2015

Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications

Uri Simonsohn
University of Pennsylvania

Joseph P. Simmons
University of Pennsylvania

Leif D. Nelson

Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications

Abstract
Empirical results often hinge on data analytic decisions that are simultaneously defensible, arbitrary, and motivated. To mitigate this problem we introduce Specification-Curve Analysis, which consists of three steps: (i) identifying the set of theoretically justified, statistically valid, and non-redundant analytic specifications, (ii) displaying alternative results graphically, allowing the identification of decisions producing different results, and (iii) conducting statistical tests to determine whether as a whole results are inconsistent with the null hypothesis. We illustrate its use by applying it to three published findings. One proves robust, one weak, one not robust at all.

Keywords
specification curve, p-hacking

Disciplines
Applied Statistics | Business | Marketing

Comments
This is an unpublished version.

This version: November 2015

Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications

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

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

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

Abstract: Empirical results often hinge on data analytic decisions that are simultaneously defensible, arbitrary, and motivated. To mitigate this problem we introduce Specification-Curve Analysis, which consists of three steps: (i) identifying the set of theoretically justified, statistically valid, and non-redundant analytic specifications, (ii) displaying alternative results graphically, allowing the identification of decisions producing different results, and (iii) conducting statistical tests to determine whether as a whole results are inconsistent with the null hypothesis. We illustrate its use by applying it to three published findings. One proves robust, one weak, one not robust at all.

Note: Supplementary materials begin on page 21

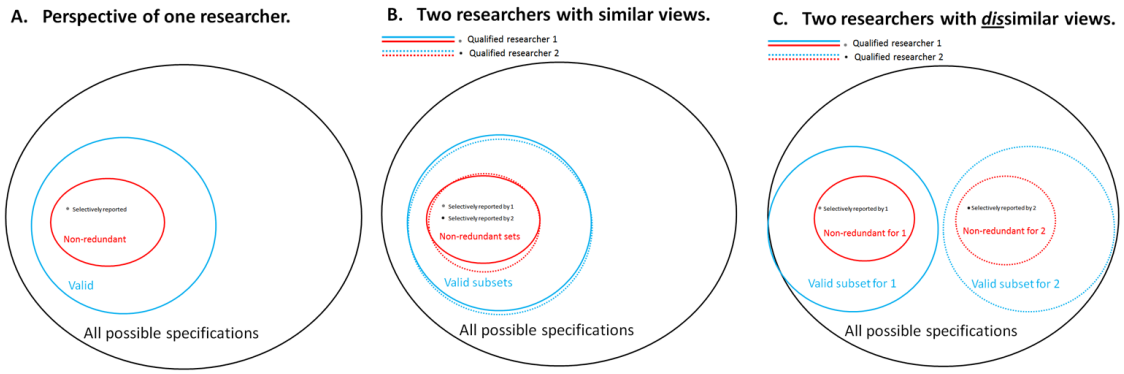
The empirical testing of scientific hypotheses requires data analysis, but data analysis is not straightforward. Instead, to convert a scientific hypothesis of interest into a testable prediction, researchers must make a number of data analytic decisions, many of which are both arbitrary and defensible.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

The abundance of valid specifications limits the conclusiveness of the results supported by any small subset of specifications, as those results may hinge on an arbitrary choice by the researcher (Leamer 1983). This problem is exacerbated by the fact that specifications are usually chosen by researchers who have a conflict of interest, reporting a result that tells a publishable story (Leamer 1983, Ioannidis 2005, Simmons, Nelson, and Simonsohn 2011, Glaeser 2006).

In this article we introduce Specification-Curve Analysis as a way to mitigate the problem. The approach consists of reporting results for all "reasonable specifications," by which we mean specifications that (1) are consistent with the underlying theory, (2) are expected to be statistically valid, and (3) are not redundant with other specifications in the set.

Figure 1 helps understand what reporting results for all reasonable specifications does, and does not, entail. 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 non-redundant alternatives (e.g., either log(x+1) *or* log(x+1.1)).

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



Currently, without specification-curve analysis, researchers selective report a few specifications in their papers (between one and a few handfuls), depicted by the small gray dot inside the red circle. Specification-curve analysis expands what gets reported from the gray dot to the entire red circle. Importantly, it does not expand beyond it. Researchers do not need to estimate specifications they consider redundant, and certainly not specifications they consider invalid. Specification-curve analysis seeks to reduce the impact of arbitrary analytical decisions while preserving the impact of non-arbitrary analytical decisions.

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).

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 vs not).

Despite having non-overlapping sets of reasonable specifications, specification-curve analysis can aid researchers 1 and 2 understand potentially different conclusions, by disentangling whether they are rooted in ex-ante disagreements of which specifications are valid, or instead in the arbitrary selectively reported results from those sets. In other words, specification curve disentangles whether the different conclusions originate in differences regarding sets of analyses deemed reasonable (different red circles), or merely in which particular few analyses the researchers reported (different gray dots).

