Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications

Uri Simonsohn
Universitat Ramon Llull
ESADE Business School
urisohn@gmail.com

Joseph P. Simmons
University of Pennsylvania
The Wharton School
jpsimmo@wharton.upenn.edu

Leif D. Nelson
University of California, Berkeley
Haas School of Business
Leif_nelson@haas.berkeley.edu

Abstract: Empirical results hinge on analytic decisions that are defensible, arbitrary, and motivated. These decisions probably introduce bias (towards the narrative put forward by the authors), and 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: (i) identifying the set of theoretically justified, statistically valid, and non-redundant specifications, (ii) displaying the results graphically, allowing readers to identify consequential specifications decisions, and (iii) conducting joint inference across all specifications. We illustrate the usefulness of this technique by applying specification curve to three findings from two different papers, one investigating discrimination based on distinctively black names, and the other investigating the effect of assigning female vs. male names to hurricanes). Specification curve reveals that one finding is robust, one is weak, one is not robust at all.

Online Supplement: http://urisohn.com/sc_supplement.pdf

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 analyze, and so on.

When reading the results of a study, people want to learn about the true relationship being analyzed, but this requires that the analyses that are reported are representative of the set of valid analysis 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 (Leamer, 1983). A probably bigger, 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 (Glaeser, 2006; Ioannidis, 2005; Leamer, 1983; Simmons, Nelson, & Simonsohn, 2011). 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 of) "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 (see e.g., Figure 2). This enables readers to visually identify both the 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.

In Section 2 we present in more detail the problem we are trying to solve. In Section 3 we present the problem more formally. In Section 4 we review previous solutions and identify what we see as our two key contributions to that literature. In Section 5 we introduce Specification Curve Analysis in detail. In this section, we use Specification Curve Analysis to analyze three findings in two published papers, one examining if hurricane names influence people's inclination to protect against them (Jung, Shavitt, Viswanathan, & Hilbe, 2014a), the other reporting an audit study of racial discrimination in the job market (Bertrand & Mullainathan, 2004). In Section 6 we explain how to apply Specification Curve Analysis to non-experimental data, where the predictor of interest could be correlated with other predictors.

2. An intuitive presentation of the problem we want to solve

To analyze 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 instead guided 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 neither theory-based nor beliefs-based.

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 Specification Curve in particular.

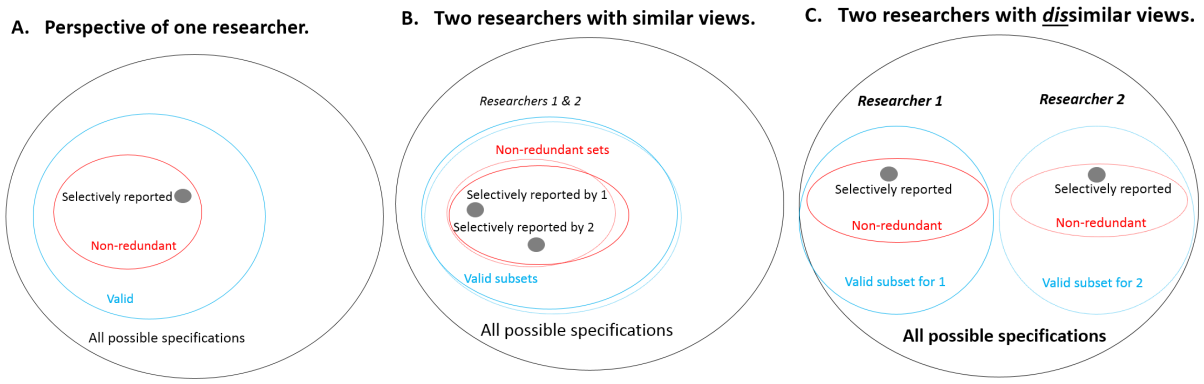
For example, a researcher interested in the causal effect of raising children on adult happiness *should* control for the marital status of the adults. Because married adults are more likely to have children than unmarried ones, the estimates of the happiness effect of raising children will (partially) include the separable effect on happiness of having a spouse (Bhargava, Kassam, & Loewenstein, 2014). Thus, reporting results with and without controlling for marital status may be interesting and informative, but it does not constitute and exercise in robustness because they 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 period (DellaVigna & Malmendier, 2006). 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 (see e.g., Stevenson & Wolfers, 2008). In a study on labor participation we should expect robustness tests on what is used as the full-time equivalence of someone working part-time (see e.g., Card & Krueger, 1994).

