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 included in 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 with three findings, from two different papers (one paper is on discrimination based on distinctively black names, the other on the impact of hurricanes with male vs female names). One finding proves robust, one weak, one not robust at all.

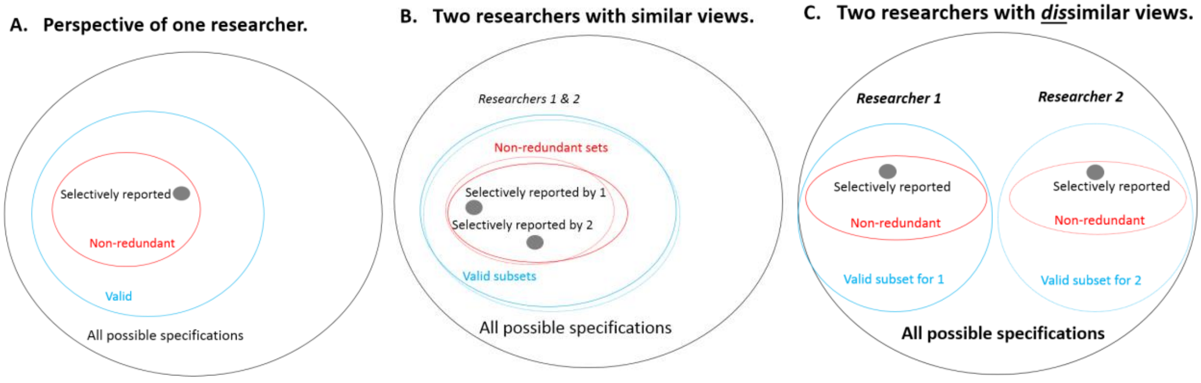
The empirical testing of scientific hypotheses requires data analysis, but data analysis is not straightforward. To convert a scientific hypothesis into a testable prediction, researchers must make a number of data analytic decisions, many of which are both arbitrary and defensible. For example, researchers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

When reading a new study, people want to learn about the true relationship being analyzed, but this requires the analyses that are reported to be representative of the set of valid analysis which could have been (validly) run. One problem is the possibility that the results may hinge on an arbitrary choice by the researcher (Leamer, 1983). A probably bigger problem is that people in general, and researchers in particular, are more likely to favor evidence consistent with the claims they are trying to make than those that are inconsistent with such claims (Glaeser, 2006; Ioannidis, 2005; Leamer, 1983; Simmons, Nelson, & Simonsohn, 2011). In other words, specifications shown in academic papers are noisier than reported (the standard errors do not include variation in estimates arising from alternative specifications), and are probably biased because they are chosen post-hocly.

In this article we introduce Specification Curve Analysis as a way to mitigate these problems of noise and bias that arise from selectively reporting only a few specifications. 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.

Figure 1 helps to illustrate what it means, and what it does *not* mean, to report the results of all 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.


A. Perspective of one researcher.
Selectively reported
Non-redundant
Valid
All possible specifications
B. Two researchers with similar views.
Researchers 1 & 2
Non-redundant sets
Selectively reported by 1
Selectively reported by 2
Valid subsets
All possible specifications
C. Two researchers with dissimilar views.
Researcher 1
Researcher 2
Selectively reported
Non-redundant
Valid subset for 1
Selectively reported
Non-redundant
Valid subset for 2
All possible specifications

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, instead, will *facilitate* those debates.

Panels B and C in Figure 1 depict researcher disagreements. Panel B considers two researchers who, despite high ex-ante agreement regarding the set of valid specifications, ex-post selectively report different results, different grey dots. With specification-curve analysis both researchers report very similar sets of analyses (very similar red circles). Their conclusions might still be different, but the origin of the disagreement will be transparently revealed

Panel C depicts two researchers with substantial ex-ante disagreement. Most specifications considered valid by Researcher 1 are deemed invalid by Researcher 2, and vice versa. This may occur if Researchers 1 and 2 base their analyses on different theories (e.g., behavioral vs neoclassical economics), disagree on the operationalization of those theories (e.g., the reference point for reference-dependent preferences), or on the appropriateness of one vs. another statistical procedure (e.g., reduced form vs. structural estimation, or, whether an identifying assumption is credible or not).

Despite having non-overlapping sets of reasonable specifications, specification-curve analysis can help Researchers 1 and 2 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).

2. Formalizing the goals of specification-curve analysis

Let's consider a relationship of interest between variables x and y, in a context where 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 includes 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., both education and economic success are), (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_x, k_Z, and k_F are indexes 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-yet seat-capacity (censored at 0).

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 (a necessarily at least somewhat subjective set; see Figure 1). 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].

