COMMENTS

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

ECONOMICS IN PRACTICE

Saying Too Little, Too Late: Public Finance Textbooks and the Excess Burdens of Taxation
Cecil E. Bohanon, John B. Horowitz, and James E. McClure

CHARACTER ISSUES

Ragnar Frisch and the Postwar Norwegian Economy: A Critical Comment on Sæther and Eriksen
Olav Bjerkholt

A Reply to Olav Bjerkholt on the Postwar Norwegian Economy
Arild Sæther and Ib E. Eriksen

WATCHPAD

Does Occupational Licensing Deserve Our Approval? A Review of Work by Morris Kleiner
Uwe E. Reinhardt

Capitalism and the Rule of Love
Clarence Philbrook
Replicability and Pitfalls in the Interpretation of Resampled Data: A Correction and a Randomization Test for Anwar and Fang

Dragan Ilić

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...
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 $\chi^2$ 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.


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. 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.\textsuperscript{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 \textit{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 \textit{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

\textsuperscript{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. 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. 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. 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. 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).

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. 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 $\chi^2$ 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 `str10seed` 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.
ently 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.
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.\textsuperscript{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

\textbf{Figure 1.} Frequency distributions of replicated average search-success rates

\begin{table}
\centering
\begin{tabular}{|c|c|c|}
\hline
Motorists’ race & White & Black & Hispanic \\
\hline
White & & & \\
\hline
Black & & & \\
\hline
Hispanic & & & \\
\hline
\end{tabular}
\end{table}

\textsuperscript{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.\textsuperscript{16}

\textbf{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)

---

\textsuperscript{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. 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. 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.

Finally (not depicted), one of the three χ² 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 χ² 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 χ² 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 χ² 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.
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. 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

<table>
<thead>
<tr>
<th>Troop</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>120</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>88</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>155</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>D</td>
<td>176</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>E</td>
<td>68</td>
<td>39</td>
<td>58</td>
</tr>
<tr>
<td>F</td>
<td>125</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>G</td>
<td>105</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>H</td>
<td>62</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>K</td>
<td>81</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>L</td>
<td>91</td>
<td>45</td>
<td>15</td>
</tr>
<tr>
<td>Q</td>
<td>41</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Troop</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.125</td>
<td>0.429</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0.170</td>
<td>0.375</td>
<td>0.667</td>
</tr>
<tr>
<td>C</td>
<td>0.581</td>
<td>0.818</td>
<td>0.923</td>
</tr>
<tr>
<td>D</td>
<td>0.682</td>
<td>0.923</td>
<td>0.800</td>
</tr>
<tr>
<td>E</td>
<td>0.882</td>
<td>0.308</td>
<td>0.138</td>
</tr>
<tr>
<td>F</td>
<td>0.240</td>
<td>0.750</td>
<td>0.445</td>
</tr>
<tr>
<td>G</td>
<td>0.286</td>
<td>0.353</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>K</td>
<td>0.926</td>
<td>0.882</td>
<td>0.556</td>
</tr>
<tr>
<td>L</td>
<td>0.989</td>
<td>0.400</td>
<td>0.800</td>
</tr>
<tr>
<td>Q</td>
<td>0.366</td>
<td>0.750</td>
<td>0.400</td>
</tr>
</tbody>
</table>

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.\textsuperscript{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.\textsuperscript{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

\textsuperscript{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.

\textsuperscript{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 \( \frac{2}{3} \) 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.\textsuperscript{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

\textsuperscript{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. 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.
### Appendix 2: An alternative randomization test

To test the estimated rates, AF employ conventional $\chi^2$ 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. 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). 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.\textsuperscript{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>
<thead>
<tr>
<th>Search-success rate differences</th>
<th>( W )</th>
<th>( B )</th>
<th>( H )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S(r_m; B) - S(r_m; H) )</td>
<td>0.0069 ± 0.0021</td>
<td>0.2145 ± 0.0106</td>
<td>0.8390 ± 0.0095</td>
</tr>
<tr>
<td>( S(r_m; H) - S(r_m; W) )</td>
<td>0.4921 ± 0.0129</td>
<td>0.7901 ± 0.0105</td>
<td>0.0013 ± 0.0009</td>
</tr>
<tr>
<td>( S(r_m; B) - S(r_m; W) )</td>
<td>0.0013 ± 0.0009</td>
<td>0.3529 ± 0.0123</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

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 \( \text{SE}_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})(1/n)} \), where \( n \) is the number of passes. As \( n \) increases, the distribution of \( \text{SE}_{\hat{p}} \) 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


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.

Discuss this article at Journaltalk:
http://journaltalk.net/articles/5853
Saying Too Little, Too Late: Public Finance Textbooks and the Excess Burdens of Taxation

Cecil E. Bohanon¹, John B. Horowitz², and James E. McClure³

LINK TO ABSTRACT

Taxation imposes manifold costs beyond the amount of revenue raised. During recent decades economists have investigated the excess burdens of taxation including the costs of ‘deadweight’ distortions, enforcement, and compliance.⁴ Our synthesis of the estimates provided by these investigations indicate that it typically costs much more than a dollar to finance a dollar of government spending. We examine whether leading public finance textbooks discuss the various excess burdens or incorporate excess burdens when calculating the optimal level of public goods. We find that most do neither.