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. Panel A 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 circle), some of which are redundant with one another (e.g., log transforming x using log(x+1) or using log(x+1.1)).

The set of reasonable specifications (the red circle) includes only the non-redundant alternatives (e.g., *either* log(x+1) *or* log(x+1.1), but not both).

Figure 1. Sets of possible specifications as perceived by researchers.



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 (i.e., because different researchers draw different circles), 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 circles), or merely in which few analyses the researchers reported (different gray dots).

3. 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 $\overleftrightarrow{\theta}$. Researchers, then, approximate y=F(x,Z) with a specification, a set of operationalizations: $\overleftrightarrow{y}_{k_y} = \overleftrightarrow{F}_{k_F}(\overleftrightarrow{x}_{k_x};\overleftrightarrow{Z}_{k_Z})$, where k_y, k_F, k_x, and k_Z are indices for single operationalizations of the respective constructs. For example $\overleftrightarrow{y}_1$ may operationalize 'economic success' with yearly salary, while $\overleftrightarrow{y}_2$ 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 $N \leq n_x * n_y * n_Z * n_F$ [1].

[1] The inequality sign is used because some combinations of operationalizations may be conceptually invalid, or practically unattainable. E.g., there may a covariate which does not vary within a subgroup, and thus the inclusion of such covariate, and the focusing on that specific subgroup is not a valid combination of operationalizations.

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 $\vec{y}_{k_y} = \vec{F}_{k_F}(\vec{x}_{k_x};\vec{Z}_{k_Z})$ 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 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). 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.

4. Existing approaches

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 with multiple columns, where each column captures a different specification, allowing readers to compare results across specifications.[2] We can think of Specification Curve analysis as an extension and formalization of that approach, one that dramatically reduces the room for selective reporting (from gray dot to red circle in Figure 1).

There have been a few 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 (e.g., when deciding between a quadratic vs cubic polynomial (Efron, 2014)). Another assesses if the best fitting model among a class of models fits better than expected by chance

[2] Gelbach (2016) has an interesting critique of how the multiple-columns are commonly interpreted, explaining how the order in which covariates are added can influence their apparent relative importance.

(White, 2000). A third proposal consists of reporting the standard deviation of point estimates across a few, carefully chosen alternatives specifications (Athey & Imbens, 2015). A fourth approach is known as "extreme bounds analysis," 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 (Leamer, 1983), or if a weighted average of the t-test in each model is itself statistically significant (Sala-i-Martin, 1997). The most recent proposal is by Young and Holsteen (2015) who, as we do here, propose the estimation of a large number of specifications, going beyond just covariates to include functional form and regression model, and who propose plotting the distribution of results obtained across specifications (see also Muñoz and Young (2018) and Young and Holsteen (2017)).

Among other differences with *all* of these approaches, Specification Curve Analysis: (i) helps identify the source of variation in results across specifications via a descriptive specification curve (see Figure 2), and (ii) provides a formal joint significance test for the family of alternative specifications, derived from expected distributions under the null. We are not aware of any existing approach that provides either feature.

A non-statistical approach to dealing with selective reporting consists of pre-analyses plans (Miguel et al., 2014; Moore, 2016). Specification Curve Analysis complements this approach, allowing researchers to pre-commit to running the entire set of specifications they consider valid, rather than 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 vs. the other conclusion by pre-committing to one vs.

another *valid* analysis, while Specification Curve allows researchers to learn which specifications the conclusion hinges on.

