EJW
ECON JOURNAL WATCH
Scholarly Comments on
Academic Economics

ECON JOURNAL WATCH 11(3)
September 2014: 250-276

Replicability and Pitfalls in the Interpretation of Resampled Data: A Correction and a Randomization Test for Anwar and Fang

Dragan Ilić[1]

LINK TO ABSTRACT

In their article "An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence," published in the *American Economic Review* in 2006, Shamena Anwar and Hanming Fang (hereafter AF) study motor vehicle stops and searches by Florida Highway Patrol officers ("troopers"). Their data include the race and ethnicity of the trooper as well as of the motorist stopped and possibly searched. A search of a stopped motorist is deemed successful if the trooper finds contraband in the vehicle. Using data on troopers and motorists of three race-ethnicity groups (white non-Hispanic, black, and white Hispanic, with others being dropped), AF compute nine trooper-on-motorist search rates and nine trooper-on-motorist search-success rates. They present a model that exploits this information to test whether troopers go beyond statistical discrimination to racial prejudice.

The model has an implication that would be unaffected by whether troopers exhibit racial prejudice. This implication is testable and concerns the rank-order of the search and search-success rates. AF report that, across the board, the data neatly fit the model's predicted inverse rank-order implication, strongly supporting the soundness of the model.

AF then apply the model to address the question of racial prejudice. They do not find evidence of racial prejudice; in my own analysis, I, too, do not find

[1] University of Basel, 4002 Basel, Switzerland; University of Zürich, 8001 Zürich, Switzerland. I would like to thank Georg Nöldeke, George Sheldon, Michael Wolf, Alex Gray, Patricia Newark, and two anonymous referees for their helpful comments. I am also grateful to Hanming Fang, Pinelopi Goldberg of the *American Economic Review*, and an anonymous *AER* referee for the extensive and constructive discussion on a previous draft.

such evidence. The present critique, then, does not arrive at results about prejudice contrary to their results.

The present critique starts by reporting that I cannot replicate their preliminary inverse rank-order findings. For each of the nine trooper-on-motorist categories, AF report the search rate and search-success rate. However, I find that replication is not possible for two of the nine reported search-success rates. Correspondingly, replication is not possible for the reported statistical significance of four of the six Z-statistics and one of the three χ^2 test statistics for the rankings of the search-success rates. These new results obliterate the reported distinct pattern of the rates and imply that the empirical support for the model's soundness is not what AF claim it to be. In consequence, our confidence in the results obtained by employing the model to test for racial prejudice should be significantly reduced.

While the problem of irreplicability is my primary point, I then move on to another matter. My replications draw attention to a neglected statistical caveat in AF's implementation of the empirical tests of racial prejudice. The replications happen to show that the novel resampling procedure employed by AF does not provide robust results. I pinpoint the empirical source of the lack of robustness, and, in an appendix, show how a simple extension to their method improves robustness. In another appendix I put forth an alternative randomization test that seems more appropriate when testing such resampled data.

With all improvements, we still do not find evidence of racial prejudice. But now we know that our knowledge about the issue is poorer than one might have guessed from reading AF.

Recap of Anwar and Fang (2006)

When a highway patrol trooper stops a motorist, he or she faces a decision of whether to search the vehicle. Consider a police force with different trooper racial groups facing motorists classified by the same races. The model postulates that each trooper racial group is characterized by a specific cost of searching motorists. We say that a given trooper racial group is racially prejudiced if their search cost depends on the race of the motorist they search. For example, consider white troopers. Suppose their cost of searching white and black motorists were the same, while their cost of searching Hispanic motorists were *lower*.[2] This would be a case

[2] Beginning with this sentence and continuing through the end of this paper, and throughout all the supplemental materials to the paper, I follow AF in using the label "white" for the group of white non-Hispanics, the label "Hispanic" for the group of white Hispanics, and the label "black" for the group comprising black Hispanics and black non-Hispanics.

of racial prejudice, although it is unclear whether we would describe the prejudice as one against Hispanic motorists or as one in favor of white and black motorists.[3]

In addition to the cost of search, a trooper's decision to conduct a search depends on the likelihood of the stopped motorist being engaged in criminal activity. The trooper infers this probability from an informative but noisy signal emitted by the motorist during a stop. This 'guilt' signal captures all possible characteristics linking the motorist to criminal activity. Given the trooper's search cost, the strength of this signal has to exceed a certain threshold in order for the trooper to expect a benefit from searching.

For every combination of motorist and trooper racial groups, there exist equilibrium search and search-success rates that are both determined by a threshold value of the guilt signal. As a trooper, if I have a high search cost, then I had better expect a motorist to be guilty with a correspondingly high probability before I consider searching him. Because the guilt signal is informative, this implies that the lower the rate at which I search motorists, the higher will be my resulting search-success rate.

To illustrate this inverse relationship in more detail, suppose I am a white trooper and I do not harbor taste-based prejudice. On the postulates of the model, this means that my cost of searching a motorist is the same regardless of whether the motorist is white or black. My cost of searching a white motorist's car is no higher than my cost of searching a black motorist's car. Now suppose that I, a white trooper, do harbor taste-based prejudice against blacks. This may be thought of as a search-cost reduction for my searching black motorists, vis-à-vis white motorists. Such a search-cost reduction would lead to a guilt probability threshold for my searching black motorists lower than such threshold for white motorists. In other words, now a lower probability of guilt on the part of a black motorist (in comparison to a white motorist) satisfies the requirements to conduct a search. On the one hand, this raises the search rate towards blacks because now a greater fraction fulfills the search criterion. On the other hand, among that larger fraction, proportionally less are actually guilty than among the searched white motorists.

For a *given* race of troopers, differing search costs against the different motorist races translates into racial prejudice. But even without racial prejudice, search costs may differ *in general* between the trooper racial groups. That is to say, some trooper racial groups may have equally higher search costs against all motorist racial groups, which does not imply racial prejudice. To put it in AF's terms of the police force being either "monolithic" and "non-monolithic," a monolithic police force would not imply that there is no racial prejudice, and a non-monolithic police force would not imply that the police are racially prejudiced. Not only do Anwar

[3] The next section makes a brief detour and elaborates on this semantic issue.

and Fang allow for non-monolithic behavior in contrast to previous models, their model actually *exploits* such behavior in order to deliver testable implications about the presence of racial prejudice. Indeed, their model is not instructive if the police are, in fact, monolithic.

To understand how the model construes and infers "prejudice," consider a police force in which black troopers have higher search costs against white motorists than white troopers do. Assuming no prejudice, it then follows that the search costs of black troopers against black motorists are the same as they are against white motorists. In addition, the search costs of white troopers against black motorists are the same as they are against white motorists. By transitivity, it follows that the search costs of black troopers against black motorists are also higher than the search costs of white troopers against black motorists. Put differently, if not prejudiced, black troopers have generally higher search costs which are not associated with the race of the motorist, and thus the race of the motorist plays no role when ranking the search costs by trooper race. This independence translates to the search and search-success rates. Recall that these rates are monotonically linked to the search costs such that when there is no prejudice, the black troopers' search rates against any given race of motorists are smaller than the white troopers' search rates; and the black troopers' search-success rates against any given race of motorists are larger than the white troopers' search-success rates.

AF's test for racial prejudice assesses this predicted *rank independence*. If the ranking of the search or search-success rates depends on the race of the motorist, then racial prejudice on the part of the police can be deduced. Note that this inference of prejudice is relative because the method cannot determine which trooper racial group(s) is (are) prejudiced. At the same time, this ranking offers a test for the soundness of the model. Regardless of whether racial prejudice exists, this other testable implication predicts that for a given race of motorists, the rank order of the search and the search-success rates should always be exactly the opposite. In the above example, black troopers should always be the ones that are least likely to search against a given motorist group, but if they do, they should always exhibit the highest success. This fundamental implication is called the model's *inverse rank order condition*.

