nature human behaviour

RESOURCE

https://doi.org/10.1038/s41562-020-0912-z

Check for updates

Specification curve analysis

Uri Simonsohn[1]✉, Joseph P. Simmons[2] and Leif D. Nelson[3]

Empirical results hinge on analytical decisions that are defensible, arbitrary and motivated. These decisions probably introduce bias (towards the narrative put forward by the authors), and they certainly involve variability not reflected by standard errors. To address this source of noise and bias, we introduce specification curve analysis, which consists of three steps: (1) identifying the set of theoretically justified, statistically valid and non-redundant specifications; (2) displaying the results graphically, allowing readers to identify consequential specifications decisions; and (3) conducting joint inference across all specifications. We illustrate the use of this technique by applying it to three findings from two different papers, one investigating discrimination based on distinctively Black names, the other investigating the effect of assigning female versus male names to hurricanes. Specification curve analysis reveals that one finding is robust, one is weak and one is not robust at all.

The empirical testing of scientific hypotheses requires data analysis, but data analysis is not straightforward. To convert a scientific hypothesis into a testable prediction, researchers must make a number of data analytic decisions, many of which are both arbitrary and defensible. For example, researchers need to decide which variables to control for, which observations to exclude, which functional form to assume, which subgroups to analyse, and so on.

When reading the results of a study, people want to learn about the true relationship being analysed but this requires that the analyses reported are representative of the set of valid analyses that could have been conducted. This is often not the case. One problem is the possibility that the results may hinge on an arbitrary choice by the researcher[1]. A probably greater, more pervasive problem is that people in general, and researchers in particular, are more likely to report evidence consistent with the claims they are trying to make than to report evidence that is inconsistent with such claims[1–4]. The standard errors around published effect sizes represent the sampling error inherent in a particular analysis, but they do not reflect the error caused by the arbitrary and/or motivated selection of specifications.

In this article we introduce specification curve analysis as a way to mitigate this problem. The approach consists of reporting the results for all (or a large random subset thereof) 'reasonable specifications', by which we mean specifications that are (1) sensible tests of the research question, (2) expected to be statistically valid and (3) not redundant with other specifications in the set.

The specification 'curve' shows the estimated effect size across all specifications, sorted by magnitude, accompanied below by a 'dashboard chart' indicating the operationalizations behind each result. This enables visual identification by the reader of both variation in effect size across specifications and its covariation with operationalization decisions. Specification curve analysis also includes an inferential component, which combines the results from all specifications into a joint statistical test. It assesses whether, in combination, all specifications reject the notion that the effect of interest does not exist.

There is a long tradition of considering robustness to alternative specifications in social science. The norm in economics and political science, for example, is to report regression results in tables in which each column reports a different specification, allowing readers to compare results across specifications. We can think of specification curve analysis as an extension and formalization of that approach, one that substantially reduces the room for selective reporting (from grey dots to red ovals in Fig. 1).

There have been other attempts to formalize this process. One proposal is that researchers modify the estimates of a given model to take into account an initial model selection process guided by fit (for example, when deciding between a quadratic versus cubic polynomial[5]). Another assesses whether the best-fitting model among a class of models fits better than expected by chance[6]. A third proposal consists of reporting the standard deviation of point estimates across a few carefully chosen alternative specifications[7]. A fourth approach is known as 'extreme bounds analysis'[1], where a regression model for every possible combination of covariates is estimated. A relationship of interest is considered 'robust' only if it is statistically significant in all models, or if a weighted average of the *t*-test in each model is itself statistically significant[8]. A more recent proposal consists of estimating a large number of specifications, going beyond just covariates to include functional form and regression model and plotting the distribution of results obtained across specifications[9,10].

Specification curve analysis contributes to these efforts by facilitating the visual identification of the source of variation in results across specifications (see Fig. 2), without imposing linearity on such effects[10]. Specification curve analysis also provides a formal joint significance test for the family of alternative specifications, derived from expected distributions under the null.

A non-statistical approach to dealing with selective reporting consists of pre-analyses plans[11,12]. Specification curve analysis complements this approach, allowing researchers to pre-commit to running the entire set of specifications they consider valid, rather than only a small and arbitrary subset of them as they must currently do. Researchers, in other words, could pre-register their specification curves.

If different valid analyses lead to different conclusions, traditional pre-analysis plans lead researchers to blindly pre-commit to one versus the other conclusion by pre-committing to one versus another valid analysis, while specification curve allows researchers to learn which specifications the conclusion hinges on.

[1]ESADE Business School, Behavioral Science, Universitat Ramon Llull, Barcelona, Spain. [2]The Wharton School, Operations Information & Decisions Department, University of Pennsylvania, Philadelphia, PA, USA. [3]Berkeley Haas School of Business Marketing Department, University of California, Berkeley, CA, USA. ✉e-mail: urisohn@gmail.com

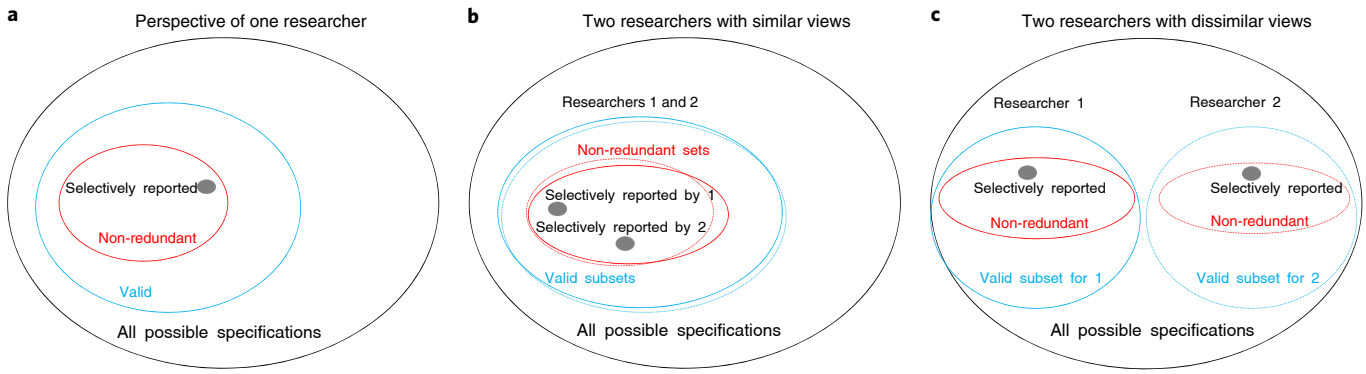

Fig. 1 | Sets of possible specifications as perceived by researchers. a, The set of specifications reported in an article are a small subset of those the researcher would consider valid to report. **b**, Different researchers may have similar views on the set of valid specifications but report quite different subsets of them. **c**, Different researchers may also disagree on the set of specifications they consider valid.