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 the problem specification-curve analysis seeks to address easy to understand.

By definition, any given $\overleftrightarrow{y}_{k_y} = \overleftrightarrow{F}_{k_F}(\overleftrightarrow{x}_{k_x};\overleftrightarrow{Z}_{k_Z})$ is considered a valid proxy for y=F(x,Z) and therefore so is the full set of all such proxies: Π. A large, random, independently drawn sample of Π would thus also be a valid proxy of 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 handful specifications in any given paper, providing a statistically noisy approximation. Second, it is a curated rather than a random sample. Researchers 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. 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. It depends in other words, on how likely it is, under the null, that one specification in π will show

[1] The inequality sign is used because some combinations of operationalizations may be conceptually invalid, or practically unattainable thus cannot be materialized onto a specification.

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, 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 determining its nature: which operationalization decisions are vs. are not consequential. Second, specification-curve analysis generates a π with fewer arbitrary inclusion decision, 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.

3. Existing approaches

There is a long tradition of considering robustness to alternative specifications in social science. The norm in economics, political science, and other fields consists of reporting regression results in tables with multiple columns, where each column captures 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 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 (White, 2000). A third proposal consists of reporting the standard deviation of point estimates across a few 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.

Among other differences with *all* of these approaches, Specification-Curve Analysis, (i) provides a step-by-step guide to generate the set or reasonable specifications, (ii) aids in the identification of the source of variation in results across specifications via a descriptive specification curve (see Figure 2), and (iii) 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 any of these three features.

A non-statistical approach to dealing with selective reporting consists of pre-analyses plans (Miguel et al., 2014). 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.

4. Arbitrary vs theory based decisions

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

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. If a specification does not make sense theoretically or statistically it does not belong in a robustness test in general, nor in specification-curve in particular. In this section we discuss what kinds of alternative specifications do and do not constitute robustness tests for specification-curve analysis purposes.

4.1 Covariates are usually not about robustness

In most settings, which covariates to include in a regression model is a decision guided by theory and/or context. The decision of what to control for is not arbitrary; economists can usually easily determine which combination of covariates is the one that leads to an estimate that most directly and adequately answers the question of interest.[2]

[2] Sometimes the set of covariates considered relevant is too large for the size of the data at hand, and machine learning or other procedures can be used to choose the combination of covariates to include (see e.g.,Belloni,

Nevertheless, it has long been customary in economics to report regression results with different sets of covariates. This is a useful practice we wished other social sciences imitated, but it is important to realize that reporting regression results that exclude covariates the researchers believe belong in the model is *not* a robustness exercise (Gelbach, 2016 proposes a more sophisticated approach to assess robustness to incrementally added covariates). The (often unspoken) goal is instead assessing the impact of *unobserved* covariates on the parameter estimate of interest in their favored specification. The heuristic is: "not controlling for *observable* heterogeneity does not alter the results very much, so the unobservable heterogeneity we are unable to control for presumably also is not altering very much." For example, DellaVigna and Malmendier (2006) report regressions predicting whether people remain enrolled in a gym based on the membership plan they purchased. They report results controlling and not controlling for gender and age (among other demographics; see their Table 6). The model that does control for such variables is strictly preferred from a theoretical perspective as it accounts for potential confounds on which plan they selected. The results without covariates are not there to assure readers that two equally valid ways to analyze the data lead to the same conclusion; indeed, the authors point out that the model without covariates is biased. Instead, they report the results without covariates to assure readers that, since excluding observed heterogeneity does not much alter the results, presumably the unobserved heterogeneity they are unable to control for also does

Chernozhukov, & Hansen, 2014). That is not a counter-example to our claim that covariate selection is generally driven by theory. It is easiest to see this once one considers what robustness would look like in this case. To assess the robustness of selecting covariates this way, a researcher would not choose another set of covariates and apply the same datamining tool to reduce their dimensionality, rather, she would start with the same set of theoretically motivated covariates and use a different datamining tool, or a different operationalization of the same datamining tool (e.g., different overfitting penalty). Using datamining to select covariates is not about covariate selection, it is about dimensionality reduction. It is not about "what should I control for?" but about "what subset of the things I would like to control for *can* I?"

not alter the results very much. Regressions with and without a certain set of covariates are not different answers to the same question, they are different answers to different questions.

Previous attempts to enhance the reporting of robustness in regression analysis have focused primarily on the set of covariates used (Leamer, 1983; Sala-i-Martin, 1997; Young & Holsteen, 2015). We believe this focus is misplaced, and that this misplaced focus at least partially explains why economists have not used the tools put forward in those articles. The decision of which covariates to include is usually guided by theory and hence should not usually be part of robustness tests, which are designed to assess the impact of decisions not guided by theory.