In their analysis, AF cannot reject the hypothesis that troopers of different races do not exhibit relative racial prejudice. That is, their data suggest that the rankings of the search and search-success rates by trooper race do not seem to depend on the race of the motorist. What is more, the inverse rank order condition is firmly satisfied in all cases. The reported Z-statistics from the rank order tests indicate distinct ranks in the predicted manner with high statistical significance ($p < 0.001$) across the board: AF report that white troopers display the highest search rates against any race of motorists, followed by Hispanic troopers. Black troopers are the least likely ones to perform a search. If black troopers search, however, they are the most successful group. In turn, Hispanic troopers have higher search-success rates than white troopers. A perfect fit, the reported pattern of these rank orders lends strong support to the descriptive validity of the model.

The validity of the empirical tests hinges on the assumption that the fraction of motorists of a given race carrying contraband does not depend on the race of the troopers searching them. The raw data, however, indicate that this assumption might not be empirically valid. White, black, and Hispanic troopers are dispersed disproportionately across the eleven regional troops in Florida and thus do not seem to face similar pools of motorists.[4] For this reason, the application of the empirical tests implements a clever novelty. AF introduce a sophisticated resampling procedure to create a reweighted data set that meets this assumption and serves as the basis for the empirical tests. To alleviate sampling error, this reweighted data set is the average of 30 independently drawn resamples using the procedure. This makes the search and search-success rates reported in AF the bootstrapped means from the corresponding rates calculated in each of the 30 draws. By the same token, every empirical test in AF is based on the average of the corresponding test statistics calculated in each of the 30 independent resamples.[5] In what follows, I refer to the execution of AF's procedure with 30 iterations as a "pass."

A few words on monolithic behavior and semantics

Before we proceed to the replication, a few words are in order for the reader that is unfamiliar to the literature. As noted already, we work with three racial groups: white, black, and Hispanic. The combinations for trooper-on-motorist make nine cells for the search and search-success rates, respectively. The previous section has shown that Anwar and Fang's model allows for the possibility that the trooper racial groups have different search costs against a given race of motorists, a behavior they dub "non-monolithic." In the context of such non-monolithic behavior, there is a basic assumption made in modeling trooper behavior, an assumption employed by AF and maintained throughout my own analysis, including my renovations. For the moment, consider only the search-success rate

[4] See Figure 1 in AF (2006, 142) for the troop locations.

[5] I return to the exact nature of the resampling procedure in a later section. I would like to thank Hanming Fang for thoroughly explaining the procedure and the empirical tests.

cells.[6] The modeling postulates, for example, that in the cell for Hispanic troopers searching white motorists, the cost of searching is the same for all troopers within that cell. That is, the postulate says that the cost to a Hispanic trooper of searching the car of a white motorist is the same, irrespective of which Hispanic trooper it is and which white motorist it is. The term non-monolithic is apt in that we have nine different combinations and the cost of search is allowed to differ among them, a generalization that sets AF's model apart from previous ones. But the term is a little misleading because *within each of the nine cells* the search-cost assumption is in fact monolithic. Put differently, there is heterogeneity across the nine cells, but homogeneity within each of them.

The reexamination shows that the data, in fact, should make us uncomfortable about the postulate of homogeneity within each cell.[7] But that is a weakness of my own analysis as well as AF's. It is, as it were, yet another reason to figure we do not really know what we seek to know (that is, whether racial prejudice plays a significant role in trooper behavior).[8]

The reader should also be alerted to the very distinct way of construing and modeling "prejudice" in this branch of the literature. I follow the semantic practice of AF and the preceding literature in talking of prejudice; see, for example, the seminal work by John Knowles, Nicola Persico, and Petra Todd (2001). In our semantics, prejudice is said to be present when troopers of a given race have search costs that depend on the race of the motorist. More precisely, a trooper is deemed prejudiced *against* group X if the search costs against a motorist of group X are lower than they are against a motorist of group Y.[9] With this approach of modeling prejudice, a biased trooper requires a lower guilt signal on the part of a group X

[6] The same reasoning applies to the search rate cells. More precisely, in what follows we are talking about the nine trooper-on-motorist search cost combinations, which uniquely determine both the search and the search-success rate combinations.

[7] In Ilić (2013, 50ff.), I elaborate on this issue of heterogeneity in the police force.

[8] The issue of homogeneity vs. heterogeneity also crops up in other dimensions. In another paper I show that aggregating police stop and search data across time and regions involves the danger of false conclusions when testing for racial prejudice with the established economic models (Ilić 2013). For example, when singling out troop G in AF's data, we cannot reject prejudice using AF's framework, a conclusion that drowns in their aggregate analysis. What is more, in troop C, the region with the largest number of searches, the inverse rank order condition predicted by AF's model is violated with statistical significance, a violation that refutes the model for these data. The same holds true for troops E and K. These violations are lost in the aggregate analysis, yet these three troops account for half the searches in the aggregate data.

[9] This notion of prejudice is based on the idea of taste-based discrimination as introduced by Becker (1957). Economists crucially distinguish between this malevolent form of discrimination and statistical discrimination (Arrow 1973; Phelps 1972). Statistical discrimination is an efficient technique of optimal signal extraction that exploits information on group membership. In contrast to taste-based discrimination, statistical discrimination does not enter the utility function of the decisionmaker and does not reflect malevolent intent.

motorist in order to trigger a search. One could also argue that the trooper draws utility from disadvantaging a motorist of group X by means of searching them. Yet by the same token, one could argue that the trooper is prejudiced *in favor of* group Y because the trooper cuts even relatively suspicious group Y motorists some slack, or because the trooper would draw disutility from annoying a group Y motorist.[10]

Although the idea of favoring a group is mentioned in the early literature, it comes up only in connection with favoring black motorists from fear of future litigation when searching them (Knowles, Persico, and Todd 2001, 227).[11] The notion of actively favoring in terms of sympathy only emerged with additional empirical information on trooper race (Close and Mason 2007). Favoring is not explicitly brought up in AF. The problem with favoring is that it would undo a researcher convention of the anchoring of treatment. As described in the above example, it might well be that a trooper is not prejudiced against motorists of group X despite the lower search costs. This is the case if these search costs actually reflect the *unbiased benchmark*. The trooper might simply favor group Y, and that is all there is to it. This semantic difference has consequences for the interpretation of the data in AF's framework. If the observed rank orders are not consistent with the hypothesis of no relative racial prejudice, then one cannot readily say whether these results imply the presence of malevolent prejudice or preferential prejudice. All one can deduce is that there is something racially non-neutral in police behavior. So when AF stress that their model can only detect relative racial prejudice (because one cannot say which trooper race(s) are biased), it should also be clarified that furthermore, the model cannot distinguish between favoritism and animus if it detects prejudice. This bears importance for policy recommendations.

[10] Construing racial prejudice by the level of the search costs is not without problems. It could be that race-specific search costs are affected by reasons other than prejudice. Suppose that it is known among the police that Hispanic motorists are the most dangerous group to search. If troopers take this into account, the search costs against Hispanic motorists will rise. This alone does not pose a problem for the analysis in AF's framework as long as all troopers feel equally threatened. For in that case, the *rank order* of the search and search-success rates against Hispanic motorists will not change. But suppose that this peril looms only or particularly for a certain racial trooper group, say white troopers. Then for this combination only, danger would affect the search and search-success rate similarly to (preferential) prejudice. A violation of the rank order independence in AF's test would then mistakenly indicate relative racial prejudice in the police force.

[11] This issue relates to footnote 10. AF's test of prejudice is not affected if the fear of litigation is shared by all troopers alike. If, however, white troopers are particularly driven by this fear, we might mistakenly infer relative racial prejudice.