An intuitive presentation of the problem we want to solve

To analyse data, we need to make decisions about specifications. Some of these decisions are guided by theory or beliefs about the phenomenon of interest. Other decisions are guided instead by convenience, happenstance, the desire to report stronger-looking results, or nothing at all. Specification curve analysis is concerned with minimizing the impact of specification decisions that are based on neither theory nor beliefs.

Some researchers object to blindly running alternative specifications that may make little sense for theoretical or statistical reasons just for the sake of 'robustness'. We are among those researchers. We believe one should test specifications that vary in as many of the potentially ad hoc assumptions as possible without testing any specifications that are not theoretically grounded. If a specification does not make sense theoretically or statistically, or if it is unambiguously inferior to alternative specifications, it does not belong in a robustness test in general, nor in a specification curve in particular.

For example, a researcher interested in the causal effect of raising children on adult happiness should control for the marital status of the adults. Because married adults are more likely to have children than unmarried ones, the estimates of the happiness effect of raising children will (partially) include the separable effect on happiness of having a spouse[13]. Thus, reporting results with and without controlling for marital status may be interesting and informative, but it does not constitute an exercise in robustness because both sets of results do not provide two a priori equally valid answers to the same research question. Only specifications that include a control for marital status represent valid tests of this hypothesis.

Nevertheless, many analytic decisions are arbitrary and no more or less defensible than any others. For instance, in an event study, we should expect robustness tests on the definition of the length of the before and after periods[14]. In a study on the effect of income on well-being we should expect robustness tests on different measures of well-being—say, happiness and life satisfaction[15]. In a study on labour participation we should expect robustness tests on what is used as the full-time equivalence of someone working part-time[16].

Figure 1 helps to illustrate what it means, and what it does not mean, to report the results of a representative set of reasonable specifications. Figure 1a depicts the menu of specifications as seen from the eyes of a given researcher. There is a large, possibly infinite, set of specifications that could be run. The researcher considers only a subset of these to be valid (the blue oval), some of which are redundant with one another (for example, log transforming x using $\log(x+1)$ or using $\log(x+1.1)$). The set of reasonable specifications (the red oval) includes only the non-redundant alternatives (for example, either $\log(x+1)$ or $\log(x+1.1)$, but not both).

Because competent researchers often disagree about whether a specification is an appropriate test of the hypothesis of interest and/or statistically valid for the data at hand (that is, because different researchers draw different ovals), specification curve analysis will not end debates about what specifications should be run: specification curve analysis will instead facilitate those debates.

Even if two researchers have non-overlapping sets of reasonable specifications, specification curve analysis can help them understand why they may have reached different conclusions, by disentangling whether those different conclusions are driven by different beliefs about which specifications are valid, or whether they are driven by arbitrary selectively reported results from those sets. In other words, specification curve disentangles whether the different conclusions originate in differences regarding which sets of analyses are deemed reasonable (different red ovals), or merely in which few analyses the researchers reported (different grey dots).

A formal presentation of the problem we want to solve

Let's consider a relationship of interest between variables x and y, in a context in which other variables, Z, may influence the relationship: $y=F(x,Z)+e$. For example, x may be education, y may be economic success and Z may include moderators (for example, school quality) and/or confounds (for example, parental education); e consists of orthogonal predictors of y (for example, luck).

Learning about $y=F(x,Z)$ poses several practical challenges: (1) x and y are often imprecisely defined latent variables (for example, education and economic success are both imprecisely defined latent variables); (2) the set of moderators and confounders in Z are often not fully known ex ante; (3) Z also contains imprecisely defined latent variables (for example, school quality is a latent and imprecisely defined predictor of economic success); and (4) the functional form $F()$ is not known. To study $y=F(x,Z)$, researchers must operationalize the underlying constructs. Let's designate the operationalization of a construct θ, with θ. Researchers, then, approximate $y=F(x,Z)$ with a specification, a set of operationalizations: $y_{k_y} = F_{k_F}(x_{k_x}; Z_{k_Z})$, where k_y, k_F, k_x and k_Z are indices for single operationalizations of the respective constructs. For example y_1 may operationalize 'economic success' with yearly salary, while y_2 with private jet seat capacity.

For each construct there are multiple statistically valid, theoretically justified and non-redundant operationalizations. Their combination leads to what we refer to as the set of reasonable specifications, which, as discussed in the previous section, may be at least somewhat subjective. Designating the total number of valid operationalizations for each construct with n_y, n_x, n_Z and n_F, the total

NOTE by Uri Simonsohn: Nature HB dropped, by mistake, some important notation in this section (after we had approved the proofs). For the author-approved version of this section, please see last page of this .pdf