I. 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, assessing if the best fitting model among a class of models fits better than expected by chance having been selected post-hocly as the best (White 2000). A third proposed approach consists of reporting the standard deviation of point estimates across alternatives specifications (Athey and Imbens 2015). A fourth approach is the most similar to ours. It is known as "extreme bounds analysis," where one estimates regression models for every possible combination of covariates. 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).

Among other differences with all four 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), (iii) and provides a formal joint significance test for the family of alternative specifications, derived from expected distributions under the null. No existing approach that we are aware of provides any of these three features.

In relation to the most well known approaches within economics in particular (Leamer 1983, Sala-i-Martin 1997), Specification-Curve Analysis considers all

operationalization decisions, not just those of covariates. Disagreements about covariates tend to involve the more conceptual discussion of what is vs is not appropriate to control for in light of the theory of interest, rather than how to operationalize a given theory. The *interpretation* of an effect with and without a covariate is often substantially different, while that from an estimate of using one vs another algorithm to define outliers or generate weights behind the dependent variable often less so. Specification-Curve Analysis seeks to reduce the impact of arbitrary operationalizations, not of non-arbitrary theorizing.

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 learning what the conclusion hinges on.

II. 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 (Jung et al. 2014a, Malter 2014, Maley 2014, Bakkensen and Larson 2014, Christensen and Christensen 2014, Jung et al. 2014b), 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 and 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.

A. 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. Note that if the resulting set is too large, 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, which covariates to include in the analysis, which regression model to use, and which functional form to assume. 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.Which covariates to include	Property damages in dollars interacted with femininity; minimum hurricane pressure interacted with femininity	Log of dollar damages; year; year interacted with damages
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

B. Step 2. Estimate & Describe Results

The descriptive specification curve serves two functions: displaying the range of estimates that are obtained through alternative reasonable specifications, and identifying analytic decisions that are 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. [1]

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 controlling for year, (ii) controlling for the *log* of dollar 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.

[1] To make point estimates for the continuous and discrete measures of femeninity 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.