Replication

The meaning of replication requires some clarification. Because the reported search and average search-success rates are calculated via AF's novel resampling procedure, they are stochastic and vary to some extent in each iteration and thus from pass to pass. The same reasoning applies to the test statistics. An exact replication of AF's results is therefore unlikely. To account for the stochastic leeway in the replication, I have automated AF's tedious task of manually processing the 30 iterations that make up one pass and have conducted 10,000 independent passes. In other words, I have calculated essentially all the possible results that the resampling procedure can produce with AF's data.[12]

The replications expose two problems in AF's paper. First, two of the nine reported average search-success rates cannot be replicated in that they do not fall within the domain of possible outcomes. Second, in the same vein, four of the six Z-statistics used in the rank order tests and one of the three χ^2 test statistics used in the preceding test of monolithic trooper behavior cannot be replicated. As a consequence, these test statistics no longer reject the respective null hypotheses of equal rates.[13] Taken together, these two issues negate the empirical support for the model.

Consider first the replication of the nine estimated average search-success rates, which, in AF, are reported in Panel B of their Table 1 (2006, 130). The frequency distributions that I obtained by the automated replications of the rates using the resampling procedure are shown in Figure 1. For ease of comparison, the arrangement is in line with the combinations of motorist and trooper racial groups in AF's Table 1. That is, the left, the middle, and the right column depict white, black, and Hispanic troopers, respectively. In turn, white, black, and Hispanic motorists are arranged by upper, middle, and lower row, respectively. So for instance, the upper left distribution shows that the bulk of the 10,000 indepen-

[12] The 10,000 automated replications are calculated using AF's original Stata resampling algorithm and employ their data, both of which are available at the *American Economic Review* website. I have used Stata version 13 and, for a previous draft, version 11. In keeping with AF's code, no specific seed was set prior to the resampling. Setting specific seeds or using truly random seeds via the Stata package *setrngseed* did not affect the general results from the replication. Appendix 3 links to an online resource that provides a more detailed description of my replications including additional data, codes, and figures. Among these additional data are the frequency distributions of the replicated search rates, which do not show any deviation from AF's reported values and are thus omitted from the discussion in this paper.

[13] This second issue does not emerge because of the first one, the two irreproducible average search-success rates. On the contrary, the rates reported in AF would even render five of the six rank orders indistinguishable.

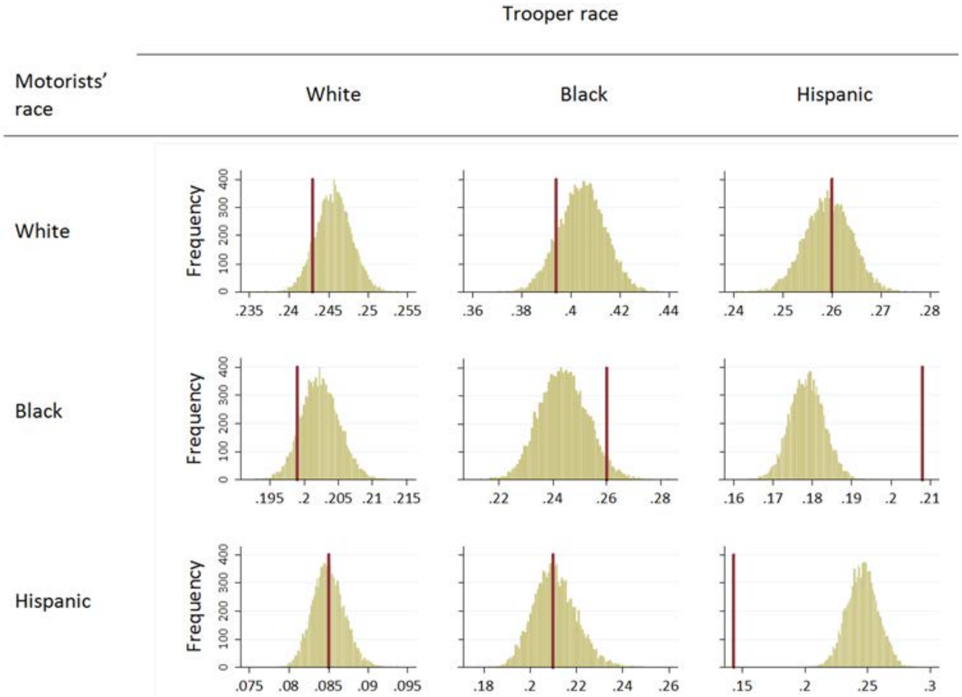
dently estimated average search-success rates of white troopers against white motorists falls between 24 and 25 percent. This is consistent with AF's particular pass that yielded 24.3 percent, indicated by the vertical red line: These lines in Figure 1 are AF's reported estimates of the average search-success rates.

Two of the nine reported rates (the red lines) cannot be replicated in this way. Figure 1 shows that AF's estimated average search-success rates of Hispanic troopers against black and Hispanic motorists, respectively, fall outside the computed ranges: In contrast to the reported 20.8 percent, the replications place the possible average search-success rates of Hispanic troopers against black motorists between 17 and 19 percent. And against Hispanic motorists, the possible rates of Hispanic troopers range from 21 to 28 percent. At 14.3 percent, the reported value lies below this spread.[14] Put differently, these two reported rates cannot be squared with the data even when accounting for the variation in possible outcomes.

In contrast to the reported pattern in AF, the replications of the average search-success rates displayed in Figure 1 no longer provide empirical support for the inverse rank order condition predicted by the model. Recall that the pattern of the *search rates* in the data predicts that, against any given race of motorists, black troopers should search with the most success, followed by Hispanic troopers. White troopers should display the lowest average search-success rates. AF's values, indicated by the red lines in Figure 1, fit this prediction perfectly. Both the two irreproducible rates, however, run afoul of this prediction. On the one hand, the replications disclose that Hispanic troopers are the least successful ones when it comes to searching black motorists. On the other hand, the replications also reveal that they are the most successful trooper group against Hispanic motorists. At first glance, this seems to have severe consequences for the model. Given the scale of the *Z*-statistics associated with the relatively small differences in means reported in AF (p. 146), the new rates would not only revoke the empirical support for the model. They would actually violate the inverse rank order condition with high statistical significance and would thus formally refute the model (p. 138).

[14] The standard errors reported in AF's Table 1 do not provide a measure for the significance of the difference between the reported values and the replications. They are the bootstrapped standard errors of 30 independently drawn means and thus reflect the volatility of the rates *within* AF's particular pass. In contrast, Figure 1 illustrates the volatility *among* independent passes.

Figure 1. Frequency distributions of replicated average search-success rates



This takes us to the second issue with respect to irreproducibility. The empirical tests reported in AF support all predicted rank orders with high statistical significance. For example, AF test whether the difference in the average search-success rates of white and Hispanic troopers against white motorists (24.3 and 26 percent, respectively) is different from zero. They report a *Z*-statistic of −324.1, making a clear case for a distinct rank order. The other reported *Z*-statistics are in the same ballpark.[15] My replications, however, show that the data cannot account for these magnitudes. On the contrary, most rank orders of the average search-success rates turn out not to be statistically significant, a result that also happens

[15] Like the average search-success rates, the empirical tests are based on average test statistics. In a first step, the test statistics are calculated independently in each of the 30 reweighted samples which make up the pass. The average of these 30 test statistics is then used to test the corresponding null hypothesis. For ease of comparison with the wording in AF, I will not explicitly refer to the test statistics as "averages." Although they do not report all (average) *Z*-statistics, AF "find that the evidence supports" all predicted rank orders (p. 146). On a more fundamental note, the implementation of AF's empirical tests raises a question of inference. It is not obvious that their implementation is applicable in the context of averaged resampled data. On that account, Appendix 2 presents a randomization test (a straightforward way to test differences of average rates in a resampling).