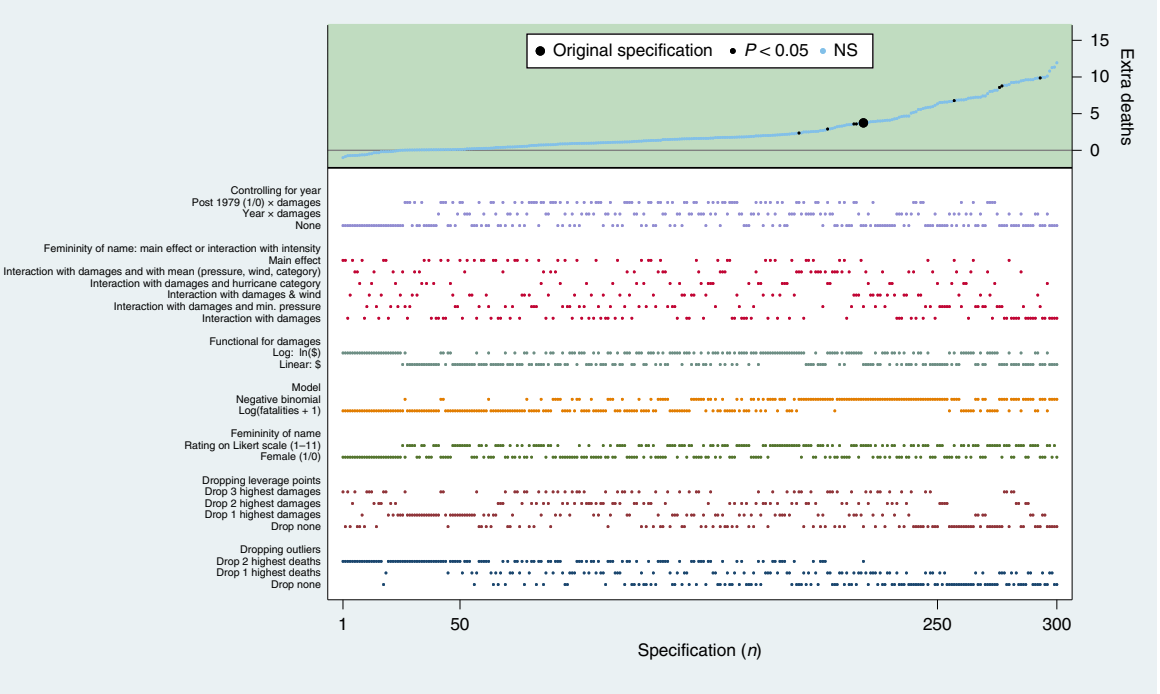

Fig. 2 | Descriptive specification curve. Each dot in the top panel (green area) depicts the marginal effect, estimated at sample means, of a hurricane having a female rather than male name; the dots vertically aligned below (white area) indicate the analytical decisions behind those estimates. A total of 1,728 specifications were estimated; to facilitate visual inspection, the figure depicts the 50 highest and lowest point estimates and a random subset of 200 additional ones, but the inferential statistics for specification curve analysis include all 1,728 specifications. NS, not significant.

number of reasonable specifications available to study $y=F(x,Z)$ is $N \leq n_x \times n_y \times n_Z \times n_F$.

Let Π be this set of N reasonable specifications, and π be the subset of specifications reported in a paper. Thinking about π as a sample of Π makes it easier to understand the problem that specification curve analysis attempts to remedy.

By definition, any given $y_{k_x} = F_{k_F}(x_{k_x}; Z_{k_Z})$ is considered a valid proxy for $y=F(x,Z)$ and therefore so is the full set of all such proxies: Π. A (1) large, (2) random and (3) independently drawn sample of Π would thus lead to a reasonable estimate of the model of interest: $y=F(x,Z)$. The problem is that π, the sample of specifications reported in a paper, has none of these three properties.

First, it is small, not large. Researchers report a few specifications in any given paper, providing a statistically noisy approximation. Second, it is a curated rather than a random sample. Researchers often choose which specifications to report knowing the results of these versus other specifications, after knowing how they, reviewers, and audience members respond to different results. Thus, π is chosen by a person seeking academic success, not by a random sampling procedure blind to the consequences of selecting one versus another specification to report.

Third, and least obvious, the specifications in π are not statistically independent. How much information is there in the fact that a result is obtained across ten rather than just three specifications? It depends on how statistically independent the alternative specifications are. In other words, it depends on how likely it is, under the null, that one specification in π will show an effect if another specification in π already does. Currently the statistical independence of robustness results is not considered, either formally or informally. Results are labelled as robust without considering how likely they are to coincide by chance alone.

Specification curve analysis addresses all three of these problems. First, it generates a much larger π, where hundreds or even thousands of specifications are reported. This increases statistical efficiency by reducing specification noise. It also makes transparent the existence of such noise, and allows readers to determine its nature (that is, which operationalization decisions are versus are not consequential). Second, specification curve analysis generates a π with fewer arbitrary inclusion decisions, and thus more closely approximates a random sample of Π. When using specification curve analysis we can more legitimately consider π as an approximation of $y=F(x,Z)$, though for the sampling to be even closer to random it would need to be performed by researchers who are blind to the consequence of choosing one versus another specification. Third, specification curve analysis allows statistical inference that takes into account the statistical dependence across alternative specifications in π.

The null hypothesis that the true effect of x on y is zero for all specifications is thus: H_0: $\frac{d(F_{k_F})}{d(x_{k_x})}=0$, $\forall\ \pi_k$ in Π, where π_k indexes the valid operationalizations in Π. For example, considering the special (though quite general) case of a general additive model where $F(x,Z) = f_x(x)+f_z(z)+f_{xZ}(xZ)$, the null is H_0: $\frac{d(f_{x_{kF}})}{d(x_{k_x})} = \frac{d(f_{xZ_{kF}})}{d(x_{k_x})}=0$, $\forall\ \pi_k$ in Π and $\forall$ observable x.

Conducting specification curve analysis

Specification curve analysis is carried out in three main steps: (1) define the set of reasonable specifications to estimate; (2) estimate all specifications and report the results in a descriptive specification curve; and (3) conduct joint statistical tests using an inferential specification curve.

