Inferential Selection Bias in a Study of Racial Bias: Reproduction of "Working Twice as Hard to Get Half as Far"

L.J Zigerell
Assistant Professor of Politics and Government
Illinois State University
Schroeder Hall 401
Normal IL 61790-4540
ljzigerell@ilstu.edu
724-561-8280

Abstract. A recent article reported evidence from a survey experiment indicating that Americans reward whites more than blacks for hard work but punish blacks more than whites for laziness. However, the present study demonstrates that these inferences were based on an unrepresentative selection of possible analyses. Combination of analyses reported in the original article with equivalent analyses reported in the present study provided evidence that black targets were punished more than white targets for laziness but were not rewarded less than white targets for hard work, at least at a statistically significant level; moreover, newly-reported evidence indicates that respondents were biased in favor of a black target over a white target when given a direct choice between equivalent targets of different races. Results suggest benefits to readers and researchers from preregistration of research design protocols.

"Researcher degrees of freedom" refers to a situation in which a hypothesis can be tested in multiple ways. This problem is often discussed in terms of p-hacking: "...it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields 'statistical significance,' and to then report only what 'worked'" (Simmons et al. 2011: 1359). But the key problem caused by researcher degrees of freedom – sensitivity of inferences to research design specifications – can still be present when researchers refrain from manipulating a research design to obtain a preferred result (Gelman and Loken 2013: 2). Consider a situation in which a hypothesis can be tested in two ways: test 1 would produce a p-value of 0.01, and test 2 would produce a p-value of 0.85. If only test 1 results are reported, the reader's confidence in the correctness of the hypothesis will be overstated compared to what the reader's confidence would have been had results from both tests been reported; however, in terms of inferential error made by a reader, it would not matter whether the researcher suppressed test 2 results or did not realize that test 2 was possible. This test 1 / test 2 scenario illustrates inferential selection bias in a hypothetical; to illustrate inferential selection bias in a real-world example, this study demonstrates how key inferences in DeSante (2013) conflict with inferences from a more representative selection of analyses.

Review of the Experiment Reported in DeSante (2013)

DeSante (2013) reported results from a sample of 1,000 adult respondents to the 2010 Cooperative Congressional Election Studies, providing evidence of a nuanced antiblack bias among members of the U.S. adult population: blacks are penalized more than whites for laziness but are rewarded less than whites for hard work. These inferences were drawn from

an experiment in which respondents were asked to divide $1500 among three alternatives: to an applicant for state assistance said to need $900, to another applicant for state assistance said to need $900, and to offset the state budget deficit. Respondents were shown an application for state assistance, with variation in two elements: first, the applicant had no name provided, a name intended to indicate a white female (Laurie or Emily), or a name intended to indicate a black female (Keisha or Latoya); second, an application section for Worker Quality Assessment was not provided, indicated a poor assessment, or indicated an excellent assessment. Table 1 reports characteristics of the experimental conditions and mean allocations in each condition. See DeSante (2013) for more detail on the research design.

Table 1. Experimental Condition Descriptions and Mean Allocations

Condition	Applicant 1			Applicant 2			State Budget Deficit	N
	Name	Worker Quality Assessment	Mean Allocation	Name	Worker Quality Assessment	Mean Allocation		
1	--	--	579	--	--	595	326	117
2	--	Excellent	644	--	Poor	416	439	67
3	--	Poor	512	--	Excellent	618	370	63
4	Laurie	--	579	Emily	--	587	334	112
5	Laurie	Excellent	682	Emily	Poor	566	250	64
6	Laurie	Poor	478	Emily	Excellent	711	311	55
7	Laurie	--	556	Keisha	--	600	345	133
8	Laurie	Excellent	620	Keisha	Poor	486	394	55
9	Laurie	Poor	500	Keisha	Excellent	607	394	70
10	Latoya	--	546	Keisha	--	567	387	133
11	Latoya	Excellent	627	Keisha	Poor	460	413	72
12	Latoya	Poor	434	Keisha	Excellent	597	469	59

Reported and Unreported Experimental Condition Comparisons in DeSante (2013)

For reporting of analyses below, the notation [X/Y] indicates an allocation to applicant X in condition Y; the reporting presumes that respondents perceived Emily and Laurie as white applicants and Keisha and Latoya as black applicants.[1] Data were not weighted in any analysis, t-tests were conducted with equal variances assumed, and – unless otherwise indicated – reported p-values are two-tailed p-values.