to render the aforementioned violation of the inverse rank order condition merely descriptive.[16]

Figure 2. Frequency distributions of replicated *Z*-statistics from AF's pairwise differences in means tests of search-success rates by trooper race for a given race of motorists (null hypothesis: no difference)

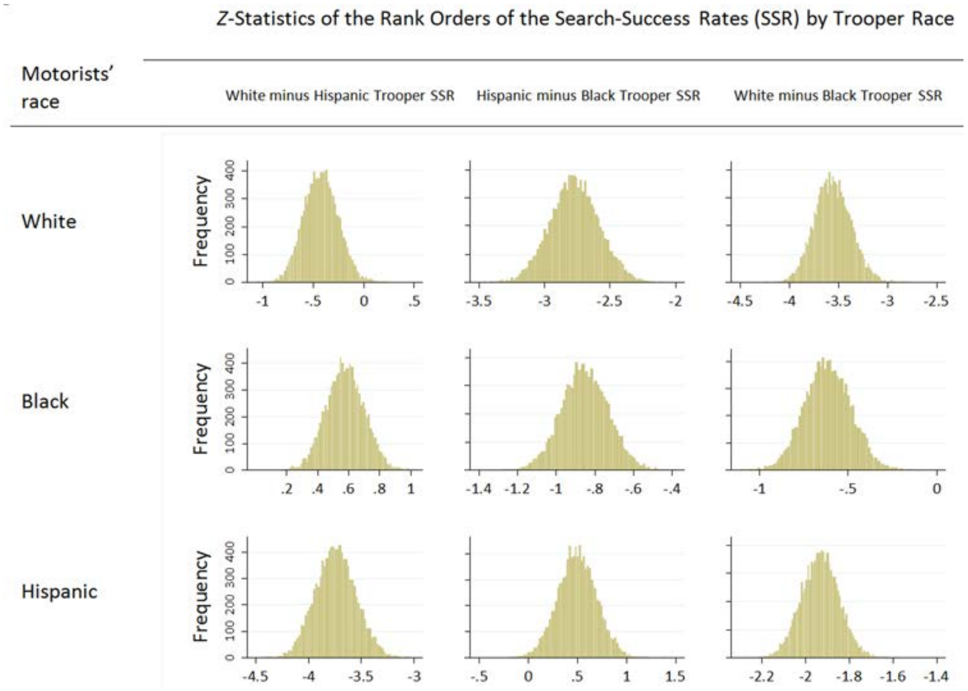


Figure 2 depicts the frequency distributions of all replicated *Z*-statistics from the pairwise rank order tests of the average search-success rates, again based on 10,000 passes. The first two columns replicate AF's six rank order tests for the average search-success rates against white, black, and Hispanic motorists, respectively, which are listed by row. The first column tests whether the difference of the average search-success rates between white and Hispanic troopers is zero. The second column does the same for Hispanic and black troopers. As additional

[16] AF repeatedly stress that if, for a given race of motorists, the ranking of the search-success rates is not "exactly the opposite" of the ranking of the search rates, the model is refuted (pp. 131, 136, 138, 140, 146). The replications show that this exact opposite is no longer observable in the data. However, a descriptive observation of violation alone does not immediately imply that the rank order condition is actually violated, which is AF's formal condition. In other words, there is an empirical difference between statistically significant violation, statistically significant support, and lack of statistical support for the inverse rank order condition.

evidence to AF's tested rank orders, the third column tests the difference between AF's first and third rank, that is to say, black and white troopers. Despite the spreads, each distribution of possible *Z*-statistics in Figure 2 paints an unambiguous picture in terms of statistical significance when considering conventional significance levels.[17] The outcomes show that the statistical significance of four of the six reported rank order tests for the average search-success rates cannot be replicated. For instance, in contrast to the aforementioned *Z*-statistic of −324.1 when testing the difference in the average search-success rates of white and Hispanic troopers against white motorists, the upper left distribution indicates possible outcomes between −0.9 and zero, values that cannot reject the null hypothesis of equal rates.[18] Two of the six rank orders remain consistent with the reported statistical significance in AF, albeit at lower levels. First, the difference in the average search-success rates of black and Hispanic troopers against white motorists. And second, as a coincidental consequence owing to the new value of the replicated average search-success rate depicted in the lower right distribution in Figure 1, the difference in the rates of white and Hispanic troopers against Hispanic motorists becomes statistically significant. In contrast, AF's value at 14.3 percent would not have rejected the null.[19]

Finally (not depicted), one of the three χ^2 test statistics from AF's test of monolithic trooper behavior with respect to the average search-success rates cannot be replicated. This test precedes the rank order tests and, in showing that the trooper racial groups exhibit a distinctive stop and search behavior on the whole, lays the foundation for the application of the rank order tests. At the same time, it highlights the model's advantage in comparison to the seminal framework by Knowles, Persico, and Todd (2001). When testing for monolithic behavior against black motorists, Table 1 in AF indicates a *p*-value of <0.001, rejecting the notion that the troopers behave differently against black motorists. Yet the replicated frequencies of successful and unsuccessful searches based on 10,000

[17] Except for maybe the lower right corner, which depicts the frequency distribution of the *Z*-statistic for the difference of the average search-success rates between white and black troopers against Hispanic motorists: four of the 10,000 passes yield an average *Z*-statistic above −1.64 and would thus fail to reject the null hypothesis at the five percent level.

[18] This *p*-value from the replication corresponds to results reported in Knowles, Persico, and Todd (2001), who test for similar differences in search-success rates with a comparable sample size. For example, when testing for the difference in the rates against black motorists (34 percent in 1,007 searches) and white motorists (32 percent in 466 searches), they cannot reject a difference of zero (by means of a χ^2 test). In comparison, using the resampled sample size from a random iteration, I cannot reject that the difference between the rates of white (24.6 percent in 1,846 searches) and Hispanic troopers (23.2 percent in 211 searches) against white motorists is zero.

[19] The replications show that the reported test statistics are also disproportionate for the search rates (see Appendix 3). But in contrast to the average search-success rates, this does not alter the corresponding significance levels.

passes yield possible χ^2 values between 1 and 2.5, implying that the three average search-success rates against black motorists are not likely different from each other. The *Z*-statistics for the rank order tests against black motorists in the second row of Figure 2 support this inference. A back-of-the-envelope calculation shows that this new value of the χ^2 test statistic is not due to the new average search-success rate estimate of Hispanic troopers against black motorists.

Upon reexamination, then, the data no longer indicate a discernible pattern of the rank orders of the search-success rates. This does not refute the model. The replications do, however, rescind the reported strong empirical support.

But the replications raise yet another issue. The variation among the estimated average search-success rates and the estimated test statistics provided by the resampling procedure gives reason to reconsider the conclusiveness derived from the empirical tests. Does robustness pose a serious problem in AF's data? Figure 2 shows that despite the spread in the estimated test statistics, the statistical inferences from AF's data (as measured by conventional significance levels) do not depend on the outcome of the resampling. Figure 1, on the other hand, indicates a slight overlap in the distributions of the estimated average search-success rates of white and Hispanic troopers searching white motorists. So depending on the particular pass, the estimated rates may give even less descriptive support for the inverse rank order condition. But by and large, things do not look bad in AF's data despite the imprecision of the estimates.