We demonstrate these three steps by applying the specification curve to two published articles with publicly available raw data. One reports that hurricanes with more feminine names have caused more deaths[17]. We selected this paper because it led to an intense debate about the proper way to analyse the underlying data[17–22], providing an opportunity to assess the extent to which specification curve

Table 1 | Original and alternative reasonable specifications used to test whether hurricanes with more feminine names were associated with more deaths

Decision	Original specifications	Alternative specifications
(1) Which storms to analyse	Excluded two outliers with the most deaths	Dropping fewer outliers (zero or one); dropping storms with extreme values on a predictor variable (for example, hurricanes causing extreme damages)
(2) Operationalizing hurricane names' femininity	Ratings of femininity by coders (1–11 scale)	Categorizing hurricane names as male or female
(3) Operationalizing hurricane strength	Property damages in dollars; minimum hurricane pressure	Log of dollar damages, hurricane wind speed.
(4) Type of regression model	Negative binomial regression	Ordinary least squares with log(deaths + 1) as the dependent variable
(5) Functional form for femininity	Assessed whether the interaction of femininity with damages was greater than zero	Main effect of femininity; interacting femininity with other hurricane characteristics (for example, wind or category) rather than damages

analysis could inform such debates. The second article reports a field experiment examining racial discrimination in the job market[23]. We selected this highly cited article because it allowed us to showcase the range of inferences that specification curves can support. We discuss in detail each of the three steps for specification curve analysis with the first example, and then apply them to the second.

Both of these examples involve a key predictor that is orthogonal to all others. In a later section we explain how to conduct inference in specification curve analysis when this is not the case (for example, when data do not originate in an experiment).

Step 1: Identify the set of specifications. The set of reasonable specifications can be generated by (1) enumerating all of the data analytic decisions necessary to map the scientific hypothesis or construct of interest onto a statistical hypothesis; (2) enumerating all the reasonable alternative ways a researcher may make those decisions; and (3) generating the exhaustive combination of decisions, eliminating combinations that are invalid or redundant. If the resulting set is too large, then in the next step (estimation) one can randomly draw from them to create specification curves.

To illustrate, in the hurricanes study[17] the underlying hypothesis was that hurricanes with more feminine names cause more deaths because they are perceived as less threatening, leading people to engage in fewer precautionary measures.

As shown in Table 1, we identified five major data analytic decisions required to test this hypothesis, including which storms to analyse, how to operationalize hurricanes' femininity, how to operationalize the severity of the hurricane, which regression model to use and which functional form to assume for the effect of hurricane name. Although the authors' specification decisions appear reasonable to us, there are many more alternatives that are just as reasonable. The combination of all operationalizations we considered valid and non-redundant makes up our red oval, a set of 1,728 reasonable specifications (see Supplementary Note 1 for details).

Step 2: Estimate and describe results. The descriptive specification curve serves two functions: displaying the distribution of estimates that are obtained through alternative reasonable specifications, and identifying which analytical decisions are the most consequential. When the set of reasonable specifications is too large to be estimated in full, a practical solution is to estimate a random subset of, say, a few thousand specifications.

Figure 2 shows the descriptive specification curve for the hurricanes example. The top panel depicts estimated effect size, in additional fatalities, of a hurricane having a feminine rather than masculine name. The figure shows that the majority of specifications lead to estimates of the sign predicted by the original authors (feminine hurricanes produce more deaths), though a very small minority of all estimates is statistically significant ($P < 0.05$). The point estimates range from −1 to +12 additional deaths. To make comparable point estimates for the continuous and discrete measures of femininity, we compute the average value of the former for the two possible values of the latter, and compute as the effect size the difference in predicted deaths for both values. Estimates are marginal effects computed at sample means.

The bottom panel of the figure tells us which analytic decisions produce different estimates. For example, we can see that obtaining a negative point estimate requires a fairly idiosyncratic combination of operationalizations: (1) not taking into account the year of the storm, (2) operationalizing the severity of the storm by the log of damages, (3) conducting an ordinary least squares regression, and so on. A researcher motivated to show a negative point estimate would be able to report 20 different specifications that do so, but the specification curve shows that a negative point estimate is atypical.

Following the publication of the hurricanes paper, the journal *Proceedings of the National Academy of Sciences* (PNAS) published four letters/critiques proposing alternative specifications under which the impact of hurricane name on fatalities disappears[18–21]. In particular, the critiques argued that outlier observations with >100 deaths should be excluded[19,21], that the regression should include an interaction between intensity of the hurricane and dollar damages as a predictor[18], and that dollar damages should not be included as a predictor at all[20].

Returning to Fig. 1, this appears to be a panel c situation. Original authors and critics disagree on which set of valid specifications to run. The specification curve results from Fig. 2 show that, while such disagreements may be legitimate and profound, we do not need to address them to determine what to make of the hurricane data. In particular, the figure shows that even keeping the same set of observations as the original study and treating damages in the same way as in the original, it is the case that modifying virtually any arbitrary analytical decision renders the original effect non-significant. Readers need not take a position on whether it does or does not make sense to include a damages × pressure interaction in the model to determine whether the original findings are robust.

When the number of specifications is large, a descriptive specification curve may be too dense for visual identification of patterns of interest; in Supplementary Note 5 (Supplementary Figs. 8–10) we propose a few alternative visualizations.

Figure 2 shows that PNAS could have published nearly 1,700 letters showing individual specifications that make the effect disappear (without deviating from the original red oval). It also could have published 37 responses with individual specifications showing the robustness of the findings. It would have been better to publish a single specification curve in the original paper. Visually inspecting Fig. 2, we learn not only about the variability of the point estimate across specifications but also about which operationalizations are consequential. For example, we learn that (1) only by logging both damages and fatalities and dropping the two outliers does a sign reversal arise; (2) the treatment of outliers is consequential but the definition of hurricane severity less so; (3) effects become larger as