4.2 Alterative arbitrary operationalizations are about robustness

Unlike which covariates to include, many decisions necessary to estimate a model are arbitrary, as they are as defensible as many other alternatives are. These are the kinds of decisions that results should be robust to. Let's return to DellaVigna and Malmendier (2006)'s gym study. They were interested in estimating the impact of choosing a monthly (rather than annual) membership plan on the probability of remaining enrolled in the gym in the long term. What's "long term"? There is no theoretically correct answer to that question. The authors arbitrarily set 15 months as "long term," but isn't 16 months, or 26 months just as valid an answer? Of course. This is precisely why the authors report the results defining long term also as 16, and also 26 months. *These* are robustness tests. We do not want results to hinge on the arbitrary definition that long-term is 15 rather than 26 months. More generally, econometric analyses often require defining a "before" and "after" period and theory is rarely sufficiently precise to determine how long these periods are. Robustness in general and specification-curve analysis in particular should examine the impact of alternative definitions of "before" and "after."

Another common arbitrary decision is exactly what operationalization of a variable to use. For example, when studying the impact of income on well-being one can measure the latter using a life-satisfaction scale, or a happiness scale. Neither is strictly more appropriate theoretically than the other, hence for robustness Stevenson and Wolfers (2008) report cross-country regressions with one and then the other dependent variable. Similarly, in their influential paper on the (non)impact of minimum wage on employment, Card and Krueger (1994) needed to define full-time equivalent employment, a decision that does not directly follow from theory. They define it as the number of full-time workers, including managers, plus 0.5 times the number of part-time workers (p. 775). Should we really include the employment of managers when measuring minimum wage effects? Shouldn't perhaps we weigh part-time employees at 40% or perhaps 60% of full time employment? These questions cannot be answered with theory, so they give rise to equally valid ex-ante specifications, which is precisely why the authors report results defining full-time employment in all of these different ways. Specification-curve analysis concerns itself with these atheoretical decisions and consists of reporting in a descriptively useful, and inferentially interpretable way, the full set of alternative arbitrary specifications that could be generated to validly examine a phenomenon of interest.

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, Shavitt, Viswanathan, & Hilbe, 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.

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 hurricanes 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

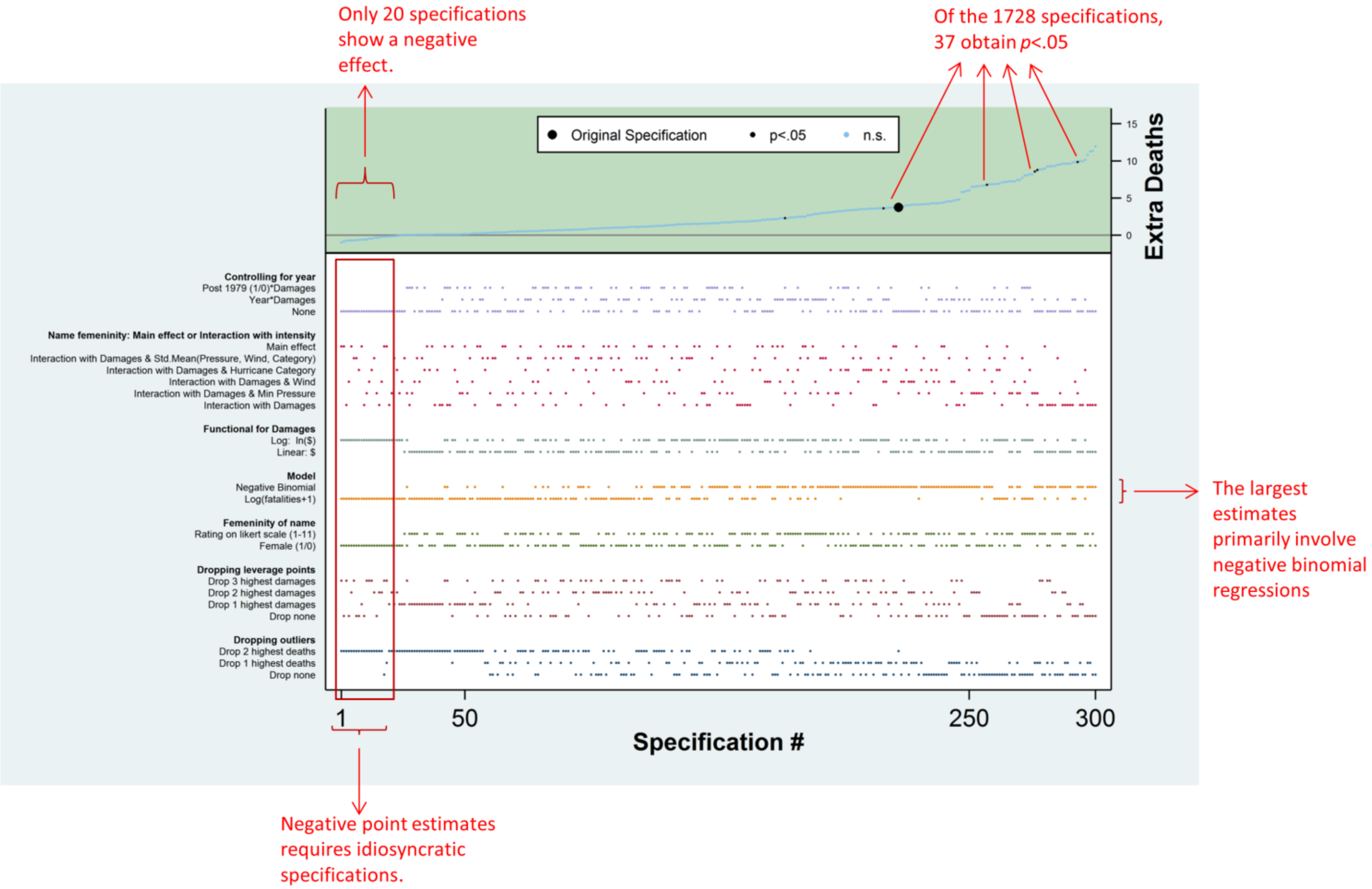
The descriptive specification curve serves two functions: displaying the distribution of estimates that are obtained through alternative reasonable specifications, and identifying which analytic decisions are the most consequential. When the set of reasonable specifications is too large to be estimated 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 examples. 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 the original analyses were statistically invalid because outlier observations, with more than 100 deaths, had been included (Christensen & Christensen, 2014; Maley, 2014), because the regression did not include an interaction between intensity of the hurricane and dollar damages as a predictor (Malter, 2014), and conversely, 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.