Other data might be less forgiving. The volatility of the estimates opens up the possibility that the same data can give rise to conflicting conclusions. For one, the rank order tests on the basis of the resampling procedure could erratically indicate the presence or absence of racial prejudice. This is primarily a concern if one uses only search data in the empirical tests.[20] Because search data have smaller sample sizes than stop data, they are more prone to volatile outcomes via the resampling procedure. Overlaps in the frequency distributions of the possible outcomes could then randomly imply (in-)dependence of the rank order for a given race of motorists, indicating the (absence) presence of racial prejudice. An additional issue arises when using both stop data and search data for additional evidence, such as AF do, i.e., to test the soundness of their model via the inverse rank order condition. When doing so, fickle outcomes might sometimes lend (some) support to the model, only to refute it in another pass by violating the inverse rank order condition with statistical significance. Such caprice is vexing. In Appendix 1, I show that raising the number of iterations is a simple solution to

[20] AF point out that, in principle, the rank order test can be implemented with only search data (2006, 131 n.11).

mitigate the risk of reaching arbitrary conclusions. The next section sheds light on the empirical source of the nonrobustness of the estimates.

Resampling procedure and disaggregated trooper data

The considerable range of possible outcomes produced by the resampling procedure raises the question of what is triggering the volatility. Toward the answer, this section first describes the resampling procedure in detail. I then look at the trooper search pattern and racial trooper locations in AF at a disaggregated level, which turn out to be the decisive empirical factors that drive the precision of the estimates.[21]

In each troop, AF's resampling procedure randomly draws a subsample (without replacement) for each trooper race in relation to their aggregated proportion in the data. As an approximation, AF use proportions of 75, 15, and 10 percent for white, black, and Hispanic troopers, respectively.[22] Through the trooper identifier, these subsamples are subsequently merged with the raw stop and search data, forming the sample stop and search data. Put differently, the resampling procedure prescribes a number of draws for each trooper race in each troop and only keeps those observations from the raw stop and search data that are carried out by troopers who were drawn in the resampling. From the sample stop and search data, the aggregate number of stops and (successful) searches are tabulated for each trooper/motorist race combination, yielding the search and search-success rates. These rates are then tested for non-monolithic behavior and for differences in means. To alleviate the sampling error caused by the random draws, AF conduct 30 iterations of independent resamplings, taking the average of the corresponding search and search-success rates and the test statistics from each iteration. The previous section has highlighted that a statistical problem arises in this procedure. Despite averaging over 30 iterations, the values provided by this method fluctuate substantially.[23]

[21] Recall that AF employ the resampling procedure because the raw data indicate that troopers of different races are not randomly assigned to motorists of different races. Depending on the data, the empirical tests may well be applicable without any prior resampling.

[22] The exact shares for these groups in the data are 76.3 percent, 13.7 percent, and 10 percent. AF maintain strict multiples of 75/15/10.

[23] In an exchange Hanming Fang mentioned that the size of the reweighted samples was an issue for their computers at that time, driving the choice for 30 samples.

TABLE 1. Trooper distribution and sample ratios by troop and race

Panel A: Trooper distribution				Panel B: Sample ratios			
Trooper race				Trooper race			
Troop	White	Black	Hispanic	Troop	White	Black	Hispanic
A	120	7	2	A	0.125	0.429	1
B	88	8	3	B	0.170	0.375	0.667
C	155	22	13	C	0.581	0.818	0.923
D	176	26	20	D	0.682	0.923	0.800
E	68	39	58	E	0.882	0.308	0.138
F	125	8	9	F	0.240	0.750	0.445
G	105	17	4	G	0.286	0.353	1
H	62	8	0	H	-	-	-
K	81	17	18	K	0.926	0.882	0.556
L	91	45	15	L	0.989	0.400	0.800
Q	41	4	5	Q	0.366	0.750	0.400

One can show that the dispersion is driven by the underlying heterogeneous trooper search behavior. AF's trooper data set contains information on 1,469 troopers conducting 8,976 searches. In the resampling, the variables of interest are their race and troop assignment. Define the sample ratio as the prescribed number of troopers of a given race in the subsample divided by their actual number in that troop. Panel A in Table 1 tabulates the race/troop allocations in the raw trooper data, which pin down the sample ratios in Panel B.

The variation in the sample ratios captures the differences in the racial composition of troopers between the troops. In each troop, the most underrepresented trooper race sets the bar for the sample ratios of the other racial groups. Consequently, troops that are disproportionate in comparison to the racial proportion of the entire police force induce lower sample ratios.[24] For example, because of the relative Hispanic dominance in troop E, a Hispanic trooper only has a 13.8 percent chance of being selected into the subsample. On the other hand, the presence of merely two Hispanic troopers in troop A severely limits the sample ratio of their white colleagues: While the Hispanic troopers in troop A do not undergo any resampling, a white trooper is drawn with a probability of 12.5 percent. Troop H illustrates the extreme case of disproportion. Its lack of Hispanic

[24] Because of AF's adherence to strict multiples of 75/15/10 and the low numbers of observations in some troops, not all troops contain a bar-setting sample ratio of one.

troopers leads to the omission of the entire troop in the resampling procedure, discarding its share of observations in the data.[25]

In addition to the racial disproportion between troops, the trooper data reveal a striking imbalance in the number of searches at the individual level. It turns out that 742 of the 1,469 troopers never search and drop out when merging the trooper subsamples from the resampling procedure with the raw search data. Of the troopers actually contributing to the aggregated search data, 727 conduct at least one search, 530 at least two, and 431 at least three searches. When considering only troopers with more than ten searches, 194 remain. The dots in Figure 3 visualize this heterogeneous search behavior. Each dot represents one of the 727 troopers that has conducted at least one search. The x-axis denotes the number of searches per trooper and the left y-axis measures their cumulative distribution. The skew highlights that most troopers rarely search, but a few do so vigorously.[26]

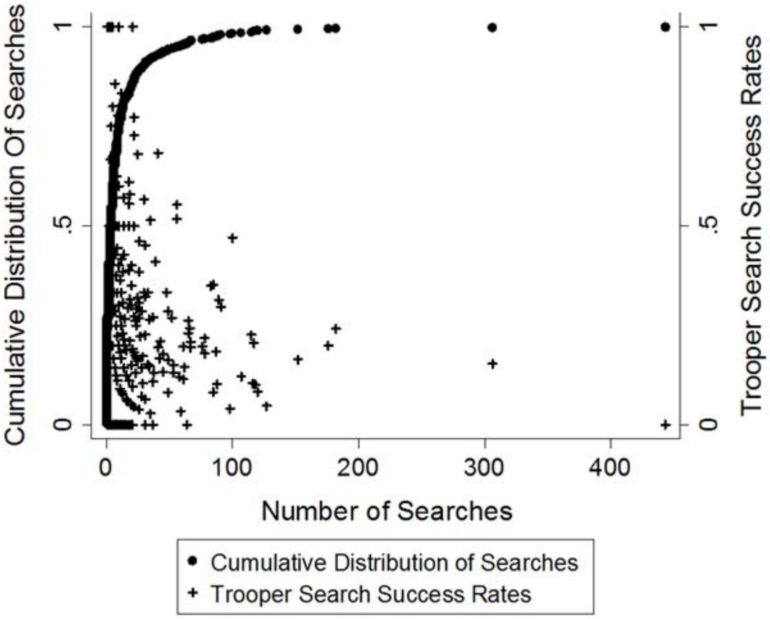
Figure 3 also incorporates data on individual search-success rates associated with the total number of searches conducted by each trooper. Measured on the right y-axis, each plus represents a trooper's search-success rate that corresponds to her dot on the same latitude. Crucially, the data suggest a negative relationship between the number of searches and the search-success rates, a finding which is independent of trooper race. In general, the more searches a trooper conducts, the smaller the overall chance is of uncovering engagement in criminal activity. This relationship affects the precision of the estimates provided by the resampling procedure because, for any troop, the draws within the racial groups give each trooper the same probability of becoming part of the subsample without consideration of her particular search-success rate and, more importantly, her number of searches. On a different note, the negative relationship between the

[25] There are two ways to increase the sample ratios. I was able to obtain an updated trooper data set from the Florida Highway Patrol, which contains information on 122 additional troopers covering the same time frame. The new data improve the racial balance in disproportionate troops, doubling most sample ratios. Moreover, troop H can be kept in the resampling procedure due to the presence of six Hispanic troopers. Alternatively, starting from the most underrepresented group, the numbers drawn in the resampling for the other groups could be rounded to the nearest integers in relation to their overall proportion. Depending on the troop and trooper race, the probability of being selected into the subsample could accordingly be increased by almost 50 percent. As in AF, the empirically testable model assumption that the troopers face the same pool of motorists determines the applicability of this and other alternative ways to increase the sample ratios. Nevertheless, neither the new data nor the laxer proportion requirement change any of the conclusions in this paper. I would like to thank John Knox and Richard Taylor from the Florida Highway Patrol for their support in obtaining the additional data.