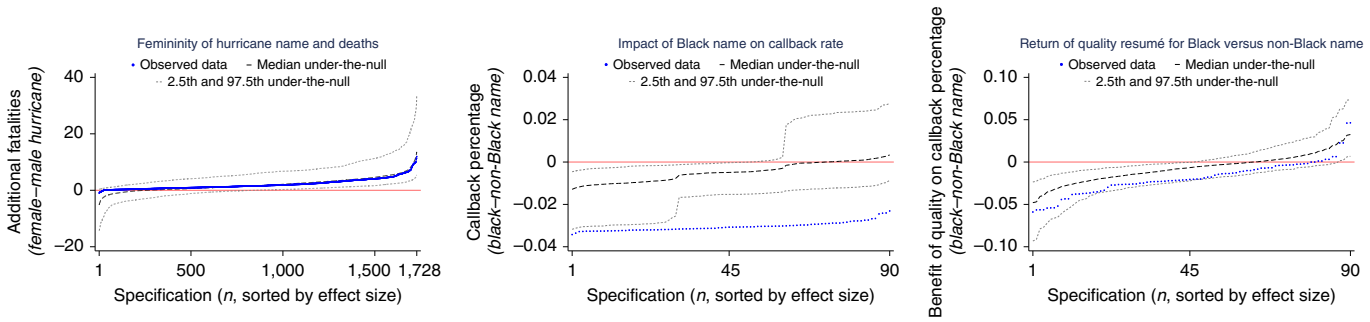

Fig. 3 | Observed and expected under-the-null specification curves for the hurricanes and racial discrimination studies. Observed and expected under-the-null specification curves for the hurricanes and racial discrimination studies. The expected curves are based on 500shuffled samples where the key predictor in each dataset (hurricane and applicant name, respectively) is shuffled. All specifications are estimated on each shuffled sample (1,728specifications for hurricanes study, 90 for racial discrimination). The resulting estimates for each shuffled dataset are ranked from smallest to largest. The dashed lines depict the 2.5th, 50th and 97.5th percentiles for each of these ranked estimates (for example, the median smallest estimate, the median second smallest estimate and so on). Specification curves under the null are typically not symmetric around zero (see Main text). Blue dots depict the specification curve for the observed data.

outliers are retained; and (4) the negative binomial gives systematically larger results than log(deaths + 1).

Inference with specification curve analysis

The third step of specification curve analysis involves statistical inference, answering the question: considering the full set of reasonable specifications jointly, how inconsistent are the results with the null hypothesis of no effect?

The null hypothesis is that effect of x on y, in $y = F(x,Z)$, is zero. Implementing the testing of this null requires a test statistic, a single scalar on which we can measure the extremity of the data, the results of all $y_{k_y} = F_{k_F}(x_{k_x}; Z_{k_Z})$ specifications in π, given the null hypothesis. We propose three test statistics for specification curve analysis. The first consists of obtaining the median effect estimated across all specifications, and then testing whether this median estimated effect is more extreme than would be expected if all specifications had a true effect of zero.

The second test statistic consists of the share of specifications that obtain a statistically significant effect in the predicted direction, testing whether such share is more extreme (higher) than would be expected if all specifications had an effect of zero. The third test statistic is similar to the second, but rather than discretizing each Pvalue into a significant versus non-significant dichotomous variable, and counting them, it aggregates all of them in a continuous fashion, by averaging the Zvalue associated with each (for example, $Z = 1.96$ for $P = 0.05$), as in Stouffer's method, and testing whether the average Zvalue across all specifications is more extreme than would be expected if the true effect were zero in all specifications. The third test statistic bypasses arbitrary discretization and is thus preferable from a statistical efficiency perspective, but the count of statistically significant specification results is a more intuitive metric that answers a question readers are more likely to ask. Rather than choosing between a simpler and a more statistically efficient result, we propose reporting both.

We do not believe it is possible to generate the distributions for any of these test statistics under-the-null analytically (that is, with statistical formulas), because the specifications are neither statistically independent nor part of a single model. Fortunately, it is simple to generate such distributions by relying on resampling under-the-null. This involves modifying the observed data so that the null hypothesis is known to be true, and then drawing random samples of the modified data. The test statistic of interest is then computed on each of those samples. The resulting distribution is the estimated distribution of the test statistic under the null[24–28].

The implementation of under-the-null resampling is more intuitive for experiments than for non-experiments, where covariates are possibly correlated with the predictor of interest. The two examples in this paper involve experiments and we thus explain resampling for experiments in this section. Resampling for observational data is discussed in more detail in in Supplementary Notes 1–5.

Because specification curve analysis relies on resampling for inference, it will be generally robust to assumption violations of the underlying specifications. For instance, if due to a violated assumption, some specifications have inflated false-positive rates—for example, exhibiting a 14% chance of obtaining $P < 0.05$ when the null is true, instead of the nominal 5%, by relying on resampling-based inference—the false-positive rate will be corrected and returned to 5%. In Supplementary Note 4 we provide a demonstration: a specification curve that combines a series of Poisson regressions, each with an inflated false-positive rate (>40%), obtains—overall—the nominal 5% false-positive rate for the specification curve that combines them all.

Example 1: inference in the hurricanes paper

Resampling experimental data under the null is simple and intuitive, as it involves shuffling the column(s) with the randomly assigned variable(s)[29–32]. In the case of the hurricanes paper, one shuffles the hurricane's name. The shuffled datasets maintain all the other features of the original (for example, collinearity, time trends, skewness and so on) except that we now know there is no link between (shuffled) names and fatalities; the null is true by construction. For each shuffled dataset we estimate all 1,728 specifications. Repeating this exercise many times gives us the distribution of specification curves under the null. The only assumption behind this test is exchangeability[31,32], that any hurricane could have received any name. The resulting Pvalues are hence 'exact', not dependent on distributional assumptions.

Sign: because many of the different specifications are similar to each other (for example, the same analysis conducted with an outlier included versus excluded), the results obtained from different specifications are not independent. Therefore, even with shuffled datasets we do not expect half the estimates to be positive and half negative on any given shuffled dataset; rather, we would expect most specifications to be of the same sign. In the extreme case, if all specifications were identical to one another, all results for any given data would be identical and thus in each shuffled dataset 100% of results would be positive or 100% negative.