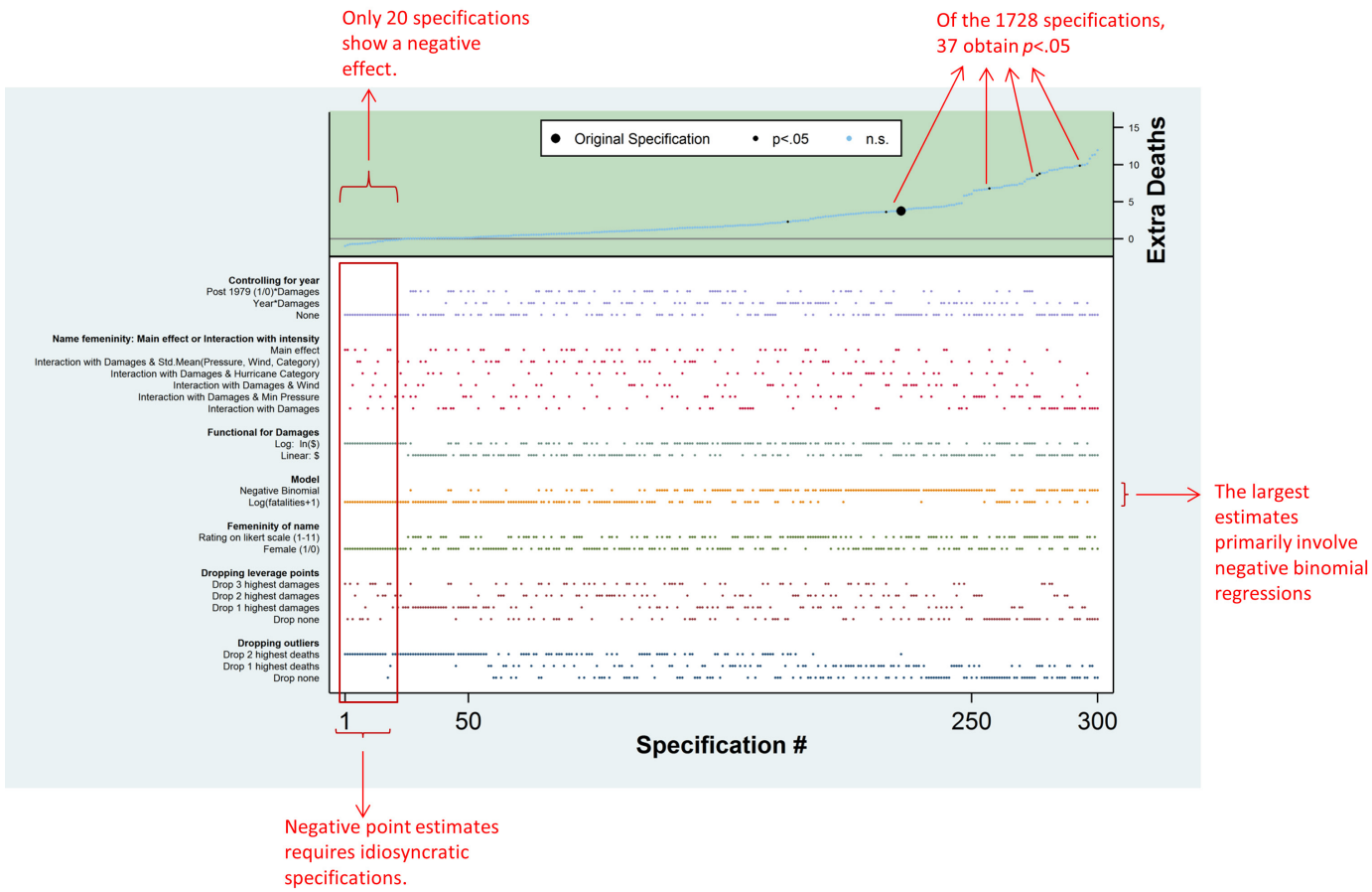


Figure 2. *Descriptive Specification Curve.* Each dot in the top panel (green area) depicts a point estimate from a different specification; the dots vertically aligned below (white area) indicate the analytic decisions behind those estimates. A total of 1728 specifications were estimated; the figure depicts the 50 highest and lowest point estimates, and a random subset of 200 additional ones.

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 (Christensen and Christensen 2014, Maley 2014, Malter 2014, Bakkensen and Larson 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 and 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 and Larson 2014).

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. Better to publish a single specification curve in the original paper.

Step 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 (Pitman 1937, Fisher 1935, Pesarin and Salmaso 2010, Ernst 2004), and bootstrapping techniques for studies without it (Davison and Hinkley 1997, Bickel and Ren 2001). These approaches generate, via resampling, the expected distribution of specification curves when the null hypothesis is true.

The hurricanes data are a natural experiment. Permutation tests applied to experimental data are extremely simple and intuitive. They consist of shuffling the column with the randomly assigned variable (Pitman 1937, Fisher 1935, Pesarin and Salmaso 2010, Ernst 2004), in this case, the hurricane's name. 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, 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, $prob(p \leq .05 \mid H_0) = .05$.

The only assumption behind permutation tests is exchangeability (Pesarin and Salmaso 2010, Ernst 2004), for example, that any hurricane could have received any name. 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 slightly different covariates), the results obtained from different specifications applied to the same dataset 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 either all positive or all 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 by, 80%, say, of specifications having the same sign, whether positing or negative, is treated as an equally extreme outcome.

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.

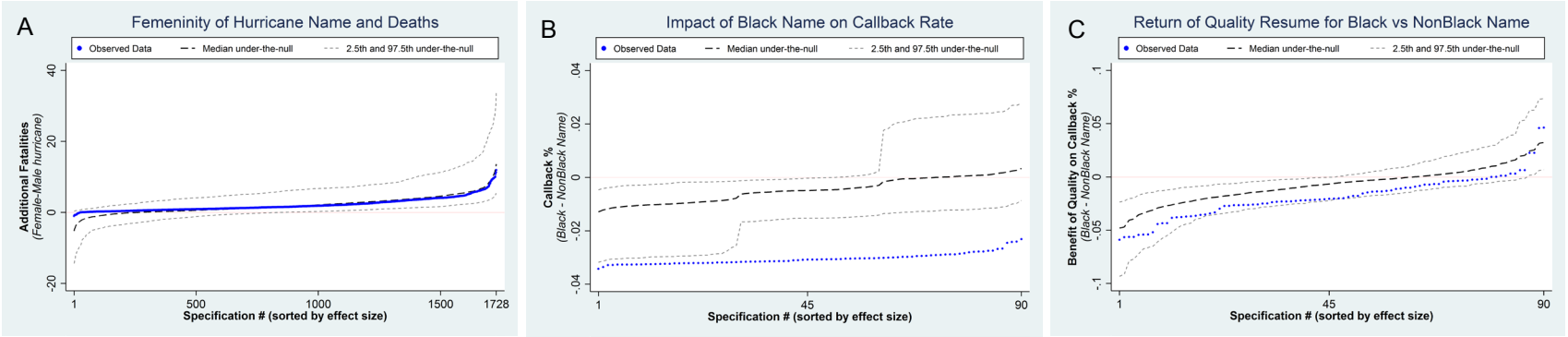


Figure. 3. Observed and expected under-the-null specification curves for the hurricanes (**A**) and racial discrimination studies (**B,C**). The expected curves are based on 500 shuffled samples, where the key predictor in each dataset (hurricane and applicant name respectively) is shuffled. All specifications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

III. Second example

Having gone through the three steps for carrying out Specification-Curve Analysis with our first example, we move on to our second example (Bertrand and 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.

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.

IV. 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 desirable in our view).

When different researchers arrive at different conclusions from the same data, the disagreement may reflect profound 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 theory (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 (this is actually not a big problem, our datasets are samples too), 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 "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 1728 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 is less imperfect than the alternatives we are aware of.