[26] This observation relates to a generalized model in Persico and Todd (2006). They prove that the test for racial bias provided in Knowles, Persico, and Todd (2001) does not break down in the presence of heterogeneity in police search costs or intensity of racial bias. However, their model rules out environments in which, for example, white troopers are biased against black motorists and, at the same time, black troopers are biased against white motorists. I would like to thank an anonymous referee from the *American Economic Review* for bringing this to attention.

number of searches and the search-success rates qualifies the model assumption of monolithic behavior within any given racial trooper/motorist group combination.

Figure 3. Trooper heterogeneity in searches and search-success rates



As an illustrative example of how this relationship affects the precision of the estimates, consider a troop with three troopers of race X. Let trooper x_1 conduct 99 searches, 33 of which are successful. Troopers x_2 and x_3 each conduct three searches, two of which are successful. Trooper x_1 searches much more often than x_2 or x_3 but, relatively, does so with less success. Let the sample ratio be ⅔ and draw the corresponding subsamples. The aggregated search-success rates for the three possible subsamples are 34.31 percent, 34.31 percent, and 66.66 percent. With independent resampling, the average search-success rate converges to 45.10 percent. The inclusion of x_1 in a subsample introduces a bias in the aggregated rate towards x_1's rate and stems from her disproportionate share in the aggregated number of searches. So the spikes in the aggregated search-success rates in the subsamples are caused by trooper x_1.

The example stresses that if most sanctions are conducted by a minority of troopers, the average rate is biased towards their rates. Should these eager troopers exhibit systematically deviating success rates (as Figure 3 indeed suggests), they increase the variance of the estimated rates among iterations and, to a lesser degree, among the average search-success rates between distinct passes. This results in a decrease of precision in the estimated rates. Figure 3 gives an idea of the impact a single trooper can exert on the average search-success rates.[27] The extent of the instability such troopers can evoke in the resampling depends on their probability of becoming part of their subsample. The lower the sample ratio, the lower is the probability of a trooper being selected. In practice, this depends on the empirical distribution in trooper race across troops, as seen in Table 1.

The selection probability also depends on the proportion of non-searching troopers. The data show that only every other trooper ever conducts searches. Accordingly, among the subsample of drawn troopers, only a fraction provides actual data for the calculation of the search-success rates. For example, of the 39 black troopers in troop E, 12 find their way into the subsample. Yet out of these 39 troopers, as few as eight conduct searches. In a random draw, it is unlikely for them to be selected simultaneously into the subsample of 12. One can show that most likely, the subsample will only include one, two, three, or four searching troopers (with probabilities of 0.17, 0.32, 0.29, and 0.14, respectively). Thus in addition to the sample ratio, non-searching troopers further limit the presence of searching troopers in the subsamples, amplifying the fluctuations of the estimates provided by the resampling procedure.

To sum up, the interaction of non-searching troopers, the sample ratios, and the negative relationship between the number of searches and the search-success rates decreases the precision of the estimates and explains the large ranges of possible outcomes provided by the resampling procedure displayed in Figure 1 and 2. Appendix 1 shows that raising the number of iterations is an obvious and easily implementable solution to enhance the precision of the estimates, mitigating the risk of *pass dependence* in general and, in turn, lowering the risk of false conclusions from the data.

Conclusion

The replications in this paper do not bear out the empirical results reported in Anwar and Fang (2006). In contrast to the predicted inverse rank orders of the

[27] One trooper in the data stands out with a total of 443 searches conducted, and all of his 443 searches are listed as being unsuccessful. These numbers are startling and raise questions about data error. Richard Taylor, Operation and Management Consultant at the Florida Highway Patrol, supports the assumption of erroneous data for this particular trooper as he could not find any corresponding drug arrest documents. However, for the purpose of this paper I have refrained from modifying AF's data set. Suffice it to say that excluding this white trooper's observations from the data raises the white troopers' average search-success rates by roughly one to two percent (depending on the motorist racial group). This does not change the conclusions from my replications.

search-success rates which are firmly buttressed by the empirical tests conducted in AF, the data no longer reveal that distinct pattern and therefore do not provide empirical support for the model. That does not take away from AF's theoretical contribution. It does point out, however, that the data do not nearly fit their model as well as previously thought. In this sense, AF's main empirical conclusion that the police do not exhibit racial prejudice stands on less firm ground.

This paper also draws attention to a neglected statistical problem that affects the interpretation of the empirical results. Because the data do not seem to satisfy a crucial condition of the model, AF make use of a novel resampling procedure to create a reweighted data set. It turns out that the estimates provided by this procedure lack precision. Although AF's replicable results are only affected qualitatively, the imprecision creates a non-negligible risk of severely misinterpreting other resampled data. Depending on the outcome of the resampling, one might infer racial prejudice when there is none (or vice versa). And more fundamentally, one might support or reject the model when there is no reason to do so. Resampling with 30 iterations as conducted by AF seems too few to yield conclusive estimates.

In Appendix 1, I show how simply raising the number of iterations improves robustness. There is no general rule how many iterations are needed for conclusive results, but the existing bootstrap literature suggests that 1,000 replicates should suffice. On another note, it is not obvious that the parametric tests employed in AF are appropriate to test the complex data obtained by the resampling procedure. To inform future research further, Appendix 2 presents a randomization test that provides an alternative and more expedient way to empirically test the observed rank orders. A randomization test seems more appropriate than conventional statistical tests for it makes no assumptions about the distribution of the resampled data. In light of today's computational power, both raising the number of iterations for higher accuracy and randomization no longer pose a problem and can be readily implemented in existing software.

The statistical problem is not confined to the empirical tests employed in AF's particular model. Any empirical test based on a theoretical framework that assumes that heterogeneous decisionmakers (troopers) face agents (motorists) from the same quality pool is a candidate for resampling when the data call for it. More precisely, when there is variation within the data suggesting that the decisionmakers are systematically assigned to different groups of agents, an aggregation problem occurs. It is because of regional assignment of the troopers that AF have resorted to resampling. Resampling the data ensures that, on average, the decisionmakers face the same pool of agents. Such resampling is not restricted to geographical location. One might also resample data along other dimensions, such as time of day, year, or cohort. The results in this paper advise researchers to take into account the accuracy of their estimates before interpreting any resampled data.

More conclusiveness is clearly desirable to mitigate the neglected risk of jumping to false conclusions, such as when assessing racial prejudice among a police force. But the robustness has yet another merit. It prevents malicious cherry-picking of a particular outcome that suits a given agenda. Suppose biased researchers are aware that the possible outcomes of the resampled data support two diametrically opposed interpretations. In that case, they might deliberately report the convenient but wrong interpretation, an interpretation which is replicable at that and which, for this very reason, would leave them unscathed.