To capture this lack of independence graphically, we refer to the sign of the majority of estimates for a given dataset as the 'dominant

Table 2 | Joint tests for inferential specification curves in the two examples

Test statistic used	Observed result	*P* value (% of shuffled samples with results as, or more, extreme)
Example 1: Female hurriance names		
(1) Median effect size	1.56 additional deaths	$P = 0.536$
(2) Share of significant results	37 of 1,728 specifications	$P = 0.850$
(3) Aggregate all *P* values	Stouffer $Z = 28.47$	$P = 0.512$
Example 2a: Black names receive fewer callbacks		
(1) Median effect size	3.1 pp fewer calls	$P < 0.002$
(2) Share of significant results	85 of 90 specifications	$P < 0.002$
(3) Aggregate all *P* values	Stouffer $Z = 35.71$	$P < 0.002$
Example 2b: Black names benefit less from higher quality CV		
(1) Median effect size	2.0 pp smaller benefit	$P = 0.162$
(2) Share of significant results	13 of 90 specifications	$P = 0.032$
(3) Aggregate all *P* values	Stouffer $Z = 9.22$	$P = 0.126$

Each overall *P* value is computed by the proportion of shuffled samples leading to a test statistic at least as extreme as in the observed sample. For *P* value calculations, we divide by two the proportion of shuffled samples, resulting in a test statistic of the exact same value as that in the observed data[34]. When no shuffled sample is as extreme as the observed, we report $P < 0.002$ because our estimate is that it is less frequent than 1 out of the 500 samples we collected. However, estimates as small as that are more susceptible to random simulation error. Stouffer's *Z* is computed by converting each *P* value to a *Z*-score (normal deviate) and then computing a weighted average, where the weight is 1 divided by the square root of the number of tests. The *P* value associated with this is also obtained via resampling, rather than from the normal distribution, to take into account the lack of independence across specifications (which is why $Z = 9.22$ (last row in Table 2) has a non-significant *P* value). pp, percentage points.

sign' and we plot results as having the dominant or non-dominant sign rather than a positive or negative sign. This allows visual capture of how similar estimates of a given dataset are expected to be across specifications. This constitutes a two-sided test where 80% of specifications—say, having the same sign—is treated as an equally extreme outcome regardless of whether it is 80% positive or 80% negative.

Results for hurricanes study: Figure 3a contrasts the specification curves from 500 shuffled samples with that from the observed hurricane data. The observed curve from the real data is quite similar to that obtained from the shuffled datasets—that is, we observe what is expected when the null of no effect is true. Table 2 reports the results of the three proposed test statistics for statistical inference: (1) median effect size, (2) share of results that are significant and (3) the average *Z*-score transformation of each *P* value (Stouffer's method).

For example, in the observed hurricane data, 37 of the 1,728 specifications are statistically significant in the predicted direction. Among the 500 shuffled samples, 425 have at least 37 significant effects in the same direction, leading to a *P* value for this joint test of $P = 425/500 = 0.85$.

Example 2: discrimination in an audit study. Having gone through the three steps for carrying out specification curve analysis with our first example, we move on to our second example[23], a field experiment in which researchers used fictitious resumés to apply for real jobs using randomly assigned names that were distinctively Black (for example, Jamal or Lakisha) or not (for example, Greg or Emily).

The authors of this article arrived at two key conclusions: applicants with distinctively Black names (1) were less likely to be called back and (2) benefited less from having a higher-quality resumé. We conducted specification curve analysis for both of these findings. For ease of exposition, we considered the same set of specifications for both, although they more naturally apply to finding (2). In particular, we considered two alternative regression models (ordinary least squares versus probit), three alternative samples (men and women, only men, and only women), and 15 alternative definitions of resumé quality. These resulted in a set of 90 reasonable specifications. We justify this set of specifications and report the descriptive specification curves in Supplementary Notes 2 and 3, respectively.

Figure 3b,c shows the inferential specification curve results for these findings. Starting with the core finding that distinctively Black names had lower callback rates (Fig. 3c), we see that the entire observed specification curve falls outside the 95% confidence interval around the null. In Table 2 we see that the null hypothesis is formally rejected.

The robustness of the second finding, that resumés with distinctively Black names benefitted less from higher quality, is less clear. The observed specification curve never crosses the 95% confidence interval (Fig. 3b), and only one of the joint tests is significant at the 5% level.

Inference with non-experimental data

To force the null on non-experimental data, we propose the following procedure, which is nearly equivalent to that of Flachaire[33]. For each specification one first estimates the model with the observed data—say, estimating the parameters a, b and c in $y = a + bx + cz + e$. Then one forces the null on the data by creating a new dependent variable, y^*, that subtracts the estimated effect of x on y—that is, $y^* = y - \hat{b}x$, where $\hat{b}$ is the sample estimate of b. With y^* we now have a model where the null is true—that is, we have $y^* = a + b^*x + cz + e$, where we know that $b^* = 0$.

To generate a distribution of expected results, the sampling distribution of $\hat{b}$ under the null, one samples with replacement rows of data by using y^* rather than y as the dependent variable. Each resample has the same sample size as the original. The resulting distribution of $\hat{b}$ across the resamples is used to assess the extremity of the observed $\hat{b}$ if the null were true. Applying this approach to specification curve analysis leads to the following six steps:

(1) Estimate all K specifications with the observed data, $y_{k_y} = F_{k_F}(x_{k_x}; Z_{k_Z})$. These will result in K different point estimates: $\hat{b}_k$. with $k = 1 \ldots K$. Note that y_{k_y} may be the same for more than one specification, even for all K of them, if the operationalization of the dependent variable is not varied across specifications.
(2) Generate K different dependent variables under the null, $y_k^* = y_k - \hat{b}_k \times x_k$. Even if there are fewer than K different y_k, there will be K different y^*_k because $\hat{b}_k$ is different across specifications and thus so is y^*_k. So now every row of data has the x values and K different y^* values.
(3) Draw at random, and with replacement, N rows from this matrix, using the same drawn rows of data for all K specifications.
(4) Estimate the K specifications on the drawn data.
(5) Repeat steps 3 and 4 a large number of times (for example, 500 or 1,000).
(6) For each bootstrapped sample we now have K estimates, one for each specification. Compute what percentage of the resampled specification curves (for example, of the 500 resamples) exhibits an overall test statistic (for example, median effect size) that is at least as extreme as that observed in the real data.