References

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

Bakkensen, LA, and W Larson. 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, and S Mullainathan. 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, Peter J, and Jian-Jian Ren. 2001. "The bootstrap in hypothesis testing." *Lecture Notes-Monograph Series, State of the Art in Probability and Statistics* 36:91-112.

Christensen, Björn, and Sören Christensen. 2014. "Are female hurricanes really deadlier than male hurricanes?" *Proceedings of the National Academy of Sciences* 111 (34):E3497-E3498.

Davison, Anthony Christopher, and D.V. Hinkley. 1997. *Bootstrap methods and their application*. Cambridge university press.

Efron, Bradley. 2014. "Estimation and accuracy after model selection." *Journal of the American Statistical Association* 109 (507):991-1007.

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

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

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

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

Jung, Kiju, Sharon Shavitt, Madhu Viswanathan, and Joseph M Hilbe. 2014a. "Female hurricanes are deadlier than male hurricanes." *Proceedings of the National Academy of Sciences*:201402786.

Jung, Kiju, Sharon Shavitt, Madhu Viswanathan, and Joseph M Hilbe. 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, HO. 1961. "Significance Tests in Discrete Distributions." *Journal of the American Statistical Association* 56 (294):223-234.

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

Maley, Steve. 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, Daniel. 2014. "Female hurricanes are not deadlier than male hurricanes." *Proceedings of the National Academy of Sciences* 111 (34):E3496-E3496.

Miguel, E, Colin F. Camerer, K Casey, J Cohen, KM Esterling, A Gerber, R Glennerster, DP Green, M Humphreys, and G Imbens. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166):30-31.

Pesarin, Fortunato, and Luigi Salmaso. 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, Xavier X. 1997. "I just ran two million regressions." *The American Economic Review*:178-183.

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

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

This version: 2015 11 12

SUPPLEMENTARY MATERIALS FOR:

"Specification Curve: Descriptive and Inferential Statistics On All Reasonable Specifications"

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

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

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

OUTLINE.

Supplement 1. Set of reasonable specifications for hurricanes study.

Jung et al. (Jung et al., 2014) hypothesized that hurricanes with more feminine names are perceived as less threatening and hence lead to fewer precautionary measures by the general public. To convert this hypothesis into a testable prediction, Jung et al carried out various operationalizations. To construct a specification curve, we examine what we judged to be the five major operationalizations (most likely to be consequential), and consider sensible alternatives. In particular, we shall examine these operationalizations:

1. The set of storms to include in the analyses
2. How to measure the femininity of storms' names
3. What regression model to run (e.g., Negative-Binomial vs OLS)
4. What's the key prediction made by the authors' hypothesis
5. What to control for

Note that these five mirror the operationalizations in Figure 1 in the main paper.

1) The set of storms to include in the analyses

1.1) Universe of storms

Jung et al. included only Atlantic hurricanes included in a NOAA list (see page in their paper).[1] The universe of named storms that cause destruction is much larger than that. First, named tropical storms (of lesser intensity than hurricanes) also lead to deaths and have gendered names (these would more than double sample size). Second,

[1] The list has been saved on the WebArchive:
http://web.archive.org/web/20140709120550/http://www.aoml.noaa.gov/hrd/hurdat/All_U.S._Hurricanes.html

hurricanes elsewhere in the world also receive names and also cause deaths.[2] Third, some Atlantic hurricanes are missing from the NOAA list used. For instance hurricane Diane from 1955, Isidore from 2002, and Ernesto from 2006 are missing from the list.

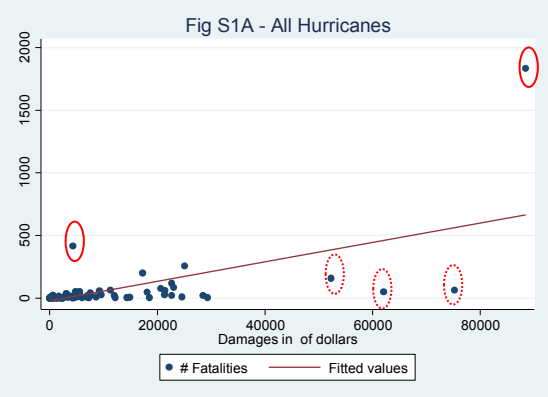
While it would be sensible and straightforward to assess the robustness of the results to different definitions of the universe of storms to study, this type of robustness is somewhat unusual (most papers cannot so easily expand their data-sources) and hence we have decided not to do so for our specification curve demonstration.