5. Conducting Specification Curve Analysis

Specification Curve Analysis is carried out in three main steps. First, define the set of reasonable specifications to estimate. Second, estimate all specifications and report the results in a descriptive specification curve. Third, conduct joint statistical tests using an inferential specification curve.

We demonstrate these three steps by applying specification curve to two published articles with publicly available raw data. One reports that hurricanes with more feminine names have caused more deaths (Jung et al., 2014a). We selected this paper because it led to an intense debate about the proper way to analyze the underlying data (Bakkensen & Larson, 2014; Christensen & Christensen, 2014; Jung et al., 2014a; Jung, Shavitt, Viswanathan, & Hilbe, 2014b; Maley, 2014; Malter, 2014), providing an opportunity to assess the extent to which Specification Curve analysis could aid such debates. The second article reports a field experiment examining racial discrimination in the job market (Bertrand & Mullainathan, 2004). We selected this highly cited article because it allowed us to showcase the range of inferences 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 section 6 we explain how to conduct inference in Specification Curve analysis when this is not the case (e.g., when data do not originate in an experiment).

5.1 Step 1. Identify the set of specifications

The set of reasonable specifications can be generated by (i) enumerating all of the data analytic decisions necessary to map the scientific hypothesis or construct of interest onto a

statistical hypothesis, (ii) enumerating all the reasonable alternative ways a researcher may make those decisions, and finally (iii) 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 (Jung et al., 2014a) 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 analyze, 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 just-as-reasonable alternatives. The combination of all operationalizations we considered valid and non-redundant make up our red circle, a set of 1,728 reasonable specifications (see Supplement 1 for details).

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 analyze	Excluded two outliers with the most deaths	Dropping fewer outliers (zero or one); dropping storms with extreme values on a predictor variable (e.g., 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	OLS 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 (e.g., wind or category) instead of damages

5.2 Step 2. Estimate & Describe Results

The descriptive specification curve serves two functions: displaying the distribution of estimates that are obtained through alternative reasonable specifications, and identifying which analytic 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 reports 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 are statistically significant ($p<.05$). The point estimates range from -1 to +12 additional deaths.[3]

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: (i) not taking into account the year of the storm, (ii) operationalizing severity of the storm by the log of damages, (iii) conducting an OLS regression, etc. A researcher motivated to show a negative point estimate would be able to report *twenty* different specifications that do so, but the specification curve shows that a negative point estimate is atypical.

Following the publication of the hurricanes paper, PNAS published four letters/critiques proposing alternative specifications under which the impact of hurricanes name on fatalities goes away (Bakkensen & Larson, 2014; Christensen & Christensen, 2014; Maley, 2014; Malter, 2014). In particular, the critiques argued that outlier observations, with more than 100 deaths, should be excluded (Christensen & Christensen, 2014; Maley, 2014), that the regression should include an interaction between intensity of the hurricane and dollar damages as a predictor (Malter, 2014), and that dollar damages should not be included as a predictor at all (Bakkensen & Larson, 2014).

[3] To make point estimates for the continuous and discrete measures of femininity comparable, 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.