Table 1 gives the locations of the treatments of (1) public goods and (2) the welfare costs of taxation found in six public finance textbooks used in top economics programs in the United States.⁵ In each of the six books, the treatment of public goods precedes the treatment of the welfare costs of taxation (tax efficiency). In fact, after treating public goods, the treatment of tax efficiency comes, on average, 11 chapters later. By the time the author(s) gets to tax efficiency, the focus has long since shifted away from the optimal provision of public goods.

¹. Ball State University, Muncie, IN 47306.
². Ball State University, Muncie, IN 47306.
³. Ball State University, Muncie, IN 47306.
⁴. See Slemrod and Gillitzer (2014) for an in-depth discussion of administrative, compliance, and welfare costs of taxation.
⁵. The process we used to choose these textbooks is explained below.
TABLE 1. Topic separation: Public goods vs. the welfare costs of taxation

<table>
<thead>
<tr>
<th>Textbook</th>
<th>Chapters in book</th>
<th>Chapter treating public goods</th>
<th>Chapter treating the welfare costs of taxation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ronald C. Fisher (2006), State and Local Public Finance, 3rd ed.</td>
<td>22</td>
<td>Chapter 2</td>
<td>Chapter 12</td>
</tr>
<tr>
<td>Jonathan Gruber (2013), Public Finance and Public Policy, 4th ed.</td>
<td>25</td>
<td>Chapter 7</td>
<td>Chapter 20</td>
</tr>
<tr>
<td>David N. Hyman (2010), Public Finance, 10th ed.</td>
<td>18</td>
<td>Chapter 4</td>
<td>Chapter 11</td>
</tr>
<tr>
<td>Richard A. Musgrave and Peggy B. Musgrave (1989), Public Finance in Theory and Practice, 5th ed.</td>
<td>34</td>
<td>Chapter 4</td>
<td>Chapter 16</td>
</tr>
<tr>
<td>Harvey S. Rosen and Ted Gayer (2008), Public Finance, 9th ed.</td>
<td>22</td>
<td>Chapter 4</td>
<td>Chapter 15</td>
</tr>
<tr>
<td>Joseph E. Stiglitz (2000), Economics of the Public Sector, 3rd ed.</td>
<td>28</td>
<td>Chapter 6</td>
<td>Chapter 19</td>
</tr>
</tbody>
</table>

Note: Tax distortions are briefly mentioned by Hyman (2010) in Chapter 2 along with other sources of economic distortions.

The sequencing and wide separation of these discussions is a manifestation of the broader problem we focus upon: Textbooks say too little, too late, about the excess burdens of taxation. Even when the textbooks do get around to treating tax efficiency, the coverage of the costs of taxation is often inadequate. Such practice is likely to lead students to underestimate the costs of government programs, predisposing them toward increased government spending. If instead students were instructed on the manifold costs of taxation and these costs were integrated into discussions of public goods, students would probably be less predisposed toward government spending.

Naive public goods theory misleads on costs

Just like the modern textbooks, Adam Smith’s Wealth of Nations discusses government expenses first and then turns to revenue (Smith 1976/1776, V.1, V.2). However, unlike the modern textbook writers, when Smith discussed expenses he consistently integrated some discussion of their financing. Smith quite consistently preferred such financing to come principally from user fees, though Smith did consider national defense to be a pure public good that should be financed by

6. In some programs (particularly, it seems to us, public-administration programs) public finance is taught over two terms, with one term on spending and one on taxes—a practice that might interrelate with how authors organize the public-finance textbooks they write. Separating course material that way may make it especially hard for students to see how excess burdens affect the efficient quantity of a public good. Students are more likely to see the connections when the basics of cost-benefit analysis, public goods and excess burdens are covered during the beginning of both terms.
general taxation. Thus for Smith the relevance of tax efficiency to public goods should perhaps be limited, while its relevance should be high to modern textbook writers, who more often tend to favor general taxation as the financing mechanism when public goods are demanded.

In their analysis of public goods, textbooks normally depict supply as a marginal cost curve that does not include excess burdens. They normally assume lump-sum taxation with full information, though typically they do not make this explicit. This depiction may be reasonable when politicians are hammering out this year’s budget and deciding how to allocate money between programs. However, the depiction is not suitable in discussions of the optimal provision of a public good. Government must raise revenue to provide public goods, and so to assume a nondistortionary lump-sum tax will lead students to overestimate optimal provision.

When textbooks leave excess burdens unmentioned, they are de facto teaching the lump-sum tax perspective. The author of a textbook might deny assuming nondistortionary taxes, declaring ‘Just because I didn’t elaborate the manifold costs doesn’t mean I have denied such costs; rather, they are implicitly represented in the marginal-cost curve.’ We think that such a defense is inadequate. Normally governments must raise the revenue to provide the public good. Raising revenue generates enforcement costs, compliance costs, and deadweight distortions. When textbook authors don’t explicitly discuss these costs when discussing the optimal provision of a public good, it leads students to forget that taxes are distortionary and ignore these welfare costs when doing their analyses.

**Adam Smith on the excess burdens of taxation**

Smith’s four maxims of taxation underscore the excess burdens of taxation. Smith’s brief presentation of these maxims comes at the very beginning of his lengthy treatment of taxation: “Before I enter upon the examination of particular taxes, it is necessary to premise the four following maxims with regard to taxes in general” (1976/1776, 825).