Table 2 of DeSante (2013: 350) reported results from eleven t-tests to compare allocations in selected experimental conditions. Test 1 compared [1/1] to [1/2], for which the allocation difference was $65 (p=0.09); however, the same test – an unnamed worker with no Worker Quality Assessment compared to an unnamed worker with an excellent Worker Quality Assessment – could have been conducted by comparing [2/1] to [2/3], for which the allocation difference was $23 (p=0.56).[2] Similarly, DeSante (2013) test 2 compared

[1] There was no difference at a conventional level of statistical significance between mean allocations to unnamed applicants in condition 1 (p=0.21) or to white applicants in condition 4 (p=0.34), but Keisha received a mean allocation $21 more than Layota did in condition 10 (p=0.03).

[2] Condition 1 compared applicant 1 (with no reported Worker Quality Assessment) to an unnamed worker with no reported Worker Quality Assessment, but condition 2 compared applicant 1 (with an excellent Worker Quality Assessment) to an unnamed worker with a poor Worker Quality Assessment, so test 1 cannot isolate the effect of an excellent Worker Quality Assessment for one applicant from the effect of a poor Worker Quality Assessment for the other applicant.

[2/1] to [2/2], for which the allocation difference was $179 (p<0.0001); but the same test – an unnamed worker with no Worker Quality Assessment compared to an unnamed worker with a poor Worker Quality Assessment – could have been conducted by comparing [1/1] to [1/3], for which the allocation difference was $67 (p=0.08). In both cases, DeSante (2013) reported t-tests that produced allocation estimates 2 to 3 times larger than equivalent un-reported t-tests described here.

DeSante (2013: 349) reported the results of Tests 3 and 4 as follows: "[n]either test shows any significant difference, meaning that white applicants are not rewarded any more than blacks on the basis of race alone." Test 3 compared [1/7] to [1/10] (Laurie, paired with Keisha, was compared to Latoya, paired with Keisha, none of whom had a Worker Quality Assessment), and test 4 compared [2/4] to [2/7] (Emily, paired with Laurie, was compared to Keisha, paired with Laurie, none of whom had a Worker Quality Assessment); neither test had a low p-value (p=0.75 and p=0.69, respectively). However, respondents in condition 7 were presented with a direct allocation decision between Laurie and Keisha, neither of whom had a Worker Quality Assessment: the mean allocation to Keisha was $44 higher than the mean allocation to Laurie (p=0.0002), representing bias in favor of the black applicant relative to the white applicant.

DeSante (2013) tests 5, 6, and 7 form a group: test 5 compared [2/4] to [2/6] to assess how much an excellent Worker Quality Assessment increased Emily's allocation relative to Laurie ($123, p=0.001); test 6 compared [2/7] to [2/9] to assess how much an excellent Worker Quality Assessment increased Keisha's allocation relative to Laurie ($7, p=0.85); and test 7 assessed the difference in these differences ($116, p=0.03). But the same assessment – regarding the effect of an excellent Worker Quality Assessment for a

white applicant compared to a black applicant – could have been conducted as follows: compare [1/7] to [1/8] to assess how much an excellent Worker Quality Assessment increased Laurie's allocation relative to Keisha ($64, p=0.09); compare [1/10] to [1/11] to assess how much an excellent Worker Quality Assessment increased Latoya's allocation relative to Keisha ($81, p=0.03); and assess the difference in these differences: in this case, the difference in differences is $16 instead of $116, and this $16 favors the hard-working black applicant relative to the hard-working white applicant (p=0.76).

DeSante (2013) tests 8, 9, and 10 form a group: test 8 compared [2/4] to [2/5] to assess how much a poor Worker Quality Assessment decreased Emily's allocation relative to Laurie ($21, p=0.55); test 9 compared [2/7] to [2/8] to assess how much a poor Worker Quality Assessment decreased Keisha's allocation relative to Laurie ($113, p=0.007); and test 10 assessed the difference in these differences ($92, p=0.09). But the same assessment could have been conducted as follows: compare [1/7] to [1/9] to assess how much a poor Worker Quality Assessment decreased Laurie's allocation relative to Keisha ($56, p=0.13); compare [1/10] to [1/12] to assess how much a poor Worker Quality Assessment decreased Latoya's allocation relative to Keisha ($112, p=0.007); and assess the difference in these differences: in this case, the difference in differences in favor of the white candidate is $56 (p=0.31).