Summary of appendices

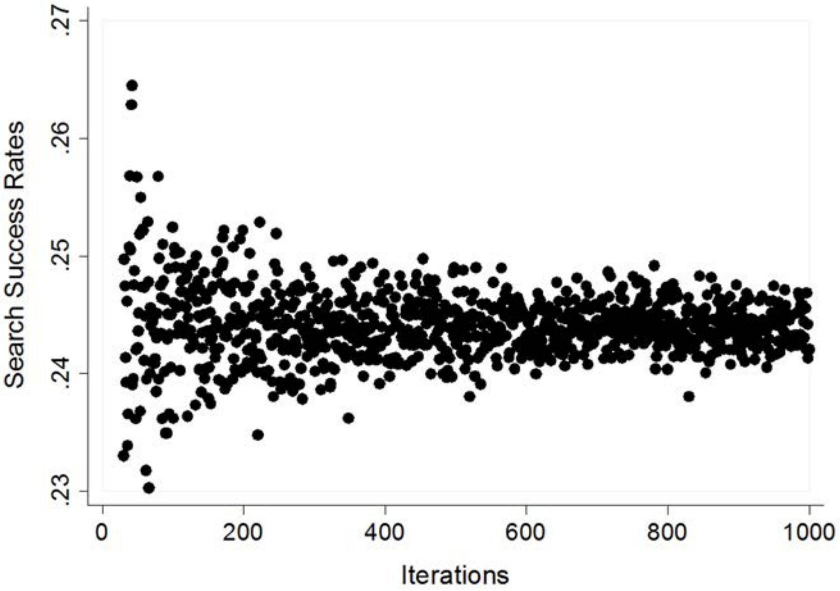
There are three appendices. The first appendix presents a straightforward and easily implementable solution to enhance the precision of the estimates provided by AF's resampling procedure. The second appendix puts forth a randomization test and argues that it is a more expedient way to test differences of average rates in a resampling. The third appendix provides a guide to the data and code files, all available for download.

Appendix 1: Generalizing the resampling procedure

AF's particular resampling procedure is reminiscent of more general bootstrap and jackknife methods. As a matter of fact, by randomly deleting a prescribed number of troopers of a given race in each troop, AF unknowingly apply a so-called delete-d jackknife. Chien Wu (1990) describes its statistical properties, such as asymptotic behavior, efficiency, and consistency. However, none of these properties are of direct use for AF's implementation, for two reasons. First, each troop undergoes three distinct delete-d jackknife draws, which are subsequently merged with the ones from the other troops to create a comprehensive mean based on aggregated individual observations. This mean is then averaged over 30 iterations. It is not readily obvious which distribution such a statistic follows. Second, the jackknife allows for inferences about the statistical properties of an original point estimator. In contrast, AF's resampling procedure makes use of its resulting distribution to construct an estimator in the first place.

All the same, akin to more general resampling techniques, the precision of the estimates provided by AF's procedure can be improved by simply raising the number of iterations in a pass. By the Central Limit Theorem, this results in the estimated average search-success rates being distributed more closely among different passes. Figure 4 illustrates this convergence by taking the example of black troopers searching black motorists. From *n* = 30 to 1,000 iterations measured on the x-axis, each dot depicts the estimated average search-success rate resulting from a pass with *n* number of iterations.

Figure 4. Estimated average search-success rates for increasing numbers of iterations

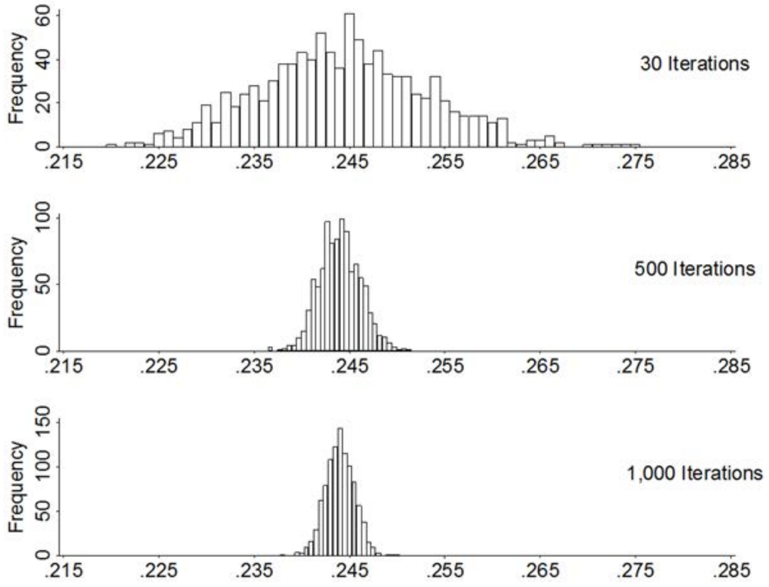


The consolidating pattern confirms that raising the number of iterations increases the precision of the estimated average search-success rate: Whereas the estimates sway from 23 to almost 27 percent when using up to 100 iterations, with a larger number the rates become increasingly bounded between 24 and 25 percent. Because only the results from one particular pass for each number of iterations are depicted, Figure 4 does not illustrate the *distribution* in possible outcomes for each number of iterations. Therefore, one cannot make out any actual confidence intervals like in Figure 1 or 2. Still, since each pass is a random draw from the probability distribution of passes with that specific number of iterations, the overall pattern of the dots gives a rough picture of the progress of the underlying precision.

Figure 1 showed the frequency distributions in average search-success rates for 30 iterations. Calculating these distributions for *all* numbers of iterations in Figure 4 is computationally not feasible, but Figure 5 shows the increase in precision of the estimated average search-success rates of black troopers searching black motorists by comparing the frequency distributions for 30, 500, and 1,000 iterations.[28] From 30 to 1,000 iterations, the 95 percent confidence interval (95-CI) for the estimated average search-success rate of black troopers against black motorists shrinks from [0.2278, 0.2613] to [0.2410, 0.2470]. Note that in raising the number of iterations, AF's reported rate of 0.26 falls outside of the estimated ranges.

Finally, Table 2 reproduces Panel B in AF's Table 1 using 1,000 instead of 30 iterations. Like the rates in AF, the rates in Table 2 stem from one particular pass and are therefore random. However, because the possible ranges into which these estimates can fall are now considerably narrower, the results are more robust.

Figure 5. Relationship between precision of estimated average search-success rate and number of iterations used



[28] A standard desktop computer completes one pass with 1,000 iterations in six minutes. Calculating the distributions for every number of iterations between 30 and 1,000 with 1,000 passes each would therefore take approximately 34 years.

TABLE 2. Estimated average search-success rates with 1,000 iterations

	Trooper race			
Motorists' race	White	Black	Hispanic	*p*-value
White	0.2456 (0.0096)	0.4056 (0.0426)	0.2600 (0.0288)	<0.001
Black	0.2025 (0.0140)	0.2420 (0.0600)	0.1789 (0.0406)	0.7318
Hispanic	0.0850 (0.0089)	0.2103 (0.0614)	0.2477 (0.0396)	<0.001

Note: Standard errors of the means are shown in parentheses.

Appendix 2: An alternative randomization test

To test the estimated rates, AF employ conventional χ^2 and difference of means tests. But although increasing the number of iterations allows for more conclusive inferences based on the estimates, it is not clear if these tests are applicable here in the first place as they assume the baseline values to be non-stochastic. Statistically speaking, there exists no formal basis for concatenating the random outcomes with the employed empirical tests. In this section, I propose an alternative rank order test for use in determining how likely it is that the observed differences in the rank orders are purely by chance.

The very nature of the resampling procedure lends itself to a preceding randomization construction.[29] In devising a null distribution from the data themselves, we can obtain an exact answer to the question of how likely the observed values would be if the null hypothesis were true. The null distribution is constructed by randomly rearranging the labels of the observations. If under the null hypothesis these labels do not matter, their permutation should not change the distribution of the original data. Such nonparametric randomization tests date back to Fisher (1935).[30] With the recent rise in computational power, they have become increasingly popular in applied statistics. The method has the advantage that it does not require specific assumptions about the underlying distributions. Moreover, it can be applied to make inferences about arbitrarily complicated test statistics, such as our resampled, aggregated, and finally averaged search-success rates.