The first maxim is that individuals’ tax payments should be “in proportion to the revenue which they respectively enjoy under the protection of the state” (ibid.). Smith associates the proportionality principle with “equality” in taxation. The most straightforward interpretation of Smith’s proportionality principle is as a proportional tax, that is, a tax where the tax rate does not vary with changes in the tax base. Although Smith does not speak of excess burdens in connection with proportionality, such connection follows to the extent that proportional taxes reduce administrative, compliance, and enforcement costs.
The second maxim is that tax obligations “ought to be certain, and not arbitrary. The time of payment, the manner of payment, the quantity to be paid, ought all to be clear and plain to the contributor, and to every other person” (ibid.). Without certainty, Smith says, the tax-gatherer “can either aggravate the tax upon any obnoxious contributor, or extort, by the terror of such aggravation, some present or perquisite to himself. The uncertainty of taxation encourages the insolence and favours the corruption of an order of men who are naturally unpopular” (ibid., 825-826). Smith argues that certain and non-arbitrary tax payments reduce the excess burdens of taxation.

The third maxim is that “Every tax ought to be levied at the time, or in the manner, in which it is most likely to be convenient for the contributor to pay it” (ibid., 826). Here Smith clearly highlights excess burden. He says that taxes “upon the rent of land or of houses” or “upon such consumable goods as are articles of luxuries” are conveniently paid.

The fourth maxim is more elaborate and broken down into four sub-points. It is entirely and explicitly about excess burden, including the psychic costs arising from “trouble, vexation, and oppression.” We quote the paragraph in full:

Every tax ought to be so contrived as both to take out and to keep out of the pockets of the people as little as possible, over and above what it brings into the public treasury of the state. A tax may either take out or keep out of the pockets of the people a great deal more than it brings into the public treasury, in the four following ways. First, the levying of it may require a great number of officers, whose salaries may eat up the greater part of the produce of the tax, and whose perquisites may impose another additional tax upon the people. Secondly, it may obstruct the industry of the people, and discourage them from applying to certain branches of business which might give maintenance and employment to great multitudes. While it obliges the people to pay, it may thus diminish, or perhaps destroy, some of the funds which might enable them more easily to do so. Thirdly, by the forfeitures and other penalties which those unfortunate individuals incur who attempt unsuccessfully to evade the tax, it may frequently ruin them, and thereby put an end to the benefit which the community might have received from the employment of their capitals. An injudicious tax offers a great temptation to smuggling. But the penalties of smuggling must rise in proportion to the temptation. The law, contrary to all the ordinary principles of justice, first creates the temptation, and then punishes those who yield to it; and it commonly enhances the punishment too in proportion to the very circumstance
which ought certainly to alleviate it, the temptation to commit the crime. Fourthly, by subjecting the people to the frequent visits and the odious examination of the tax-gatherers, it may expose them to much unnecessary trouble, vexation, and oppression; and though vexation is not, strictly speaking, expence, it is certainly equivalent to the expence at which every man would be willing to redeem himself from it. It is in some one or other of these four different ways that taxes are frequently so much more burdensome to the people than they are beneficial to the sovereign. (Smith 1976/1776, 826-827)

No contemporary public finance textbook that we are aware of even comes close to discussing excess burdens as comprehensively as Smith did in 1776.

### Some analysis and estimates of the excess burden

One would like to think that economists can neatly distinguish the components of the excess burden of taxation, provide a precise estimate of the magnitude of each component, and then add up the component estimates to arrive at an estimate of the total excess burden. Unfortunately, for a number of reasons it is not that simple. Consider some of the complicating factors. First, cost depends on how the relevant choice is contextualized. Second, taxation takes many forms. Third, it is very difficult to arrive at monetary values for the subjective costs from fear, anxiety, anger, and frustration from what Smith called “unnecessary trouble, vexation, and oppression.” Fourth, there is no definitive way to divide the components; for example, should enforcement be separated from compliance? Fifth, some potential components, such as ones having to do with tax evasion, tax avoidance, or black markets, might mitigate other components, such as suppressed work or opportunity. Sixth, empirical estimation is necessarily very crude and inexact. Seventh, the costs vary over time; for example, perhaps technology is making it easier for people to comply with tax law.

One component of excess burden is compliance cost, the costs of conforming to often complex and changing tax laws. Joel Slemrod and Jon Bakija estimate that “individual taxpayers spend as much as 3 billion hours of their own

---

7. For a breakdown of components and estimates, see James Payne (1993, 150, 247-248). Incidentally, Payne insists that all of his estimates are conservative, lower-bound estimates (ibid., 9).