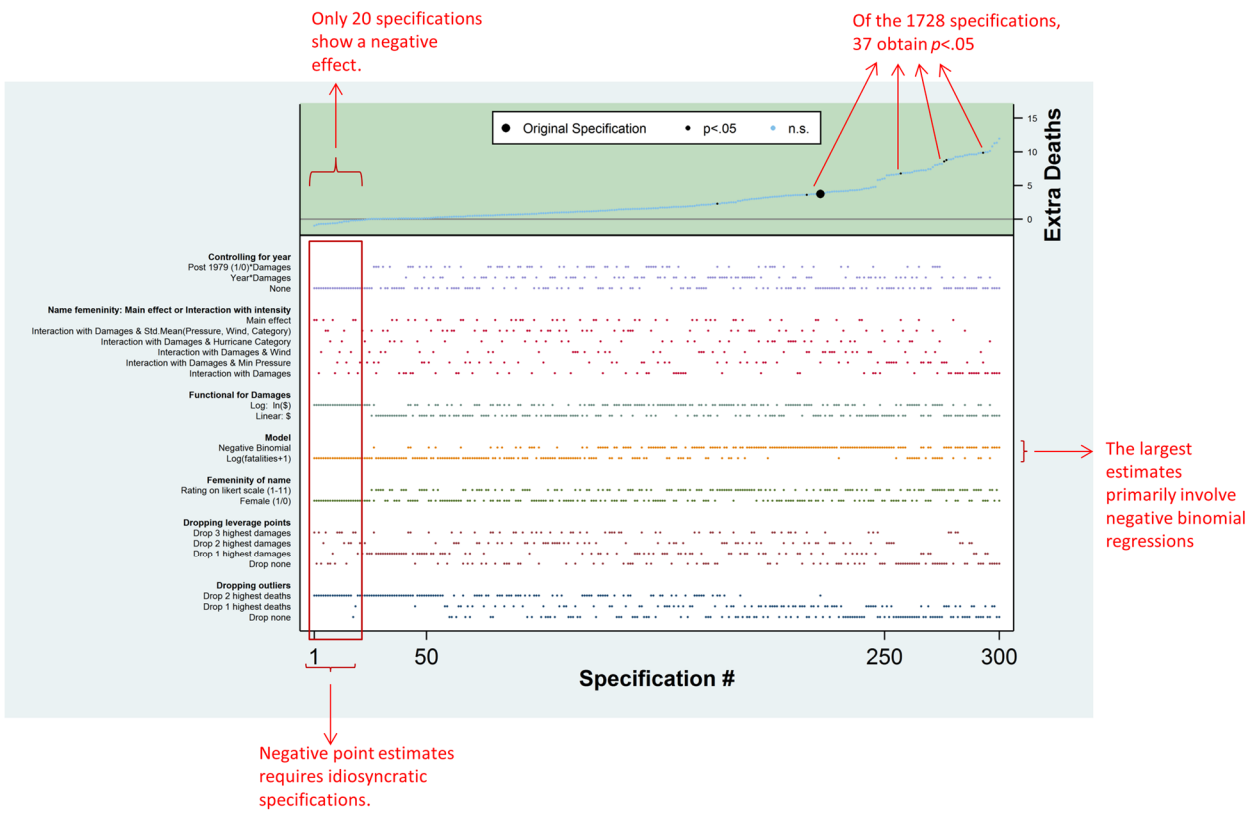


Figure 2. *Descriptive Specification Curve.* Each dot in the top panel (green area) depicts the marginal effect, estimated at sample means, of a hurricane having a female rather than male name; the dots vertically aligned below (white area) indicate the analytic decisions behind those estimates. A total of 1728 specifications were estimated; the figure depicts the 50 highest and lowest point estimates, and a random subset of 200 additional ones.

Returning to Figure 1, this appears to be a Panel C situation. Original authors and critics disagree on the set of valid specifications to run. The specification curve results from Figure 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 hurricanes 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 treated in the original, modifying virtually any arbitrary analytical decision renders the original effect nonsignificant. Readers need not take a position on whether it does or does not make sense to include a damages x pressure interaction in the model to determine if the original findings are robust.

Figure 2 shows that PNAS could have published nearly 1,700 letters showing individual specifications that make the effect go away (without deviating from the original red circle). It also could have published 37 responses with individual specifications showing the robustness of the findings. It would be better to publish a single specification curve in the original paper.

5. 3. 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 $\overleftrightarrow{y}_{k_y} = \overleftrightarrow{F}_{k_F}(\overleftrightarrow{x}_{k_x};\overleftrightarrow{Z}_{k_Z})$ 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) that would be expected if all specifications had an effect of zero. The third test statistic is similar to the second, but instead of discretizing each *p*-value into a significant vs non-significant dichotomous variable, and counting them, it aggregates all of them in a continuous fashion, by averaging the Z-value associated with each (e.g., Z=1.96 for *p*=.05), as in Stouffer's method, and testing whether the average Z value 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 (i.e., 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 under-the-null bootstrapping, where the observed data are modified so that the null hypothesis is known to be true, and then random samples of the modified data are drawn, the test statistic of interested computed on each of them, and the resulting distribution is the estimated distribution of the test-statistic under the null (Bickel & Ren, 2001; Boos, 2003; MacKinnon, 2009; Paparoditis & Politis, 2005; Romano, 1989).