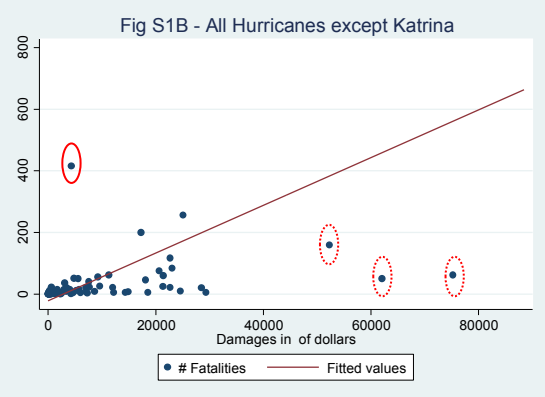
1.2) *Outliers*

The paper excludes hurricanes Katrina and Audrey from the analyses, considering them "outliers."[3] The exclusion decision was made after the authors run the regressions with them included, and is motivated in the paper as seeking to eliminate over-dispersion from the model (rather than seeking to eliminate observations that may be invalid).[4,5]

The two excluded observations have 1833 and 416 deaths, see solid circles in Figure S1. The same figure highlights other candidate outlier observations (dotted circles). The can also be though thought of as leverage points, observations with extreme predictor values.

[2] See e.g. this list saved on the WebArchive:
http://web.archive.org/web/20150216102357/http://www.nhc.noaa.gov/aboutnames.shtml
[3] Jung et al. reads *"We removed two hurricanes, Katrina in 2005 (1833 deaths) and Audrey in 1957 (416 deaths), leaving 92 hurricanes for the final data set. Retaining the outliers leads to a poor model fit due to overdispersion*." (p.4)
[4] There are many other ways to address over-dispersion. For example, a textbook on binomial regressions, written by one of the authors of the PNAS paper, suggests 35 different methods to deal with overdispersion; adjusting for outliers is just one of them, and there are in turn many ways to deal with outliers (Hilbe (2011) "*Negative Binomial Regressions*," Second Edition, p. 158)
[5] Because those observations are known to be legitimate, the overdispersion these "outliers" create probably is better addressed by modifying the model rather than ignoring those valid datapoints.





Notes: The solid circles are the two "outliers" drop from the analyses by Jung et al., the dashed circles identify additional potential outliers (or leverage points, extreme values of X variable).

In light of the above, we examine the following alternatives to deal with outliers (new operationalizations in light blue, original in black):

Alternative operationalizations for dealing with "outliers:"

1) Exclude 0 observations.
2) Exclude 1 most extreme on deaths (drop Katrina with 1833, keep Audrey)
3) Exclude 2 most extreme on deaths (drop Katrina & Audrey)
4) Exclude 2 most extreme on deaths and remaining 1 most extreme on damages
5) Exclude 2 most extreme on deaths and remaining 2 most extreme on damages
6) Exclude 2 most extreme on deaths and remaining 3 most extreme on damages

2) How to measure the femininity of storms' names

The authors used 9 raters to judge in a 1-11 scale the femininity of the storm names, and also a binary (1=female, 0=male) gender indicator. [6,7]

Operationalizations for quantifying femininity: [8]

[6] Throughout we use the term "linearity" abstracting from the fact that estimated regression is a negative binomial and hence a linear term is not really assuming a linear effect. The key point is that a linear term imposes a strong functional form assumption, rather than that assumption is of a linear effect per-se.
[7] An additional concern worth mentioning is that femininity of a name may be correlated with other attributes of the name, such as how strong, evil, or harmful names are perceived. E.g., male name Adolf vs female name Angel. This would require controlling for other name attributes, something that would be a distraction for our purposes but necessary to properly interpret the original results.
[8] Because the authors do not report femininity for Katrina and Audrey, we conducted an MTUrk survey with 32 participants asking them to rate all 94 storms using the same scale from Jung et al. The ratings were correlated $r = .98$ with those used by Jung et al. However, within gender the ratings are much lower: $r = .67$

1) Femininity rating
2) Binary gender indicator

3) What regression model to run

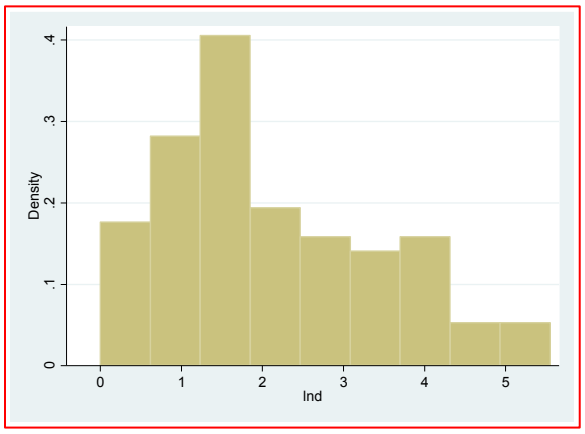
Jung et al. estimated a negative binomial regression which is often used for count data with over-dispersion (that is, where the Poisson assumption of $\mu=\sigma$ does not apply). Some papers examining deaths from natural disasters employ a zero-inflated negative binomial (Czajkowski, Simmons, & Sutter, 2011), where a higher proportion of observation with 0 values (higher than that implied by a regular negative binomial) is observed. To perform a zero inflated one needs to identify variables that predict whether there are any deaths but not how many there are. This complication, paired with just 10% of observations having 0 deaths, leads us to not include a zero-inflated negative binominal in the specification curve.