Discussion

Specification curve analysis provides a (partial) solution to the problem of selectively reported results. Readers expecting a judgement-free solution, one where researchers' viewpoints do not influence the conclusions, will be disappointed by this (and any other) solution. Only an expert, not an algorithm, can identify the set of theoretically justified and statistically valid analyses that could be performed and different experts will arrive at different such sets, and hence different specification curves (see Fig. 1). The goal to eliminate subjectivity is unattainable (and not, in our view, desirable).

Specification curve analysis has several limitations. First, as identified by the review team, its default inferential analysis gives equal weight to all included specifications. While all included specifications should be theoretically justified, statistically valid and non-redundant, researchers may nevertheless consider some specifications superior to others and that some should be given greater weight than others. This limitation can be addressed in principle. If desired weights were identified, one could easily modify the three test statistics we propose—(1) median effect across specifications, (2) share of significant effects and (3) aggregated overall *P* value (using the Stouffer method)—by instead computing (1) a weighted median, (2) a weighted proportion of significant effects and (3) a weighted Stouffer test. In practice, we believe it is generally difficult to identify meaningful numerical weights to give each specification.

The second limitation of specification curve analysis is that it cannot realistically include all valid analyses that even a given researcher might be in favour of running, in part because that number can be too large to estimate in full and in part because a researcher may not immediately think of all the analyses they would consider valid to run. Second, because specification curve analysis merely reduces and does not eliminate ambiguity, researchers, as motivated thinkers, will still be inclined to report results that do versus do not further their goals or satisfy expectations. We believe that specification curve analysis will reduce, but not eliminate, this problem.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets used for both demonstrations have been deposited at OSF: https://osf.io/9rvps/

Code availability

The code used to generate all figures and calculations, including those in the Supplementary information, has been deposited at OSF: https://osf.io/9rvps/

Received: 14 December 2018; Accepted: 9 June 2020;
Published online: 27 July 2020

References

1. Leamer, E. E. Let's take the con out of econometrics. *Am. Econ. Rev.* **73**, 31-43 (1983).
2. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, 696–701 (2005).
3. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
4. Glaeser, E. L. Researcher incentives and empirical methods. *NBER Technical Working Paper Series* https://doi.org/10.3386/t0329 (2006).
5. Efron, B. Estimation and accuracy after model selection. *J. Am. Stat. Assoc.* **109**, 991–1007 (2014).
6. White, H. A reality check for data snooping. *Econometrica* **68**, 1097–1126 (2000).
7. Athey, S. & Imbens, G. A measure of robustness to misspecification. *Am. Econ. Rev.* **105**, 476–480 (2015).
8. Sala-i-Martin, X. X. I just ran two million regressions. *Am. Econ. Rev.* **87**, 178–183 (1997).
9. Muñoz, J. & Young, C. We ran 9 billion regressions: eliminating false positives through computational model robustness. *Sociol. Methodol.* **48**, 1–33 (2018).
10. Young, C. & Holsteen, K. Model uncertainty and robustness: a computational framework for multimodel analysis. *Sociol. Methods Res.* **46**, 3–40 (2017).
11. Miguel, E. et al. Promoting transparency in social science research. *Science* **343**, 30–31 (2014).
12. Moore, D. A. Preregister if you want to. *Am. Psychol.* **71**, 238–239 (2016).
13. Bhargava, S., Kassam, K. S. & Loewenstein, G. A reassessment of the defense of parenthood. *Psychol. Sci.* **25**, 299–302 (2014).
14. DellaVigna, S. & Malmendier, U. Paying not to go to the gym. *Am. Econ. Rev.* **96**, 694–719 (2006).
15. Stevenson, B. & Wolfers, J. Economic growth and subjective well-being: reassessing the Easterlin Paradox. *Brookings Pap. Econ. Act.* **2008**, 1–87 (2008).
16. Card, D. & Krueger, A. B. Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania. *Am. Econ. Rev.* **84**, 772–793 (1994).
17. Jung, K., Shavitt, S., Viswanathan, M. & Hilbe, J. M. Female hurricanes are deadlier than male hurricanes. *Proc. Natl Acad. Sci. USA* **111**, 8782–8787 (2014).
18. Malter, D. Female hurricanes are not deadlier than male hurricanes. *Proc. Natl Acad. Sci. USA* **111**, E3496 (2014).
19. Maley, S. Statistics show no evidence of gender bias in the public's hurricane preparedness. *Proc. Natl Acad. Sci. USA* **111**, E3834 (2014).
20. Bakkensen, L. & Larson, W. Population matters when modeling hurricane fatalities. *Proc. Natl Acad. Sci. USA* **111**, E5331 (2014).
21. Christensen, B. & Christensen, S. Are female hurricanes really deadlier than male hurricanes? *Proc. Natl Acad. Sci. USA* **111**, E3497–E3498 (2014).
22. Jung, K., Shavitt, S., Viswanathan, M. & Hilbe, J. M. Reply to Christensen and Christensen and to Malter: pitfalls of erroneous analyses of hurricanes names. *Proc. Natl Acad. Sci. USA* **111**, E3499–E3500 (2014).
23. Bertrand, M. & Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* **94**, 991–1013 (2004).
24. Boos, D. D. Introduction to the bootstrap world. *Stat. Sci.* **18**, 168–174 (2003).
25. Bickel, P. J. & Ren, J.-J. The bootstrap in hypothesis testing. *Proj. Euclid* **36**, 91–112 (2001).
26. MacKinnon, J. G. in *Handbook of Computational Econometrics* (eds Belsley, D. A. & Kontoghiorghes, E. J.) 183–213 (Wiley, 2009).
27. Paparoditis, E. & Politis, D. N. Bootstrap hypothesis testing in regression models. *Stat. Probab. Lett.* **74**, 356–365 (2005).
28. Romano, J. P. Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. P Stat.* **17**, 141–159 (1989).
29. Pitman, E. J. G. Significance tests which may be applied to samples from any populations. *J. R. Stat. Soc.* **4**, 119–130 (1937).
30. Fisher, R. A. *The Design of Experiments* (Oliver and Boyd, 1935).
31. Pesarin, F. & Salmaso, L. *Permutation Tests for Complex Data: Theory, Applications and Software* (John Wiley & Sons, 2010).
32. Ernst, M. D. Permutation methods: a basis for exact inference. *Stat. Sci.* **19**, 676–685 (2004).
33. Flachaire, E. A better way to bootstrap pairs. *Econ. Lett.* **64**, 257–262 (1999).
34. Lancaster, H. Significance tests in discrete distributions. *J. Am. Stat. Assoc.* **56**, 223–234 (1961).