The implementation of under-the-null bootstrapping 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 bootstrapping for experiments in

this section. Bootstrapped Specification Curve analysis for non-experiments is explained in section 6.

Because Specification Curve Analysis relies on bootstrapping 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, e.g., exhibiting a 14% change of obtaining $p<.05$ when the null is true, instead of the nominal 5%, by relying on bootstrapping based inference, the false-positive rate will be corrected and returned to 5%. In supplement 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.

5.4 Inference in the Hurricanes example

Bootstrapping experimental data under the null is simple and intuitive, as it involves shuffling the column(s) with the randomly assigned variable(s) (Ernst, 2004; Fisher, 1935; Pesarin & Salmaso, 2010; Pitman, 1937). In the case of the hurricanes paper, one shuffles the hurricane's name.[4] The shuffled datasets maintain all the other features of the original one (e.g., collinearity, time trends, skewness, etc.) except we now know there is no link between (shuffled) names and

[4] The hurricanes dataset consists of a natural rather than a lab experiment and there is thus some ambiguity as to what random procedure is the bootstrapping supposed to emulate. Hurricane names follow alphabetical order and alternate gender. If one were to keep that aspect of the naming procedure fixed across bootstrapped samples, there would be no room for a randomization test because every resample would assign the observed gender to all hurricanes. Under both the null and the alternative, however, this aspect of the naming procedure is inconsequential, because hurricanes are assumed to be independent events whose severity does not depend on the severity of past hurricanes, except for possible time trends (e.g., high-danger areas having population increases over the years). Thus we resample by shuffling the name column, asking "if hurricane names were entirely unrelated to their consequences, how surprising would results at least as extreme as those observed be?" rather than "if hurricanes came in the same sequence as they came, and names came in the same gender sequence as they were given, how surprising would the results be?" This second question cannot be answered, as we cannot create counterfactuals for it.

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 (Ernst, 2004; Pesarin & Salmaso, 2010), that any hurricane could have received any name (see footnote 4). The resulting *p*-values are hence 'exact,' not dependent on distributional assumptions.

Sign. Because many of the different specifications are similar to each other (e.g., the same analysis conducted with an outlier included vs 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, 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 sign,' and we plot results as having the dominant or non-dominant sign, rather than positive or negative sign. This allows us to visually capture 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: (i) median effect size, (ii) share of results that are significant, and (iii) the average Z-score transformation of each *p*-value (Stouffer's method).

For example, in the observed hurricanes data, 37 of the 1728 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 $p = 425/500 = .85$.

5.5. Second example: Bertrand & Mullainathan (2004)

Having gone through the three steps for carrying out Specification Curve Analysis with our first example, we move on to our second example (Bertrand & Mullainathan, 2004), a field experiment in which researchers used fictitious resumes to apply to real jobs using randomly assigned names that were distinctively Black (e.g., Jamal or Lakisha) or not (e.g., Greg or Emily).

The authors of this article arrived at two key conclusions: applicants with distinctively Black names (i) were less likely to be called back, and (ii) benefited less from having a higher quality resume. 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 the finding (ii). In particular, we considered two alternative regression models (OLS vs probit), three alternative samples (men and women, only men, and only women), and fifteen alternative definitions of resume quality. These resulted in a set of 90 reasonable specifications. We justify this set of specifications and report the descriptive specification curves in Supplements 2 and 3, respectively.

Figures 3B and 3C display the inferential specification curve results for these findings. Starting with the core finding that distinctively Black names had lower callback rates (Panel C) 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.