8. The Laffer curve illustrates why context matters. If people believe that tax rates are high, then they are also likely to believe that cutting tax rates will raise revenue.
time on tax matters, or about 27 hours per taxpayer on average. That is the equivalent of over 1.5 million full-time (but hidden and unpaid) IRS employees!” (2008, 3-4, emphasis in original). The IRS (2012, Table 2.1) reports that tax preparation fees reported as itemized deductions were about $6.9 billion. Slemrod and Bakija (2008, 162) report that their “best estimate of the total annual cost of enforcing and complying with the federal corporate and personal income taxes in tax year 2004 is $135 billion. This amounts to slightly more than 10 cents per dollar raised.” In other words, for each dollar collected via income taxes, the inclusion of compliance costs alone would bring the total burden to $1.10.9

Tax wedges are another important cost of taxation. In labor markets, the average taxpayer works less when she faces higher marginal tax rates. Although average tax rates are easily used to calculate one’s tax bill from gross income, the effects of taxes on one’s decisions to work and save are determined by the overall marginal tax rate (MTR) from federal, state, and local taxes.10 Edward Prescott (2004) reported that in 1970 labor supplies were nearly equal in the United States and Europe. Also in 1970, MTRs were similar in the United States and Europe. By the mid-1990s, MTRs in Europe increased to about 60 percent, compared to 40 percent in the United States—and Europeans were working about a third less than Americans. Prescott (2004, 8) finds that much of the difference in labor supply is explained by the differing MTRs.

Some analysts discuss disincentives to save as another cost of taxation; higher MTRs reduce the incentive to save. Taxes on dividends, capital gains, interest income, and corporate and business profits reduce savers’ rates of return. Although there is little agreement on how much these taxes affect savings, Jonathan Gruber (2013) notes that more recent studies suggest that consumption decisions are strongly affected by after-tax interest rates. Edgar Browning (2008) argues that one reason Americans save less than many other countries is the relatively high American MTRs on capital income. Progressive taxes also place the largest tax burden on higher income people who tend to save the most. Browning (2008) says that total savings is especially reduced by progressive taxes that reduce the return to savings for high-income individuals, who tend to save the most.

9. Payne (1993) made a much higher estimate of compliance costs borne by households and businesses, about 24 cents per tax-revenue dollar. Payne separately estimated enforcement costs, meaning the “governmental cost of tax collection,” but he found they “prove to be relatively small.” He added: “Virtually all of the costs of operating the U.S. tax system are shifted onto the private sector” (Payne 1993, 9, 29, see also 119-126).

10. Browning (2008, 154) reports that the average MTR in the United States is about 40–45%. Gruber (2009) notes that a reasonable estimate for the labor supply elasticity of primary workers is 0.1 whereas the labor supply elasticity of secondary workers is much higher at 0.5–1.0, with most of the response coming from changes in labor force participation. If the labor supply elasticity were 0.5 and if MTRs increase by 10 percentage points, then the increase in tax rates would decrease labor supply by 5 percent.
Distortions arising from reductions in the tax base via exemptions and deductions are another matter sometimes treated as a cost of taxation. According to the Internal Revenue Service (2013, Table 5), adjusted gross income since 1970 is 15% to 25% less than personal income. Taxpayers have the incentive to move their income into areas that are not taxed. This distorts taxpayers’ choices.

The welfare costs of taxation cannot be measured with great precision. Bev Dahlby (2008) presents various estimates of the marginal costs of public funds (MCF) for various taxes. But MCF is only one of several frameworks that have been used in measuring the welfare costs of taxation; others include marginal excess burden (MEB), marginal efficiency costs (MEC), and marginal welfare costs (MWC). This makes it difficult to summarize and compare the results from the various studies.

Table 2 lists the results from nine publications that assess the costs of raising tax revenue from labor income in the United States. Over the last thirty years, individual income and payroll taxes have been the source of at least eighty percent of all federal tax revenue (OMB 2014, 34-35). Column 2 shows estimates of the cost of raising a dollar in income tax revenue when compliance costs are excluded; estimates range from $1.07 to $1.52, with one outlying high estimate of $3.00. As discussed, Slemrod and Bakija (2008) estimate that there is a $0.10 compliance cost when raising a dollar in tax revenue; column 3 thus includes compliance costs by adding ten cents to the estimates in column 2.

<table>
<thead>
<tr>
<th>Source for estimate</th>
<th>Estimate (excludes compliance costs) (2)</th>
<th>Estimate plus compliance costs of $0.10 (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmed and Croushore (1996)</td>
<td>$1.12 to $1.17</td>
<td>$1.22 to $1.27</td>
</tr>
<tr>
<td>Ballard, Shoven, and Whalley (1985)</td>
<td>$1.16 to $1.31</td>
<td>$1.26 to $1.41</td>
</tr>
<tr>
<td>Browning (1987)</td>
<td>$1.32 to $1.47</td>
<td>$1.42 to $1.57</td>
</tr>
<tr>
<td>Feldstein (1999)</td>
<td>$3.00</td>
<td>$3.10</td>
</tr>
<tr>
<td>Fullerton and Henderson (1989)</td>
<td>$1.17 to $1.25</td>
<td>$1.27 to $1.35</td>
</tr>
<tr>
<td>Gruber and Saez (2002)</td>
<td>$1.29</td>
<td>$1.39</td>
</tr>
<tr>
<td>Jorgenson and Yun (1991)</td>
<td>$1.38 to $1.52</td>
<td>$1.48 to $1.62</td>
</tr>
<tr>
<td>Jorgenson and Yun (2001)</td>
<td>$1.35 to $1.40</td>
<td>$1.45 to $1.50</td>
</tr>
<tr>
<td>Stuart (1984)</td>
<td>$1.07</td>
<td>$1.17</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td><strong>$1.29 to $1.33</strong></td>
<td><strong>$1.39 to $1.41</strong></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>$1.43 to $1.50</strong></td>
<td><strong>$1.53 to $1.60</strong></td>
</tr>
</tbody>
</table>

Note: The elasticity of taxable income with respect to the after tax share was 1.04 in Feldstein (1999). In comparison, Gruber and Saez’s (2002) elasticity was 0.4. See Slemrod and Gillitzer (2014, 90-91) for an insightful discussion of the reasons for these differences.
Considering all nine publications, the median estimate of the cost to raise a dollar in revenue through a tax on labor income (column 2) is about $1.30; including ten cents of compliance costs, the figure is about $1.40 (column 3). A $1.40 cost to raise a dollar in income tax revenue implies that textbook writers who assume that it costs a dollar to raise a dollar are ignoring about 30 percent of the costs.

