if the true second slope were flat, 50% of estimates will be negative (and 41% of them as steeply negative as observed; that's the meaning of the p=.41). The data are inconclusive: consistent with a u-shaped relationship, consistent with ideology and bullshit receptivity being uncorrelated among higher values of the former, and consistent with a monotonic effect. Again, the u-shape prediction was secondary to the authors. The paper's core prediction is consistent with the first line in Panel B: "free market ideology was significantly but modestly associated with bullshit receptivity" (abstract).

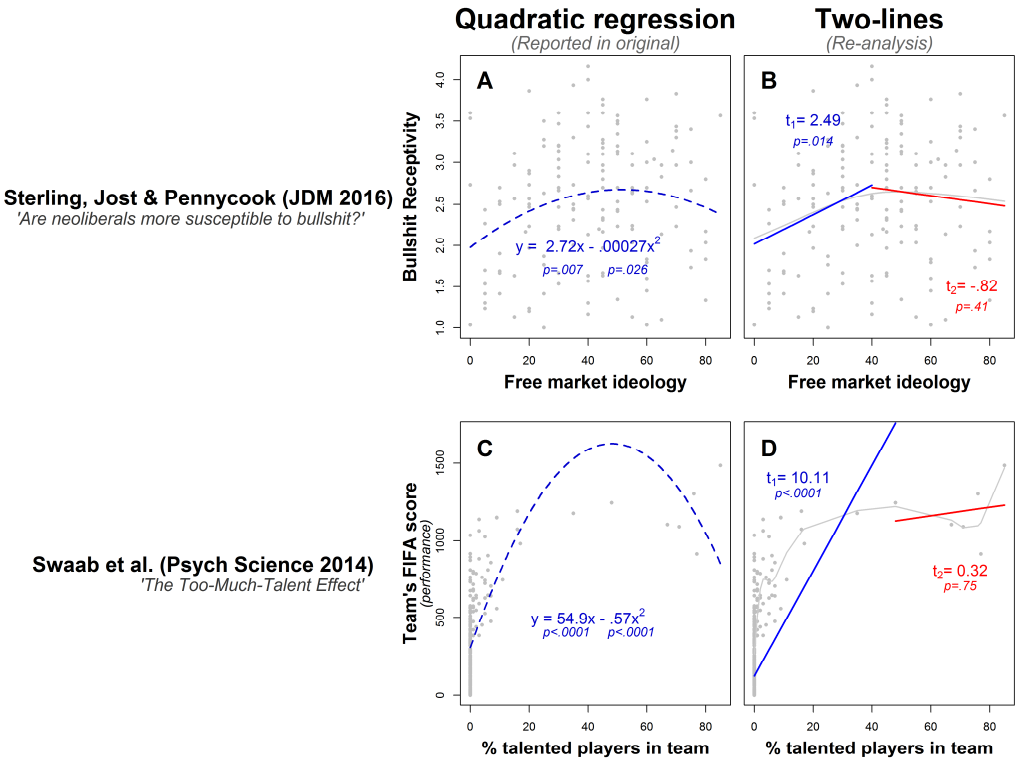

Fig. 10 *Quadratic vs Two-Lines applied to data from published papers*
Notes: In **A** & **B** each dot depicts a participant in a survey, y is how profound participants rated a series of "vague and meaningless statements," x their endorsement of "neoliberal" principles. In **C** & **D** each dot is a country, y is its FIFA rating, x the share of players in the country's team that play for a top professional team (e.g., *Arsenal*). Thin continuous lines in **B** and **D** are fitted values from cubic splines.
R Code to reproduce figure: https://osf.io/hkt2a

Continuing with Panels C & D: Swaab, Schaerer, Anicich, Ronay, and Galinsky (2014b), in their Study 2, examined the relationship between the number of elite players in a country's soccer team and its international FIFA rating. Their results, they write, "revealed a significant quadratic effect of top talent: Top talent benefited performance only up to a point, after which the marginal benefit of talent decreased *and turned negative*" (p.1584; emphasis added). I successfully replicate those results with independently obtained data (see Panel C), but in Panel D the second line is also positive. These data do not support the conclusion that there is such thing as 'too-much-talent' in international soccer.

Conclusions

The use of quadratic regressions to test u-shaped relationships is as invalid as it is common. The two-lines test is the first valid test of u-shape relationships in the literature and is arguably the most straightforward test of the hypothesis of interest: that the average effect of x on y is of opposite sign for high vs low values of x.

The two-lines test is expected to perform well as long as the true relationship of interest has at most two regions where the impact of x on y has opposite signs, that is, if the relationship of interest is: (i) flat overall (no effect), (ii) (weakly) monotonic, or (iii) u-shaped. It will not perform well, at least in terms of interpretability, if the true relationship has more than two regions with different signs, for instance, if it is N-shaped, X-shaped or W-shaped, rather than U-shaped. These relationships, it is worth noting, invalidate the quadratic regression as well.

The paper includes a supplement. Table 1 summarizes its contents.

Table 1.
Index of supplementary materials (available from https://osf.io/t6twm/)

Section	Pages
Supplement 1. False-positive u-shapes if true relationship is logistic	1-2
Supplement 2. Scatterplots for raw data behind specifications used in Figure 9B	3

References

Bai, J., & Perron, P. (1998). "Estimating and Testing Linear Models with Multiple Structural Changes". *Econometrica*, 47-78.