Another alternative to the negative binomial is to run OLS with log(count+1) as the dependent variable. Logging count data has been discouraged (O'Hara & Kotze, 2010). The discouragement is based on simulations that show that if the true data are binomial or Poisson, then log(count) performs worse than a Poisson or negative binomial model, but the whole point is that one may doubt the underlying data are adequately captured by a negative binomial or Poisson model, and wishes to run the log model for robustness. As the authors of the paper discouraging log(counts) write "…*our result may not generalize to real data, which rarely has (sic) as balanced a design as our simulations"*

---

for male names, $r = .83$ for female names. This suggests the measure of femininity beyond the binary gender variable adds considerable noise to the model. We use the MTurk ratings in our analyses.

Simple linear models are known to be robust to a broad range of violations of assumptions, this is not the case for non-linear models like the negative binomial. In addition, log(deaths) has a rather reasonable distribution, that a linear model should obtain valid estimates form.



Moreover, a well cited paper of predicted hurricane deaths uses a OLS with log(deaths+1) as the dependent variable (Toya & Skidmore, 2007). [9] We therefore include OLS regressions with log(count+1) as the dependent variable in our specification curve.

Regression models
1) Negative binomial
2) OLS regression with log(deaths+1) as the dependent variable

4) What's the key prediction made by the authors' hypothesis

The key hypothesis in the paper is that "*a hurricane with a feminine vs. masculine name will lead to less protective action and more fatalities*." (p.1) This prediction

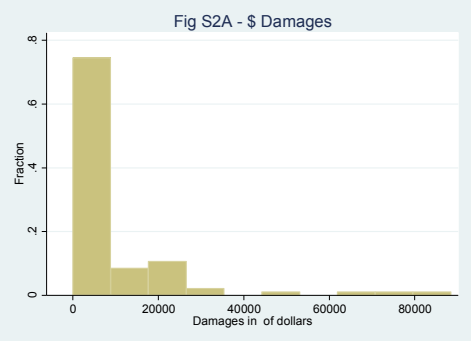
[9] 279 Google cites as of January 2015.

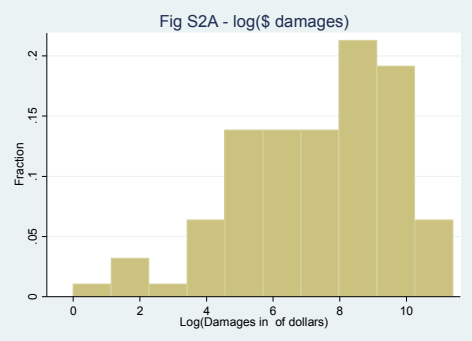
suggests a **main effect** of gender of hurricane name on fatalities but the paper tests instead an interaction of femininity with dollar damages (such that the effect is stronger for hurricanes high in dollar damages).

One can justify this specification by considering that fatalities is a proxy for insufficient protective measures, and the proxy is only sensible when people who fail to take protective measures have a higher probability of death. Failing to protect against harmless storms should not lead to higher probability of death. Nevertheless, there are many alterantive ways to operationalize this potential dependency of the effect of gender on dollar damages (including ignoring it). We discuss these alternatives below.

4.1 Damages vs log(damages)

The operationalization in the paper has gender interacted with damages measured in dollars "linearly" [10] mapping on to deaths. As the histograms in Figure S2 show, damages measured in dollars is an extremely skewed distribution, while log(damages) has a reasonable degree of skew. We hence shall include specifications with log(damages) also.





[10] See footnote 4

Functional form for $ damages:

1) $ Damages
2) Log($ Damages)

4.2 Interactions

Beyond functional form, note that damages is being used as a proxy for potential fatalities. There are various other ways to proxy how deadly a storm would be expected to be. Rather than use only dollar damages, for instance, one may include hurricane characteristics as well, such as its category, it maximum wind speed, its year, etc. Indeed Jung et al used an interaction with minimum pressure as a covariate. We consider operationalizations that add or replace wind, and hurricane category as proxies for expected fatalities. [11]

4.3 Main effect

Finally, while the justification for the interaction prediction is reasonable, it seems ex-ante also reasonable to examine the *main effect* of gender, on various grounds. First, as shown above, the stated hypothesis in the paper stipulated a main effect of femeninity. Second, 90% of hurricanes lead to at least one fatality, suggesting lack of protective measures could be fatal for most observations. Third, when the authors control for third variables (e.g., hurricane year), for which the same logic to require an interaction applies, they do not include the interaction, suggesting further than main effect estimates are sensible. Fourth, prior papers running models predicting fatalities from natural disasters like hurricanes and earthquakes test main effects (of variables like per-capita GDP) rather

[11] Jung et al indicate that wind data are not available prior to 1979, but we were able to locate such data and use it in our analyses.

than interactions with disaster intensity (Czajkowski, Simmons, & Sutter, 2011; Kahn, 2005; Toya & Skidmore, 2007). It hence seems ex-ante sensible to estimate a gender main effect to test the hypothesis of interest.