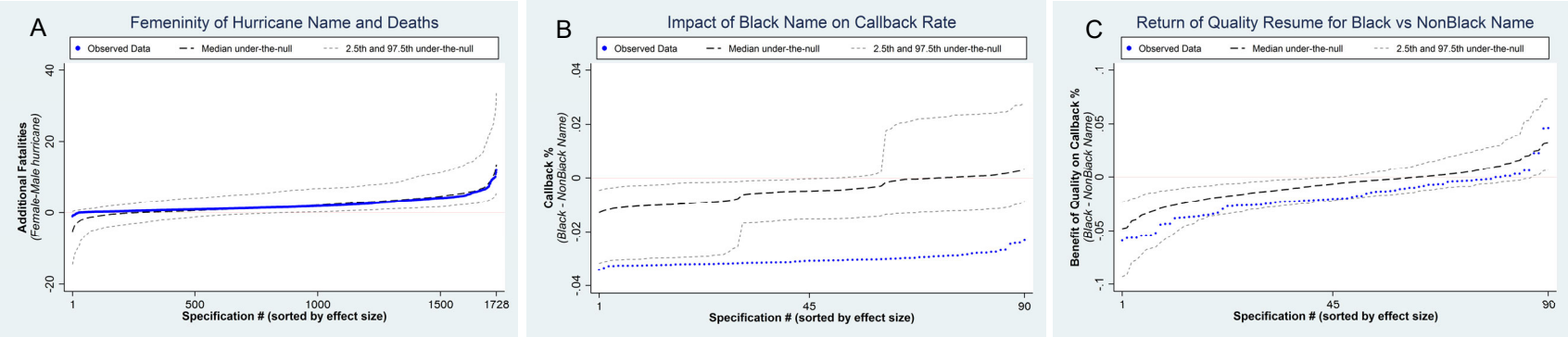


Figure. 3. Observed and expected under-the-null specification curves for the hurricanes (**A**) and racial discrimination studies (**B,C**). The expected curves are based on 500 shuffled samples, where the key predictor in each dataset (hurricane and applicant name respectively) is shuffled. All specifications are estimated on each shuffled sample (1,728 specifications 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 (e.g., the median smallest estimate, the median 2nd smallest estimate, etc.). Specification curves under the null are typically not symmetric around zero (see main text). The blue dots depict the specification curve for the observed data.

Test Statistic Used	Observed Result	***p*-value** *(% of shuffled samples with as or more extreme results)*
	Example 1. Female Hurriance Names	
(1) Median effect size	1.56 additional deaths	p = .536
(2) Share of significant results	37 out of 1728 specifications	p = .850
(3) Aggregate all p-values	Stouffer Z = 28.47	p = .512
	Example 2a. Black names receive fewer callbacks	
(1) Median effect size	3.1 *pp* fewer calls	p < .002
(2) Share of significant results	85 out of 90 specifications	p < .002
(3) Aggregate all p-values	Stouffer Z = 35.71	p < .002
	Example 2b. Black names benefit less from quality CV	
(1) Median effect size	2.0 *pp* smaller benefit	p = .162
(2) Share of significant results	13 out of 90 specifications	p = .032
(3) Aggregate all p-values	Stouffer Z = 9.22	p = .126

Table 2. Joint tests for inferential specification curves in the two examples. *pp*: percentage-points. Each overall *p*-value is computed by the proportion of shuffled samples leading to at least as extreme a test statistic 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 (Lancaster, 1961). When no shuffled sample is as extreme as the observed, we report p<.002, for our estimate is that it is less frequent than 1 out of the 500 samples we collected. But estimates that small 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 it is also obtained via bootstrapping, 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).

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

6. Optional Reading: Inference with non-experimental data

Bootstrapping for hypothesis testing purposes involves modifying the dataset so that the null is true, and then computing the test statistic on random samples from that modified dataset (Bickel & Ren, 2001; Boos, 2003; MacKinnon, 2002; Paparoditis & Politis, 2005). For example, to do bootstrapped inference on the equality of means between two samples, one can add the mean difference between samples to every observation in one sample, so that the means are now identical, and then bootstrap by drawing from these modified data. where the null of equality of means is true.