Berman, S. L., Down, J., & Hill, C. W. (2002). "Tacit Knowledge as a Source of Competitive Advantage in the National Basketball Association". *Academy of Management Journal, 45*(1), 13-31.

Bowman, A., Jones, M., & Gijbels, I. (1998). "Testing Monotonicity of Regression". *Journal of computational and Graphical Statistics, 7*(4), 489-500.

Choi, K., & Kirkorian, H. L. (2016). "Touch or Watch to Learn? Toddlers' Object Retrieval Using Contingent and Noncontingent Video". *Psychological science*. doi: 10.1177/0956797616636110

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.

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.

Hall, P., & Heckman, N. E. (2000). "Testing for Monotonicity of a Regression Mean by Calibrating for Linear Functions". *Annals of Statistics*, 20-39.

Jaspers, E., & Pieters, R. (2016). "Materialism across the Lifespan: An Age-Period-Cohort Analysis". *Journal of Personality and Social Psychology, 111*(3), 451-473. doi: 10.1037/pspp0000092

Johnson, P. O., & Neyman, J. (1936). "Tests of Certain Linear Hypotheses and Their Application to Some Educational Problems". *Statistical Research Memoirs, 1*, 57-93.

Josef, A. K., Richter, D., Samanez-Larkin, G. R., Wagner, G. G., Hertwig, R., & Mata, R. (2016). "Stability and Change in Risk-Taking Propensity across the Adult Life Span". *Journal of Personality and Social Psychology, 111*(3), 430-450. doi: 10.1037/pspp0000090

Koopmann, J., Lanaj, K., Wang, M., Zhou, L., & Shi, J. (2016). "Nonlinear Effects of Team Tenure on Team Psychological Safety Climate and Climate Strength: Implications for Average Team Member Performance". *Journal of Applied Psychology, 101*(7), 940-957. doi: 10.1037/apl0000097

Lind, J. T., & Mehlum, H. (2010). "With or without U? The Appropriate Test for a U-Shaped Relationship". *Oxford Bulletin of Economics and Statistics, 72*(1), 109-118.

Loschelder, D. D., Friese, M., Schaerer, M., & Galinsky, A. D. (2016). "The Too-Much-Precision Effect When and Why Precise Anchors Backfire with Experts". *Psychological Science*. doi: 10.1177/0956797616666074

Marsh, L. C., & Cormier, D. R. (2001). *Spline Regression Models* (Vol. 137): Sage.

McClelland, G. H. (1997). "Optimal Design in Psychological Research". *Psychological Methods, 2*(1), 3-19.

Miller, J. W., Stromeyer, W. R., & Schwieterman, M. A. (2013). "Extensions of the Johnson-Neyman Technique to Linear Models with Curvilinear Effects: Derivations and Analytical Tools". *Multivariate behavioral research, 48*(2), 267-300.

Muggeo, V. M. (2003). "Estimating Regression Models with Unknown Break-Points". *Statistics in medicine, 22*(19), 3055-3071.

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). "Replicable Effects of Primes on Human Behavior". *Journal of Experimental Psychology: General, 145*(10), 1269-1279.

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.

Seidman, G. (2012). "Positive and Negative: Partner Derogation and Enhancement Differentially Related to Relationship Satisfaction". *Personal Relationships, 19*(1), 51-71.

Simonton, D. K. (1976). "Biographical Determinants of Achieved Eminence: A Multivariate Approach to the Cox Data". *Journal of personality and social psychology, 33*(2), 218.

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.

Sterling, J., Jost, J. T., & Pennycook, G. (2016). "Are Neoliberals More Susceptible to Bullshit?". *Judgment and Decision Making, 11*(4), 352-360.

Swaab, R. I., Schaerer, M., Anicich, E. M., Ronay, R., & Galinsky, A. D. (2014a). Response to Post, "Thirty-Somethings Are Shrinking and Other Challenges for U-Shaped Inferences", from http://web.archive.org/web/20160418215725/http://datacolada.org/wp-content/uploads/2014/09/AuthorsResponse3.pdf

Swaab, R. I., Schaerer, M., Anicich, E. M., Ronay, R., & Galinsky, A. D. (2014b). "The Too-Much-Talent Effect Team Interdependence Determines When More Talent Is Too Much or Not Enough". *Psychological Science, 25*(8), 1581-1591.

Ungemach, C., Stewart, N., & Reimers, S. (2011). "How Incidental Values from the Environment Affect Decisions About Money, Risk, and Delay". *Psychological Science, 22*(2), 253-260.

von Bastian, C. C., Souza, A. S., & Gade, M. (2016). "No Evidence for Bilingual Cognitive Advantages: A Test of Four Hypotheses". *Journal of Experimental Psychology: General, 145*(2), 246-258.

Wilson, K. S., DeRue, D. S., Matta, F. K., Howe, M., & Conlon, D. E. (2016). "Personality Similarity in Negotiations: Testing the Dyadic Effects of Similarity in Interpersonal Traits and the Use of Emotional Displays on Negotiation Outcomes". *Journal of Applied Psychology, 101*(10), 1405-1421. doi: 10.1037/apl0000132

Wood, S. (2006). *Generalized Additive Models: An Introduction with R* (1st ed.): Chapman & Hall/CRC Monographs on Statistics and Applied Probability.