Operationalizations for key prediction.

Codebook

Zmin:[12] -1*(standardize minimum pressure)
Fem: femininity of name of hurricane
Dam: damages of hurricane (log or $)
zCat: standardize category of hurricane (1-5)
zWin: standardized maximum wind of hurricane
z3: average(zMin, zCat, ZWin)

Specifications to test the impact of femininity.
Via interactions:

1. Fem*Dam
2. Fem*Dam Fem*zMin
3. Fem*Dam Fem*zWind
4. Fem*Dam Fem*zCat
5. Fem*Dam Fem*z3

Via main effect:

6. Fem Dam z3

5) What to control for

Hurricane names are randomly assigned, hence covariates might be expected to not play an important role in the regression (omitted variables shouldn't correlate with randomly assigned names). The key predictor in most specifications, however, involves an interaction, and the interaction term, Damages, is not randomly assigned and hence could have confounds. Moreover, as shown above it is highly skewed, introducing possible specification error into some models, additional controls interacted with damages may alleviate or at least facilitate identifying these problems.

[12] Minimum pressure is multiplied by -1 so that a higher number is associated with higher intensity and is hence easy to combine with the other indicators.

The dataset only contains year as a plausible covariate (in addition to the hurricane intensity variables from (3)). Considering that time effects can often be non-linear, and that year has an important discontinuity in 1979, prior to 1979 all hurricanes had male names we consider the following set of covariate:

Specifications for covariates

1) No covariate[13]
2) Year * Damages
3) Dummy for year after 1979 * Damages

[13] Jung, et al. (2014) did also run models controlling for year and write that it "was dropped for the main analysis as its effect was nonsignificant in all models." (p.4)

Supplement 2. Set of reasonable specifications for racial discrimination study.

Bertrand and Mullainathan (Bertrand & Mullainathan, 2004) manipulated both the name of the fictitious candidates whose names they used and the quality of the resumes they sent. To manipulate perceived race, they used 18 distinctively African American names, and 18 non-distinctively African-American names. Half the names were male. The way in which resumes' quality was manipulated varied from ad to ad, creating some ambiguity in terms of how to quantify the quality of any given resume. This ambiguity is the primary source of the alternative specifications we consider.

We generate the set or reasonable specifications by considering alternative operationalizations involving:

1) How to deal with potential heterogeneity of the main effect across genders
2) How to measure quality of resume
3) What regression model to employ (OLS vs Probit)

1) How to deal with potential heterogeneity of the main effect across genders

Considering Bertrand and Mullainathan report some results broken down by gender, that it would be reasonable to observe and report discrimination only in one of the genders, or of different magnitude across gender, we report results for the entire sample, only for males, and only for females.

2) How to measure quality of resume

To most help-wanted ads, Bertrand and Mullainathan (Mullainathan, 2002) sent four resumes. Two high and two low quality ones (orthogonally varying race and gender also). Whether to be of high and low quality is randomly assigned, but how to implement

how vs low quality is decided on a case-by-case basis in light of the nature of the ad. Resumes were made of higher/lower quality by varying holes in employment history, having a certification degree, possessing foreign language skills, etc.

Bertrand and Mullainathan operationalize quality in their regression results in two main ways: using a 1/0 predictor for having been randomly assigned to high vs low quality, using a continuous predictor of quality (see e.g. Panels A and B in their Table 4). The continuous predictor was created by estimating the regressions in two stages. In the first stage, using 1/3 of the sample, they predict call-back rates using all measures of quality they manipulated. They then use the fitted values for call-back probabilities as the quality index for the remaining 2/3 of the sample, using a median split for high vs low predicted call-back rates as the alternative measure of quality.

We expand these two to fifteen alternative operationalizations of quality. We modify their two-stage estimation in a way that increases power. In particular, rather than use 1/3 of the sample to obtain fitted values, we use 1/2. In addition, we do not drop observations, our second stage includes all observations, one half of fitted values are obtained from the other half. One could increase power further with more refined techniques (e.g, jackknife) but it is not necessary for the purposes of our demonstration.

As operationalizations of quality we added the simple sum of 0/1 quality indicators (that is, the number of quality variables that were changed to create a higher quality resume). A second alternative was the median split of this variable.

The remaining alternatives are rely on the two-stage estimation approach alluded to above. We varied whether this first stage included covariates or not (the specification in the paper includes as covariates gender, city, occupation code for the job, and dummy

variables for required skills), and whether it was estimated on all names, or only distinctively Black or White names. The logic for this later variation in operationalizations is that Blacks and non-Blacks may benefit differently from different quality measures (e.g., some quality measures may alleviate negative stereotypes for Black names, but have no effect on White names).