These estimates do not seem to be controversial. In 2005, the President’s Council of Economic Advisers reported: “A recent study estimated that the excess burden associated with increasing the individual income tax by one dollar is 30 to 50 cents. In other words, the total burden of collecting $1.00 in additional income taxes is between $1.30 and $1.50, not counting compliance costs” (Bush CEA 2005, 77). In a *Journal of Economic Literature* review, Slemrod (2005, 817) said this estimate cited by the CEA “is a reasonable characterization of where the literature stands.”

The costs of raising funds through a general sales tax are similar. Charles Ballard, John Shoven, and John Whalley’s (1985) estimate of the MEB for consumer sales taxes puts the cost of raising a dollar of public funds between $1.25 and $1.39. Dale Jorgenson and Kun-Young Yun (1991) estimate the MEC of a sales tax on consumer and investment goods to be about $1.26; in later work, Jorgenson and Yun (2001) estimated the MEC to be about $1.18. All of these estimates omit compliance and enforcement costs, so the full cost of raising a dollar of public funds would be even higher.

However, there are cases where the marginal cost of raising a dollar in public funds can be much lower. Examples include Pigovian taxes and raising revenue by reducing distortionary tax subsidies. Ballard and Steven Medema (1993) calculated that a Pigovian tax that reduced a negative externality had a marginal cost to raise a dollar of $0.73. Don Fullerton and Yolanda Henderson (1989) calculated that raising a dollar by reducing an investment tax credit that distorted asset choices had a marginal cost of $0.62. Ballard and Fullerton (1992) develop a model in which the marginal cost of raising a dollar in public funds can be a dollar or less. Excess burdens may be negative in these cases, but because compliance and administrative costs are ignored in these studies the full costs would be higher than the authors’ estimates. Policymakers may be able to reduce marginal excess burdens by choosing those types taxes that have lower excess burdens. All else equal, replacing high MEB taxes with low MEB taxes is a good idea. This is one reason that Henry George argued for taxes on land (see Foldvarg 2005). But one might be suspect of a reform proposal to replace more distortionary taxes with less distortionary taxes, from concern that by the end of the legislative process the more distortionary taxes will be left in place with the less distortionary taxes merely added on top of them.
Integrating the full cost into public good analysis

We noted that public finance textbooks have segregated public goods theory from tax issues. Textbooks present the efficient quantity of public goods in the opening chapters and taxation is presented toward the end of the text. The derivation of the efficient quantity of public goods follows the standard formulation (Samuelson 1954; 1955), where the efficient quantity occurs where the marginal social benefit (MSB) equals the marginal resource cost (MC). The MSB is the vertical summation of the individual MB schedules; the vertical summation is because of the non-rival aspect of a public good.

For simplicity, in Figure 1, the marginal cost curve is horizontal, implying a constant marginal resource cost to produce \( Q \). Here \( Q \) is defined as a composite public good that can be thought of as encompassing defense, schooling, et cetera. In Figure 1, \( Q \) can increase either because the size of government programs increase or there is an increase in the number of government programs. Students are told that resources are optimally allocated at \( Q_1 \) where MSB and MC intersect.

**Figure 1.** How welfare costs affect the optimal level of public goods

---

11. In the early chapters of the textbooks, welfare costs and compliance costs are ignored; these chapters emphasize market failure in the provision of public goods caused primarily by free-rider issues which imply that markets underprovide public goods.
When the funds are not raised voluntarily, however, compliance and other excess burdens are especially important. When excess burdens are included in the analytics, the relevant cost curve is $MSC$ which is the sum of $MC$ and the marginal welfare costs of taxation. A familiar proposition in public finance is that welfare costs of taxation increase by the square of the tax rate (Musgrave and Musgrave 1989, 281-285). On such proposition, a doubling of revenue would quadruple the total welfare cost of taxation. As public good production expands, then, the marginal welfare cost of taxation should be thought to increase, and to increase at an increasing rate. In Figure 1, then, $MSC$ not only lies above $MC$ but also increases at an increasing rate. The efficient output is $Q^*$, which is less than $Q_1$.

Figure 2 shows the case where $Q$ is one public good, such as ‘basic science’ (let us grant that it is a pure public good). The marginal cost of funds (not shown in the figure) is determined by the total level of government spending. Assuming that basic science funding is a small part of overall spending, the marginal cost of funds is not substantially affected by increasing expenditures on basic science. In this case, $MSC$ is a horizontal line above $MC$, but still $Q^* < Q_1$. Then, if the size of other programs or the number of government programs were to be increased, the marginal cost of funds would rise, which would shift the $MSC$ curve upward in Figure 2, reducing the optimal $Q^*$.