Acknowledgements

The authors received no specific funding for this work.

Author contributions

U.S., J.P.S. and L.D.N. jointly developed the ideas surrounding specification curve analysis and wrote the manuscript. U.S. developed and implemented the inferential approach to specification curve analysis and conducted all analyses.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-0912-z.

Correspondence and requests for materials should be addressed to U.S.

Peer review information Primary handling editor: Stavroula Kousta

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

A formal presentation of the problem we want to solve

Let's consider a relationship of interest between variables x and y, in a context in which other variables, Z, may influence the relationship; y=F(x, Z)+e. For example, x may be education, y may be economic success, and Z may include moderators (e.g., school quality) and/or confounds (e.g., parental education). e consists of orthogonal predictors of y (e.g., luck).

Learning about y=F(x,Z) poses several practical challenges: (i) x and y are often imprecisely defined latent variables (e.g., education and economic success are both imprecisely defined latent variables), (ii) the set of moderators and confounders in Z are often not fully known ex-ante, (iii) Z also contains imprecisely defined latent variables (e.g., school quality is a latent and not precisely defined predictor of economic success), and (iv) the functional form F() is not known. To study y=F(x,Z) researchers must operationalize the underlying constructs. Let's designate the operationalization of a construct Θ, with $\overrightarrow{\Theta}$. Researchers, then, approximate y=F(x,Z) with a specification, a set of operationalizations: $\overrightarrow{y}_{k_y} = \overrightarrow{F}_{k_F}(\overrightarrow{x}_{k_x};\overrightarrow{Z}_{k_Z})$, where k_y, k_F, k_x, and k_Z are indices for single operationalizations of the respective constructs. For example $\overrightarrow{y}_1$ may operationalize 'economic success' with yearly salary, while $\overrightarrow{y}_2$ with private jet seat capacity.

For each construct there are multiple statistically valid, theoretically justified, and non-redundant operationalizations. Their combination leads to what we refer to as the set of reasonable specifications, which, as discussed in the previous section, may be at least somewhat subjective. Designating the total number of valid operationalizations for each construct with n_y, n_x, n_Z and n_F, the total number of reasonable specifications available to study y=F(x,Z) is $N \leq n_x * n_y * n_Z * n_F$.

Let Π be this set of N reasonable specifications, and π be the subset of specifications reported in a paper. Thinking about π as a sample of Π makes it easier to understand the problem Specification Curve analysis attempts to remedy.

By definition, any given $\overrightarrow{y}_{k_y} = \overrightarrow{F}_{k_F}(\overrightarrow{x}_{k_x};\overrightarrow{Z}_{k_Z})$ is considered a valid proxy for y=F(x,Z) and therefore so is the full set of all such proxies: Π. A (i) large, (ii) random, and (iii) independently drawn sample of Π would thus lead to a reasonable estimate of the model of interest: y=F(x,Z). The problem is that π, the sample of specifications reported in a paper, has none of these three properties.

First, it is small, not large. Researchers report a few specifications in any given paper, providing a statistically noisy approximation. Second, it is a curated rather than a random sample. Researchers often choose which specifications to report after knowing the results of these vs other specifications, after knowing how they, reviewers, and audience members respond to different results. Thus, π is chosen by a person seeking academic success, not by a random sampling procedure blind to the consequences of selecting one vs. another specification to report.

Third, and least obvious, the specifications in π are not statistically independent. How much information is there in the fact that a result is obtained across ten rather than just three specifications? It depends on how statistically independent the alternative specifications are. In other words, it depends on how likely it is, under the null, that one specification in π will show an effect if another specification in π already does. Currently the statistical independence of robustness results is not considered, neither formally nor informally. Results are labeled as robust without considering how likely the results are to coincide by chance alone.

Specification Curve analysis addresses all three of these problems. First, it generates a much larger π, where 100s or even 1000s of specifications are reported. This increases statistical efficiency by reducing specification noise. It also makes transparent the existence of such noise,

and allows for readers to determine its nature (i.e., which operationalization decisions are vs. are not consequential). Second, Specification Curve analysis generates a π with fewer arbitrary inclusion decisions, and thus more closely approximates a random sample of Π. When using Specification Curve analysis we can more legitimately consider π as an approximation of y=F(x,Z), though for the sampling to be even closer to random, it would need to be performed by researchers who are blind to the consequence of choosing one vs another specification. Third, Specification Curve analysis allows statistical inference that takes into account the statistical dependence across alternative specifications in π.

The null hypothesis that the true effect of x on y is zero for all specifications is thus $H_0: \frac{d(\vec{F}_{k_F})}{d(\vec{x}_{k_x})} = 0$, $\forall\ \pi_k$ in Π, where π_k indexes the valid operationalizations in Π. For example, considering the special (though quite general) case of a general additive model where $F(x,Z)=f_x(x)+f_z(z)+f_{xZ}(xZ)$, the null is $H_0: \frac{d(\vec{f}_{x_{k_F}})}{d(\vec{x}_{k_x})} = \frac{d(\vec{f}_{xz_{k_F}})}{d(\vec{x}_{k_x})} = 0$, $\forall\ \pi_k$ in Π, and $\forall$ observable x .