Each of these three 2-stage approaches was implemented with and without covariates in the 1st stage , and entered as a continuous or median split predictor in the 2nd stage, resulting in 3x2x2=12 operationalizations. Combined with the previously mentioned 3 we arrive at the 15 alternative specifications of quality.

3) Regression model

The paper (Bertrand & Mullainathan, 2004) reported probit regression for the 1st stage, classified observations into a high and low quality bin, and conducted simple $\chi^2(1)$ difference of proportion tests on the resulting cells (see their Table 4). Because we consider continuous predictors of quality we rely on regression models throughout, reporting results for both probit and OLS regressions. We should note that they key prediction is one of an interaction: is the benefit of a quality resume higher for nonblack names? (Gelman & Stern, 2006) But the authors only report the two simple effects (significant effect for nonBlacks, but not for Blacks). The non-reported interaction is not significant in either of the specifications included in the paper.

Supplement 3. Descriptive Specification Curves for Discrimination Study

Figure S3. Descriptive Specification Curve – Main effect of distinctively black name on call back rate.

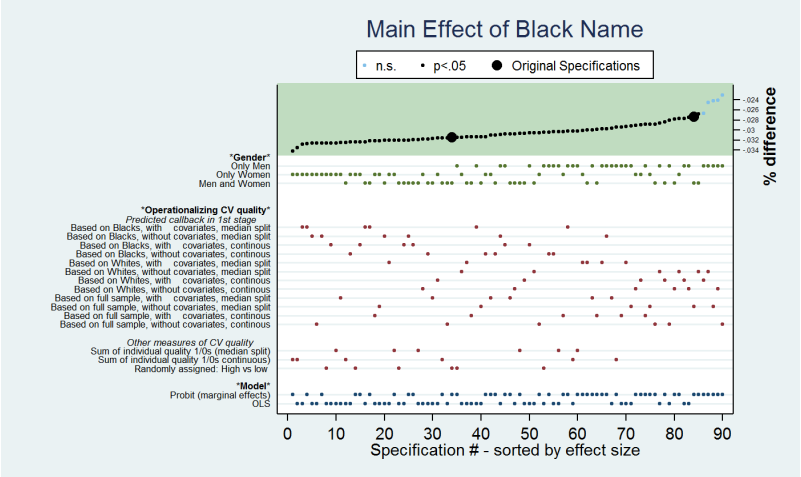
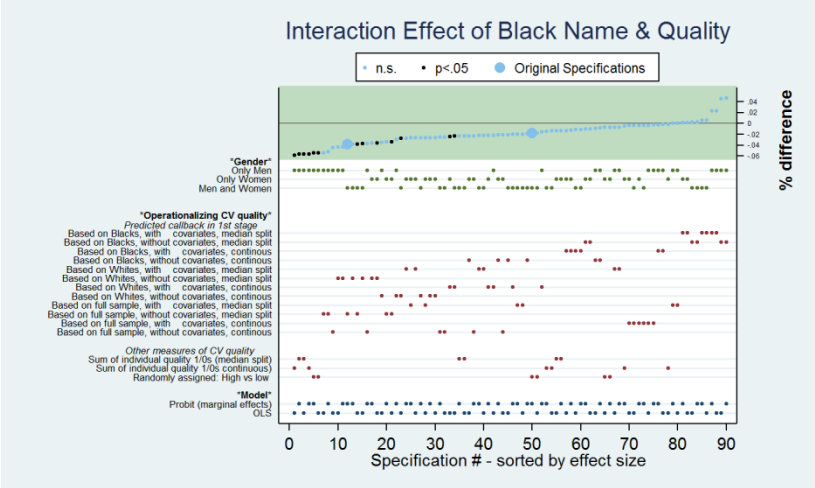


Figure S4. Descriptive Specification Curve – Interaction of distinctively black name and resume quality, on call back rate.



References

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.

Czajkowski, J., Simmons, K., & Sutter, D. (2011). "An Analysis of Coastal and Inland Fatalities in Landfalling Us Hurricanes." *Natural hazards, 59*(3), 1513-1531.

Gelman, A., & Stern, H. (2006). "The Difference between "Significant" and "Not Significant" Is Not Itself Statistically Significant." *The American Statistician, 60*(4), 328-331.

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). "Female Hurricanes Are Deadlier Than Male Hurricanes." *Proceedings of the National Academy of Sciences*, 201402786.

Kahn, M. E. (2005). "The Death Toll from Natural Disasters: The Role of Income, Geography, and Institutions." *Review of Economics and Statistics, 87*(2), 271-284.

Mullainathan, S. (2002). "A Memory-Based Model of Bounded Rationality." *Quarterly Journal of Economics, 117*(3), 735-774.

O'Hara, R. B., & Kotze, D. J. (2010). "Do Not Log-Transform Count Data." *Methods in Ecology and Evolution, 1*(2), 118-122.

Toya, H., & Skidmore, M. (2007). "Economic Development and the Impacts of Natural Disasters." *Economics Letters, 94*(1), 20-25.