**Figure 2.** How welfare costs affect the optimal level of a public good

As there are compliance and welfare costs associated with raising the revenue necessary to finance a public good, the optimal-provision analyses presented in public finance textbooks are incomplete if not misleading. Without being shown a curve such as $MSC$ in our Figures 1 and 2, the student will have no hint that excess burdens affect the choice of $Q$, unless perhaps the surrounding text articulates the
point. But textbooks generally fail to do even that; they generally don’t discuss the welfare costs of taxation until at least six chapters later.

**An examination of texts and supplements used at top schools**

We examine the treatment of public goods, the costs of taxation, and related cost-benefit analysis in six textbooks and eight supplemental readings used in public finance courses at five top schools. It is likely that most public finance courses in the United States use one of these textbooks. We record whether each text explicitly notes that the cost of raising an additional dollar of tax revenue typically costs more than a dollar. Special attention was paid to whether cost-benefit analysis of public goods includes the welfare costs of taxation. Table 3 summarizes our findings.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher (2006)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Gruber (2013)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Hyman (2010)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Musgrave and Musgrave (1989)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Rosen and Gayer (2008)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Stiglitz (2000)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

12. The textbooks and supplemental readings included were determined as follows. First, we accessed the 2009 *U.S. News & World Report* rankings of the top ten graduate programs in economics. Next we conducted online searches at each school’s website to find any syllabi for public finance courses that included a course textbook. This yielded five syllabi, from five of the schools (Harvard, Massachusetts Institute of Technology, Stanford, Northwestern, and University of Pennsylvania). There are six textbooks and ten supplemental readings on the list; we examine all six textbooks and the eight supplemental readings that we were able to access. All syllabi except for Columbia and Princeton were from undergraduate courses. The syllabi from Columbia and Princeton were from applied master’s programs. See the Appendix for the complete list of schools and readings.
Though all texts included separate-chapter discussions of cost-benefit analysis, public goods, and the welfare costs of taxation, Table 3 clearly shows that public finance textbooks commonly segregate cost-benefit analysis from discussion of excess burdens. No text incorporated excess burdens into cost-benefit analyses or the optimal provision of public goods. However, at the end of chapter 16, Richard Musgrave and Peggy Musgrave (1989) mention that the welfare cost of taxation does increase the cost of public spending. They note that:

…the deadweight loss of the marginal tax dollar…must be known to determine the proper size of the budget, because it sets the costs (tax dollar plus deadweight loss) which need be measured against the benefits derived from the marginal expenditure dollar. (Musgrave and Musgrave 1989, 293)

Also, Jonathan Gruber (2013, 600-601) discusses tax efficiency before his discussion of optimal taxation. Gruber cites Jerry Hausman’s (2000) study that estimated that a wireless communications tax generates a marginal welfare cost of 72 to 90 cents per dollar raised. However, unlike the Musgrave and Musgrave (1989) discussion, there is nothing in Gruber (2013) that links the welfare cost of taxation to public spending.

Taken together Tables 1 and 3 make the case that public finance textbooks commonly segregate public good theory and tax theory. Again, an author might respond to our critique by saying something like: ‘What do you mean? Following tradition, in the earlier chapters I discuss public goods. Then in later chapters I explain that a dollar of revenue typically costs about $1.40 [or whatever]. Readers should be able to apply the concepts that they learn to earlier chapters.’ Such an explanation, however, is never presented in their texts; expenditure theory is never explicitly integrated into tax theory in these public finance textbooks.

An author could also say: ‘My book looks at minimizing the welfare costs of taxation given some target level of public funds.’ This misses the point that the target level of public funds depends crucially on the excess burdens of the tax. Figure 1 shows that it is analytically easy to incorporate excess burden into public goods theory.

In Table 4, we consider materials specified in the course syllabi as supplemental. These supplements focus on specific issues in public finance ranging

13. Slemrod and Gillitzer (2014, esp. ch. 7), who integrate compliance and welfare costs into their tax analysis, also assume that the target level of public funds is exogenous.
14. The framework of Figure 1 follows the framework outlined in Brennan, Bohanon, and Carter (1984) which proposes a public finance pedagogy along the lines suggested in this article.
from tax issues in developing economies to the institutional history of the American tax system.

**TABLE 4. Supplemental materials, welfare costs of taxation, and cost-benefit analysis**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bird and Oldman, eds. (1990), <em>Taxation in Developing Countries</em>, 4th ed.</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Case (1986), <em>Economics and Tax Policy</em></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Cordes, Ebel, and Gravelle, eds. (2005), <em>The Encyclopedia of Taxation and Tax Policy</em></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Goode (1984), <em>Government Finance in Developing Countries</em></td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Lewis (1984), <em>Taxation for Development</em></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Prest (1985), <em>Public Finance in Developing Nations</em>, 3rd ed.</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Slemrod and Bakija (2008), <em>Taxing Ourselves: A Citizen’s Guide to the Debate Over Taxes</em>, 4th ed.</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Steuerle (2004), <em>Contemporary U.S. Tax Policy</em></td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Notes: Case (1986) discusses the provision of public goods (pp. 40, 120-121, 140) and how taxes cause welfare costs (pp. 122, 142-143). In the encyclopedia edited by Cordes et al. (2005), Watson (2005, 121-122) discusses welfare costs. Goode (1984) discusses cost-benefit analysis (pp. 62-74) and the consequences of heavy taxation (p. 95). Lewis (1984, 15-16) discusses public goods and the welfare costs of taxation. Prest (1985, 35), Slemrod and Bakija (2008, 3-4, 144-146, 160-163), and Steuerle (2004, 12-13) discuss welfare costs and collection costs of taxation.