The null hypothesis in AF's rank order test states that the search-success rates against a given race of motorists do not depend on the race of the troopers (AF 2006, 146). To implement this null hypothesis in the randomization test, I

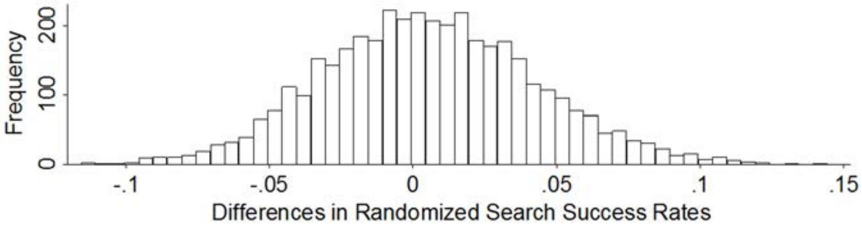
[29] I would like to thank Michael Wolf for pointing me in this direction.

[30] Romano (1990) provides a formal recap.

reshuffle the trooper identifier labels in the raw search data prior to the merger with the trooper subsamples.[31] Confining the reshufflings separately within troop and motorist race blocks picks up any potentially specific effects. This preceding randomization mirrors the idea that if the search-success rates do not depend on the race of the troopers searching them, reassigning the searches to troopers of different races should have no effect on the distribution of the search-success rates.[32]

Our observed values of the test statistic are the pairwise differences in search-success rates for a given race of motorists from Table 2. Under the null, these differences are zero. The corresponding null distributions are constructed by running a large number of independent passes, each of which is preceded by the randomization. After each pass, the differences in the search-success rates for a given race of motorists are recorded, providing the null distributions in which trooper race is exchangeable. In each null distribution, I calculate the exact *p*-value as the proportion of random values that are at least as extreme as the observed value. If trooper race does not matter, one should rarely find differences as large as the observed one. As an example, Figure 6 shows the frequency distribution in randomly obtained differences of the rates of Hispanic and white troopers against black motorists. It is easy to see that the observed value in Table 2, 0.1789 − 0.2025 = −0.0236, is not unusual when compared to this null distribution.

Figure 6. Null distribution of differences in search-success rates of Hispanic and white troopers against black motorists



[31] In contrast to my replication of AF's resampling procedure, the implementation of the randomization test required truly random seeds, which were obtained via the Stata package *setrngseed*.

[32] It would seem intuitive to randomly exchange the race labels in the trooper data set before the resampling procedure and the merger with the raw search data. However, this approach yields highly skewed null distributions because troopers exhibit heterogeneous search patterns (recall Figure 3). More precisely, reshuffling race in the trooper data set permutes *bundles* of observations in the raw search data, not single observations.

Panel A in Table 3 contains the estimated p-values for all differences in average search-success rates from the randomization test using 10,000 passes (with 1,000 iterations each). The p-values include their 99-CI.[33] For ease of comparison with AF's parametric test, the p-values from the replicated Z-tests based on Table 2 are shown in Panel B of Table 3. I follow AF's notation of search-success rates $S(r_m; r_p$, where r_m and $r_p \in \{W, B, H\}$ denote the motorist and trooper races, respectively. For a given race of motorists, the first column in Table 3 tests whether we can reject the null hypothesis of equal search-success rates for black and Hispanic troopers in favor of the one-sided alternative that black troopers exhibit a higher rate. The second column tests for inequality between Hispanic and white troopers. In addition, the third column tests for inequality between black and white troopers—the first and third rank.

TABLE 3. ***P*-values of differences in search-success rates**

	Search-success rate differences		
r_m	$S(r_m; B) - S(r_m; H)$	$S(r_m; H) - S(r_m; W)$	$S(r_m; B) - S(r_m; W)$
	Panel A: *P*-values from randomization test		
W	0.0069 ± 0.0021	0.4921 ± 0.0129	0.0013 ± 0.0009
B	0.2145 ± 0.0106	0.7901 ± 0.0105	0.3529 ± 0.0123
H	0.8390 ± 0.0095	0	0.0589 ± 0.0061
	Panel B: *P*-values from *Z*-test		
W	0.0023	0.3176	<0.0001
B	0.1919	0.7087	0.2607
H	0.6956	<0.0001	0.0217

By and large, the statistical inferences from the randomization tests are consistent with the ones from AF's empirical tests based on AF's generalized resampling procedure in Appendix 1. The p-values retain their levels of significance, with the exception of one rank. Using the randomization test, we cannot formally reject equality between the average search-success rates of black and white troopers against Hispanic motorists at a five percent level of significance: The p-value from the Z-test, 0.022, rises to 0.056. But for all intents and purposes, it still remains unlikely that this difference has been brought about purely by chance.

[33] Calculating all possible permutations would yield exact p-values but is computationally not feasible. Even so, a randomization test is asymptotically equivalent to such an exact test when the number of randomized passes is large enough. The precision of the estimated p-value, $\hat{p}$, increases with the number of passes. From the binomial distribution, the standard error of $\hat{p}$ is given by $SE_{\hat{p}} = [\hat{p}(1 - \hat{p})(1/n)]^{\frac{1}{2}}$ where n is the number of passes. As n increases, the distribution of $SE_{\hat{p}}$ approximates a normal distribution, from which the confidence intervals in Table 3 are devised. With the given data, 10,000 passes yield conclusive results in terms of statistical significance on a 99-CI.

Appendix 3: Data and code files

On the *Econ Journal Watch* website is **a guide to all the data and code files used in this research**.

References

Anwar, Shamena, and Hanming Fang. 2006. An Alternative Test of Racial Prejudice in Motor Vehicle Searches: Theory and Evidence. *American Economic Review* 96(1): 127-151. **Link**

Arrow, Kenneth J. 1973. The Theory of Discrimination. In *Discrimination in Labor Markets*, eds. Orley Ashenfelter and Albert Rees, 3-33. Princeton, N.J.: Princeton University Press.

Becker, Gary R. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.

Close, Billy R., and Patrick L. Mason. 2007. Searching for Efficient Enforcement: Officer Characteristics and Racially Biased Policing. *Review of Law and Economics* 3(2): 263-321.

Fisher, Ronald A. 1935. *The Design of Experiments*. London: Oliver & Boyd.

Ilić, Dragan. 2013. Spatial and Temporal Aggregation in Racial Profiling. *Swiss Journal of Economics and Statistics* 149(1): 27-56.

Knowles, John, Nicola Persico, and Petra Todd. 2001. Racial Bias in Motor Vehicle Searches: Theory and Evidence. *Journal of Political Economy* 109(1): 203-229.

Persico, Nicola, and Petra Todd. 2006. Generalising the Hit Rates Tests to Test for Racial Bias in Law Enforcement, with an Application to Vehicle Searches in Wichita. *Economic Journal* 116: F351-F367.

Phelps, Edmund S. 1972. The Statistical Theory of Racism and Sexism. *American Economic Review* 62: 659-661.

Romano, Joseph P. 1990. On the Behavior of Randomization Tests Without a Group Invariance Assumption. *Journal of the American Statistical Association* 85(411): 686-692.

Wu, Chien F. J. 1990. On the Asymptotic Properties of the Jackknife Histogram. *Annals of Statistics* 18(3): 1438-1452.

About the Author

Dragan Ilić is lecturer and senior researcher in the Economic Theory Group of the Faculty of Business and Economics at the University of Basel. His research interests are applied microeconomics and social economics. His email address is dragan.ilic@unibas.ch.

Go to archive of Comments section
Go to September 2014 issue

Discuss this article at Journaltalk:
http://journaltalk.net/articles/5853