For regression models there are two main approaches to modify the data and impose the null. Forcing the null and then shuffling residuals across rows, or alternatively, forcing the null and then sampling rows at random (see e.g., Flachaire, 1999). For Specification Curve analysis we rely on the latter, sometimes referred to as "case resampling", for two main reasons discussed in this footnote.[5]

Specifically, for each specification one first estimates the model with the observed data, say estimating the parameters a,b and c in $y=a + b\,x + cz + e$. Then one forces the null on the data by creating a new dependent variable, y*, that subtracts the estimated effect of x on y. So, $y^*=y - \hat{b}x$. Now we have a model where we know the null is true. That is, we know that $b^*=0$ in this model: $y^*=a + b^*\,x + cz + e$. To generate a distribution of expected results, of expected $\hat{b}$ under the null, one samples with replacement rows of data, but where y^* rather than y is the dependent variable. Each resample has the same sample size as the original. The distribution of $\hat{b}$s across the resamples

[5] Residual based bootstrapping has the general problem that it assumes homoskedasticity, in that any residual can be paired with any row of data, thus assuming the same error distribution for each observation. For Specification Curve Analysis in particular there is the additional problem that if specifications differ in the number of complete observations, when residuals are shuffled, the number of missing observations will increase (as the missing residual from one row, will be assigned to a row which did not have missing data). We thus don't think residual bootstrapping is a valid approach to bootstrapping for Specification Curve Analysis.

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 6 steps:

1. Estimate all K specifications with the observed data, $\vec{y}_{k_y} = \vec{F}_{k_F}(\vec{x}_{k_x};\vec{Z}_{k_Z})$. These will result in K different point estimates: $\hat{b}_{k.}$ with k=1…K. Note that $\vec{y}_{k_y}$ 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, $y_k^* = y_k - \hat{b}_{k.}$ Even if there are less than K different y_k, there will be K different y^*_k because $\hat{b}_k$ is different across specifications and thus so is $y^*_k = y - b^*_k x_k$. 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 & 4 a large number of times (e.g., 500 or 1000).

6. For each bootstrapped sample we now have K estimates, one for each specification. Compute what percentage of the K specification curves exhibit a test statistic that is at least as extreme as that which is observed in the real data.

7. Conclusions

Specification Curve Analysis provides a (partial) solution to the problem of selectively reported results. Readers expecting a judgment-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 Figure 1). The goal to eliminate subjectivity is unattainable (and not, in our view, desirable).

When different researchers arrive at different conclusions from the same data, the disagreement may reflect profoundly different views on what they consider to be theoretically justified or statistically valid analyses, or they may reflect superficial and arbitrary decisions on how they operationalized those same views they share (blue vs red circles in Figure 1). Specification Curve analysis helps identify the subset of disagreements that belong to the second category, and helps us reach consensus on that second subset. For the first set, the solution is not more or different data analysis, but rather, more or different theories (or training).

Something that is unsatisfying about Specification Curve is that it will never include *all* valid analyses even a given researcher could be in favor of running. Not only because sometimes the number is too big to be estimated in full and we must settle for a random subset, but also because one cannot in one sitting think of all possibilities. Looking back at one's own specification curve one may think "I guess I could have also run a probit, not just a logit" or "maybe I should also evaluate robustness to the size of the time window" or "I just thought of a really clever way to operationalize resume quality," etc.

The set of operationalizations one could think of and deem valid is sometimes, perhaps often, infinite, while the set of operationalizations one did consider valid at a given point in time, is never infinite. The only solace for this imperfection is that it is less imperfect with Specification Curve Analysis than it is with any alternative. While the 1,728 specifications for the impact of hurricane name on fatalities is not infinite, it is orders of magnitude larger than the number of specifications

typically reported in papers (1 to 20 say). Moreover, it is a set that contains much less post-hoc selection based on results (gray dot vs. red circle in Figure 1). It is harder to undetectably selectively report families of analyses than it is to do so with individual combinations. In sum, Specification Curve is an imperfect solution to the problem of selective reporting, but it represents a big improvement to the status quo.