As shown in column 6, only Wayne Thirsk (1990, 192-199) in the text edited by Richard Bird and Oliver Oldman (1990) explicitly takes account of welfare costs in his cost-benefit analysis.15 Two of the supplements offer passing verbal comments that suggest that welfare costs of taxation are intertwined with spending policy. Karl Case (1986, 122), in his book designed for an international law course, writes: “When a tax distorts or alters economic decisions…the burden imposed

15. This can be seen immediately where Thirsk lists a parameter for excess burden in his Table 15.1 (1990, 194).
on society is greater than the amount of tax collected.” The book by C. Eugene Steuerle (2004) is designed to review the evolution of federal tax policy since WWII; it goes into little depth on tax or public good theory. Nevertheless, Steuerle acknowledges that: “Taxes by their very nature distort behavior… Even when distortions are minimized for some level of tax collection, those taxes, because of their remaining effect on behavior must be justified by the gains from the programs they support” (2004, 12).

**Are intermediate texts in microeconomics different?**

We have focused on public finance textbooks because we assume that compliance costs and welfare costs are more likely to be integrated into discussions of the optimal provision of public goods in these area-specific texts than in more general economics texts. To check the reasonableness of this assumption, we chose as our sample four undergraduate intermediate micro texts authored by renowned economists: Edgar Browning and Mark Zupan (2009); Austan Goolsbee, Steven Levitt, and Chad Syverson (2013); Hal Varian (2010); and Steven Landsburg (2011). For each text we examined the discussion of the optimal provision of a public good. The selection of the four textbooks was done casually, and it is meant only to illustrate a problem the extent of which, among the very many intermediate micro texts, is uncertain.

All four of these texts ignore the welfare cost of taxation when discussing the optimal quantity of a public good. Browning and Zupan (2009, 576), Goolsbee, Levitt, and Syverson (2013, 673) and Varian (2010, 703) all illustrate the optimal provision of a public good with a diagram similar to our Figure 1 but without an MSC curve. Landsburg (2011) does not graphically show the optimal provision of a public good, but in his discussion of public-goods provision he too ignores the welfare costs of taxation:

> Because nonexcludable and nonrivalrous goods are supplied inadequately by the marketplace, they are often provided by the government. If it would cost $300 to build a streetlight that 100 neighbors value at $10 apiece, we have seen that the market can fail to provide the streetlight. A government, however, can assess a tax of $3 per neighbor and use the proceeds to build the light, yielding a clear gain in social welfare. (Landsburg 2011, 468)
The failure of these general texts to include welfare costs in their discussions of the optimal provision of a public good is consistent with the failure of specialty public finance texts to integrate taxation’s welfare costs into public-good provision analyses.

**Results and conclusion**

The coercive extraction of tax dollars is a costly activity. Estimates of the total cost of raising a dollar from the U.S. personal income tax are substantial in magnitude. So one would think that public finance textbook discussions of cost-benefit analysis and the optimal provision of public goods would explicitly incorporate the welfare costs and compliance costs of taxation. But in our sample of the public finance textbooks used at elite institutions, none integrates the welfare costs of taxation into discussions of the optimal provision of public goods or cost-benefit analysis. When these costs are ignored, the purportedly optimal quantity of public goods is in fact too high. Fortunately, this bias could—quite easily—be eliminated from public finance textbooks.

**Appendix**

Top ten graduate economics programs, 2009 ranking by *U.S. News & World Report*

1. Harvard University
2. Massachusetts Institute of Technology
3. Princeton University
4. University of Chicago
5. Stanford University
6. University of California, Berkeley
7. Yale University
8. Northwestern University
9. University of Pennsylvania
10. Columbia University

Harvard University required texts
- Bird and Oldman (1990)
- Case (1986)
- Cordes, Ebel, and Gravelle (2005)
- Fisher (2006)
- Goode (1984)
• Gruber (2007)
• Gupta, Clements, and Inchauste (2004)
• Lewis (1984)
• Musgrave and Musgrave (1989)
• Newbery and Stern (1987)
• Prest (1985)
• Rosen and Gayer (2007)
• Shome (1995)
• Steuerle (2004)
• Stiglitz (2000)

Massachusetts Institute of Technology required text
• Gruber (2009)

Princeton University
From syllabus “There are no textbooks …”

University of Chicago
No syllabus found

Stanford University required text
• Rosen and Gayer (2008)

University of California, Berkeley
No syllabus found

Yale University
No syllabus found

Northwestern University required texts
• Gruber (2009)
• Slemrod and Bakija (2008)

University of Pennsylvania required text
• Hyman (2010)

Columbia University
No text information readily available from online syllabus

References


Goolsbee, Austan D., Steven Levitt, and Chad Syverson. 2013. Microeconomics.
   New York: Worth.
Internal Revenue Service (IRS). 2013. SOI Tax Stats—Historical Data Tables. Internal Revenue Service (Washington, D.C.). Link


About the Authors

Cecil Bohanon is a Professor of Economics at Ball State University. He obtained his B.A. from Wilmington College (Ohio) and his Ph.D. from Virginia Tech. He has published over 30 refereed professional articles, notes, and comments, and over 100 popular articles, policy monographs and newspaper editorials. His research interests include public choice, applied microeconomics, and economic education. His email address is cbohanon@bsu.edu.