Only 20 specifications show a negative effect.

Of the 1728 specifications, 37 obtain *p*<.05

The largest estimates primarily involve negative binomial regressions

Negative point estimates requires idiosyncratic specifications.

Figure 2. *Descriptive Specification Curve.* Each dot in the top panel (green area) depicts a point estimate from a different specification; 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

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?*

It is difficult to answer this question analytically (i.e., with formulas) because the specifications are neither statistically independent nor part of a single model. Fortunately, it is simple to answer this question using permutation techniques for data with random assignment (Ernst, 2004; Fisher, 1935; Pesarin & Salmaso, 2010; Pitman, 1937), and bootstrapping techniques for studies without it (Bickel & Ren, 2001; Davison & Hinkley, 1997). These approaches generate, via resampling, the expected distribution of specification curves when the null hypothesis is true. The two examples in this paper involve experiments, in Supplement 1 we review the traditional regression bootstrapping tools needed to apply it to non-experimental data.[4]

The hurricanes data are a natural experiment. Permutation tests applied to experimental data are extremely simple and intuitive. They involve shuffling the column with the randomly assigned variable (Ernst, 2004; Fisher, 1935; Pesarin & Salmaso, 2010; Pitman, 1937), in this case, the hurricane's name. [5] 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 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.

A valuable property of inference with specification curve is that even if the underlying specifications are parametric or sensitive to the validity of assumptions more generally (as is the case with the negative-binomial regression used for the hurricanes data), hypothesis testing for the curve as a whole, based on the permutation test, is not. For instance, if due to a violated assumption,

[4] In a nutshell, to bootstrap a regression model under the null that x does not affect y, there are two common alternatives: residual resampling and case resampling. Only the latter lends itself to specification curve. It consists of the following: one first estimates the specification on the data, say estimating $y=a+bx+cz+e$. Then one forces the null on the data by creating a new dependent variable that subtracts the estimated effect of x, say $y^*=y-\hat{b}x$. Then one samples with replacement the same number of observations in the data, re-running the specification with y^* rather than y as the dependent variable, and repeats a number of times. The resulting set of $\hat{b}$s consists of the expected distribution of $\hat{b}$ under the null that $b=0$.

[5] 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 permutation test supposed to emulate. Hurricanes names follow alphabetical order and alternate gender. If one were to keep that aspect of the naming procedure fixed across permutations, 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. It is conditioning on all the outcomes that could be rerandomized.

say, some specifications have inflated false-positive ratios, $prob(p \leq .05 \mid H_0) > .05$, the permutation test based on the specification curve will retain a false-positive rate of .05, $prob(p \leq .05|H_0) = .05$.