DeSante (2013) test 11 compared the mean allocation to offset the state budget deficit when both applicants were white (conditions 4, 5, and 6) to the mean allocation to offset the state budget deficit when both applicants were black (conditions 10, 11, and 12): respondents allocated on average $107 more to offset the state budget deficit across conditions in which both applications were black compared to conditions in which both appli-

cants were white (p=0.005). Consistent with results reported in DeSante (2013), respondents on average allocated $64 more to offset the state budget deficit across conditions in which only one applicant was black compared to conditions in which both applicants were white (p=0.09).

The Influence of Racial Resentment on Experimental Results in DeSante (2013)

Table 3 of DeSante (2013: 352) reported results from regressions modeling allocations to offset the state budge deficit with racial resentment (RR) as an explanatory variable and/or as an explanatory variable interacting with the race of applicants in a condition, with WW indicating a condition with two white applicants, WB indicating a condition with one white applicant and one black applicant, and BB indicating a condition with two black applicants. Table 2 in this study reports results from models presented in the first three numeric columns of DeSante (2013) Table 3. DeSante (2013: 351) interpreted the model with the full set of interaction terms (Model 3) as follows:

> As seen by the large negative sign for racial resentment interacted with two white applicants (RR x WW), the presence of white applicants attenuates the effect of racial resentment on a desire for fiscal responsibility. Clearly, race matters when evaluating applicants for welfare and, when given an "acceptable" alternative to spending the money, those who are most racially resentful will allocate money to decrease a state's deficit, but at a far lesser rate when evaluating white applicants for welfare. In summation, those who are most racially resentful

are willing to spend much more on welfare when the applicants are both white than when applicants are black.

But the omitted category of conditions in Model 3 is the condition in which none of the applicants were given a name; this condition will be referred to as NN. Because the omitted category was NN, the included category variables must be interpreted in relation to this NN category (for included condition terms) or to the omitted RRxNN category (for included interaction terms): therefore, Model 3 results indicate only that respondents allocated less to offset the state budget deficit when both applicants were white than when both applicants were unnamed. Another computation must be conducted to assess the difference between the coefficient in conditions in which both applicants were white (RRxWW coefficient of -338) and the coefficient in conditions in which both applicants were black (RRxBB coefficient of -196); model 4 represents such a model: BB and RRxBB were omitted, so included categories should be interpreted relative to BB or RRxBB. The model 4 coefficient for RRxWW is -141 ($p=0.41$); the high p-value for this coefficient indicates a lack of evidence to support the inference of a difference in the effect of racial resentment on allocations to offset the state budget deficit in conditions in which both applicants were white compared to conditions in which both applicants were black.

Table 2. The Effect of Racial Resentment on Allocations to Offset the State Budget Deficit

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Intercept	-298.71* (100.58)	-291.31* (104.16)	-364.89* (123.96)	-194.55 (122.72)	-242.35* (96.39)
Conservative Ideology	56.49* (21.85)	56.97* (21.87)	58.09* (21.94)	58.09* (21.94)	47.57* (17.89)
Republican Partisanship	5.46 (11.30)	5.15 (11.28)	3.61 (11.32)	3.61 (11.32)	1.02 (9.36)
Household Income	9.30* (5.28)	9.54* (5.28)	9.54* (5.27)	9.54* (5.27)	7.20 (4.51)
Education	7.99 (13.09)	8.57 (13.08)	7.98 (13.07)	7.98 (13.07)	10.86 (10.92)
Age	2.81* (1.21)	2.76* (1.21)	2.60* (1.21)	2.60* (1.21)	2.45* (0.97)
Female	1.28 (34.63)	-3.03 (34.64)	-7.45 (34.70)	-7.45 (34.70)	17.40 (29.43)
Racial Resentment (RR)	413.13* (81.25)	417.93* (81.25)	551.42* (134.40)	355.00* (126.53)	416.82* (108.33)
Two White Applicants (WW)	---	-67.43 (48.89)	158.62 (129.23)	-11.72 (123.73)	71.04 (95.96)
Mixed Race Pair (WB)	---	-23.15 (46.75)	-11.83 (119.33)	-182.17 (113.96)	-132.09 (95.10)
Two Black Applicants (BB)	---	41.05 (46.25)	170.34 (118.14)	---	73.20 (92.43)
Unnamed Applicants (NN)	---	---	---	-170.34 (118.14)	---
RR x WW	**---**	**---**	**-337.92* (179.38)**	**-141.49 (172.13)**	**-179.17 (139.23)**
RR x WB	---	---	-17.47 (167.31)	178.96 (159.61)	157.28 (138.57)
RR x BB	---	---	-196.43 (165.60)	---	-17.99 (135.47)
RR x NN	---	---	---	196.43 (165.60)	---
Observations	627	627	627	627	820
R^2	0.16	0.17	0.18	0.18	0.16
Adjusted R^2	0.15	0.16	0.16	0.16	0.15