References

Athey, S., & Imbens, G. (2015). A Measure of Robustness to Misspecification. *American Economic Review: Papers & Proceedings, 105*(5), 476-480.

Bakkensen, L., & Larson, W. (2014). Population matters when modeling hurricane fatalities. *Proceedings of the National Academy of Sciences of the United States of America, 111*(50), E5331.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991-1013.

Bhargava, S., Kassam, K. S., & Loewenstein, G. (2014). A reassessment of the defense of parenthood. *Psychological Science, 25*(1), 299-302.

Bickel, P. J., & Ren, J.-J. (2001). The bootstrap in hypothesis testing. *Lecture Notes-Monograph Series, State of the Art in Probability and Statistics, 36*, 91-112.

Boos, D. D. (2003). Introduction to the bootstrap world. *Statistical Science, 18*(2), 168-174.

Card, D., & Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review, 84*(4), 772-793.

Christensen, B., & Christensen, S. (2014). Are female hurricanes really deadlier than male hurricanes? *Proceedings of the National Academy of Sciences, 111*(34), E3497-E3498.

DellaVigna, S., & Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review, 96*(3), 694-719.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American statistical association, 109*(507), 991-1007.

Ernst, M. D. (2004). Permutation methods: a basis for exact inference. *Statistical Science, 19*(4), 676-685.

Fisher, R. A. (1935). The Design of Experiments (8th: Oliver and Boyd, Edinburgh.

Flachaire, E. (1999). A better way to bootstrap pairs. *Economics Letters, 64*(3), 257-262.

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics, 34*(2), 509-543.

Glaeser, E. L. (2006). Researcher incentives and empirical methods. *NBER Technical Working Paper Series*(329).

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine, 2*(8), 696-701.

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014a). Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 201402786.

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014b). Reply to Christensen and Christensen and to Malter: Pitfalls of erroneous analyses of hurricanes names. *Proceedings of the National Academy of Sciences, 111*(34), E3499-E3500.

Lancaster, H. (1961). Significance Tests in Discrete Distributions. *Journal of the American statistical association, 56*(294), 223-234.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 31-43.

MacKinnon, J. G. (2002). Bootstrap inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'économique, 35*(4), 615-645.

MacKinnon, J. G. (2009). Bootstrap hypothesis testing. *Handbook of computational econometrics, 183*, 213.

Maley, S. (2014). Statistics show no evidence of gender bias in the public's hurricane preparedness. *Proceedings of the National Academy of Sciences, 111*(37), E3834-E3834.

Malter, D. (2014). Female hurricanes are not deadlier than male hurricanes. *Proceedings of the National Academy of Sciences, 111*(34), E3496-E3496.

Miguel, E., Camerer, C. F., Casey, K., Cohen, J., Esterling, K., Gerber, A., . . . Imbens, G. (2014). Promoting Transparency in Social Science Research. *Science, 343*(6166), 30-31.

Moore, D. A. (2016). Preregister if you want to. *American Psychologist, 71*(3), 238.

Muñoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology, 48*(1), 1-33.

Paparoditis, E., & Politis, D. N. (2005). Bootstrap hypothesis testing in regression models. *Statistics & probability letters, 74*(4), 356-365.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*: John Wiley & Sons.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society, 4*(1), 119-130.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics, 17*(1), 141-159.

Sala-i-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, 178-183.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359-1366.

Stevenson, B., & Wolfers, J. (2008). Economic growth and subjective well-being: Reassessing the Easterlin Paradox. *Brookings Papers on Economic Activity, 2008*(1), 1-87.

White, H. (2000). A reality check for data snooping. *Econometrica, 68*(5), 1097-1126.

Young, C., & Holsteen, K. (2015). Model Uncertainty and Robustness A Computational Framework for Multimodel Analysis. *Sociological Methods & Research*, 0049124115610347.

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research, 46*(1), 3-40.