The only assumption behind permutation tests is exchangeability (Ernst, 2004; Pesarin & Salmaso, 2010), for example, 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. Thus, even with shuffled datasets, we would 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 the exact same regression, all results would be identical, and thus in each shuffled dataset they would either all be positive or all be negative.

Because of this, 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 equally extreme an outcome, 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. We can carry out formal joint significance tests by defining a test-statistic (i.e., a single number) to summarize the entire specification curve, and then comparing the observed value of this statistic with its distribution under the null.

As with any dataset whose dimensionality is reduced to a single summary statistic, there are multiple alternatives, e.g., in two-cell experiments one may compare means, medians, ranks, means of logs, etc. We consider three joint test statistics: (i) the median overall point estimate, (ii) the share of estimates in specification curve that are of the dominant sign, and (iii) the share that are of the dominant sign and also statistically significant (*p*<.05). For example, in the observed hurricanes data, 37 of the 1728 specifications are statistically significant (all with the dominant sign). Among the 500 shuffled samples, 425 have at least 37 significant effects, leading to a *p*-value for this joint test of *p* = 425/500 = .85. See Table 2.

7. 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.

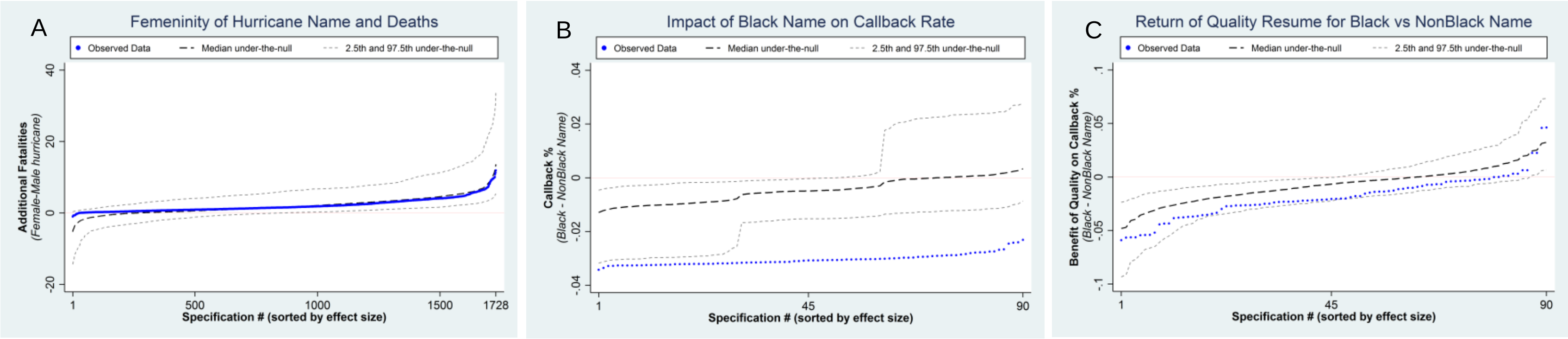


Figure. 3. Observed and expected under-the-null specification curves for the hurricanes (**A**) and racial discrimination studies (**B,C**). The expected curves are based on 500 shuffled samples, where the key predictor in each dataset (hurricane and applicant name respectively) is shuffled. All specifications 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.

	Observed Result	***p*-value** *(% of shuffled samples with as or more extreme results)*
	Example 1. Female hurricanes are deadlier	
(i) Median effect size	1.63 additional deaths	p = .459
(ii) Share of results w/predicted sign	1704 / 1728	p = .156
(iii) Share of results w/predicted sign & p <.05	37 / 1728	p = .850
	Example 2a. Black names receive fewer callbacks	
(i) Median effect size	3.1 *pp* fewer calls	p < .002
(ii) Share of results w/predicted sign	90 / 90	p = .125
(iii) Share of results w/predicted sign & p <.05	85 / 90	p < .002
	Example 2b. Black names benefit less from quality CV	
(i) Median effect size	2.0 *pp* smaller benefit	p = .030
(ii) Share of results w/predicted sign	79 / 90	p = .13
(iii) Share of results w/predicted sign & p <.05	13 / 90	p = .162

Table 2. Joint tests for inferential specification curves in the two examples. *pp*: percentage-points. 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)

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.

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 disagreement 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 operationalization 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.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives, 28*(2), 29-50.

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.

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.

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.

Davison, A. C., & Hinkley, D. V. (1997) *Bootstrap methods and their application*: Cambridge university press.

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.

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.

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.

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.

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.