Note: Numeric cell entries are coefficients with standard errors in parentheses; to mirror asterisks in DeSante (2013), an asterisk (*) indicates statistical significance at the $p \leq 0.10$ level (two-tailed test). The sample in models 1 to 4 was restricted to respondents coded as white; the sample for model 5 was all respondents. Boldface indicates the key row regarding the inference that racial resentment interacted with the race of the applicants.

DeSante (2013) conducted the eleven t-tests described above using observations from all 1,000 respondents, which included 751 respondents coded as white, 96 respondents coded as black, 84 respondents coded as Hispanic, 12 respondents coded as Asian, 7 respondents coded as Native American, 20 respondents coded as Mixed, and 30 respondents whose racial category was coded as Other; however, DeSante (2013) reported the racial resentment analysis in Table 3 using observations from only respondents coded white. Model 5 reports the results of Model 3 when, as in the t-tests, the analysis used observations from all respondents: the coefficient on the key variable of RRxWW had a high p-value (p=0.20).

Combining Reported and Unreported Results from DeSante (2013)

Results reported in this study illustrate inferential selection bias: inferences drawn from analyses reported in DeSante (2013) were different than inferences drawn from a different set of analyses using the same data. But the unreported results are no better or worse than the reported results: the superior method for reporting results – and thus for fostering correct inferences – is to combine results for each hypothesis; this combination of results was conducted with the Stata 11 metan command, with the fixed effect option.[3]

The combined -$45 effect (p=0.10) for test 1 indicates evidence at a moderate level of statistical significance that the mean allocation was lower to unnamed applicants with no Worker Quality Assessment than to unnamed applicants with an excellent Worker Qual-

[3] Respective combined effect sizes and p-values for each test using a random effects model were: -$45 (0.10), +$123 (0.03), -$24 (0.15), -$50 (0.45), and -$74 (0.055).

ity Assessment. The combined +$122 effect ($p<0.001$) for test 2 indicates that the mean allocation was higher to unnamed applicants with no Worker Quality Assessment than to unnamed applicants with a poor Worker Quality Assessment. The combined -$34 effect ($p=0.001$) for tests 3 and 4 indicates strong evidence that black applicants with no Worker Quality Assessment received a higher mean allocation than white applicants with no Worker Quality Assessment. The combined -$49 effect ($p=0.20$) for test 7 indicates a lack of evidence at conventional levels of statistical significance that black applicants were rewarded differently than white applicants for an excellent Worker Quality Assessment, but the combined -$74 effect ($p=0.056$) for test 10 indicates evidence at a moderate level of statistical significance that black applicants were punished more than white applicants for a poor Worker Quality Assessment.

Reducing Inferential Selection Bias

These analyses support three recommendations to improve the reporting of research. First, researchers should pre-register research design protocols when possible: pre-registration permits readers to distinguish confirmatory tests from exploratory analyses and protects researchers from claims that reported results reflect p-hacking in search of a preferred result. Second, researchers should conduct a range of tests for a hypothesis when multiple tests are possible and should report at least a representative set of these tests so that readers receive a sense of the robustness of results to alternate reasonable research design specifications. Third, journals and other publication outlets should require researchers to make publicly available <u>all</u> collected data and code necessary to reproduce the analyses, so

that readers can assess for themselves whether reported results are representative of the range of possible results.

References

DeSante, Christopher D. 2013. "Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America's Deserving Poor." *American Journal of Political Science* 57(2): 342-356. http://hdl.handle.net/1902.1/20351 UNF:5:EEexoDfcqPKwaPVr7DS6Ow== V1 [Version].

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'p-Hacking' and the Research Hypothesis was Posited Ahead of Time." Draft dated November 14. Retrieved June 30, 2014, from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

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.