John Horowitz (Ph.D., Texas A&M University, 1988) is Professor and Chair of the Department of Economics at Ball State University. His research focuses on microeconomics, public policy, and health economics. His publications have appeared in Economic Inquiry, Review of Economics and Statistics, Health Economics, Public Choice, and others. His email address is jhorowitz@bsu.edu.

James E. McClure (Ph.D., Purdue University, 1983) is Professor of Economics at Ball State University, where his research focus is microeconomics. His publications appear in journals such as Economic Inquiry, Econ Journal Watch, Journal of Economic Behavior and Organization, and Southern Economic Journal. His email address is jmcclure@bsu.edu.

Discuss this article at Journaltalk: http://journaltalk.net/articles/5854
Ragnar Frisch and the Postwar Norwegian Economy: A Critical Comment on Sæther and Eriksen

Olav Bjerkholt

Arild Sæther and Ib Eriksen (2014), professors from the University of Agder in Norway, argue with great zeal that Ragnar Frisch, who was professor at the University of Oslo from 1931 to 1965, exerted a strong and dogmatic influence on Norwegian economists as an ideologue of central economic planning, with important and detrimental impact on the Norwegian economy in the post-WWII period. Upon my initial reading of their argument I found them far-fetched and unreasonable, and I communicated with the editor of this journal about the matter. I am happy to accept the editor’s invitation to scrutinize their facts and give my assessment. I will not keep from the readers that I have been closer to Frisch’s world than have Sæther and Eriksen. I began to study economics at the University of Oslo at the time Frisch retired and spent many years in a team of researchers in Statistics Norway working inter alia on developing model tools for use by the Norwegian government in the conduct of macroeconomic policy.

Sæther and Eriksen assert that Ragnar Frisch in the late 1930s propounded an economic-political system for Norway with “no place for private investors or entrepreneurs in the system” and in which “[e]conomists should make the important investment decisions” (2014, 54). They claim that, due to Frisch’s influence, the postwar Norwegian economy “became close to the centrally planned economies of Eastern Europe” (2014, abs.), with detrimental consequences for economic development. The authors provide only weak evidence for these strong assertions, which run counter to much historical research and general knowledge.

1. University of Oslo, 0317 Oslo, Norway.
The legacy of Ragnar Frisch

Paul A. Samuelson wrote after Frisch’s death in 1973 as follows:

Ragnar Frisch dominated analytical economics from the early 1930’s founding of the Econometric Society to his wartime internment in a Nazi concentration camp. He combined fertility and versatility with depth. Although this was the golden decade in which the world came to realize his genius, both in the years before, and the years after, Frisch made numerous important contributions to (a) economic theory, (b) economic measurement, (c) economic policy, and (d) scientific methods in statistics, mathematics, and economics. (Samuelson 1974, 7)

As an epitaph of excellence in economics this statement is surpassed by few. Samuelson’s article was a tour de force packed with dense mathematical exposition. Samuelson made the pertinent observation that the word “model” in its present economic sense was actually introduced by Frisch. It can be traced to the following passage of a lecture given by Frisch at Yale University in 1930:

The observational world itself, taken as a whole in its infinite complexity and with its infinite mass of detail, is impossible to grasp. Taken in its entirety, in its immediate form of sense impressions, it resembles, so to speak, a jelly-like mass on which the mind cannot get a grip. In order to create points where the mind can get a grip, we make an intellectual trick: In our mind we create a little model world of our own, a model world which is not too complicated to be overlooked, and which is equipped with points where the mind can get a grip, so that we can find our way through without getting confused. And then we analyze this little model world instead of the real world. This mental trick is the thing which constitutes the rational method, that is, theory. (Frisch 2010/1930, 31-32)

The rational, instrumental use of models is indeed characteristic of Frisch’s approach to economic analysis: no dogmas, just stringent reasoning. Frisch’s “depth” as mentioned by Samuelson was not least the ability to resolve a problem from a penetrating insight into its inner mathematical structure. He also hinted at Frisch’s deep econometric insight on the confrontation between theory and data, the role of structure, simultaneity, and autonomy (Samuelson 1974, 22). Frisch’s position as an internationally renowned economist need not be further elaborated.

2. Sæther and Eriksen (2014) provide some information about Frisch’s life and career but could have been more helpful for foreign readers, as only one of nine cited papers by Frisch is in English.
Like many of his contemporaries Frisch was deeply affected by the depression of the 1930s, on which he commented that “poverty is imposed on us in the midst of a world of plenty” (Frisch 1934, 259). The depression imprinted on him an almost passionate opinion that an economist had a duty to work for the betterment of society. As a teacher, for more than forty years he infused his students with an instrumental view of economics and equipped them with skills and tools to exercise it. Naturally he took a great interest in economic policy in his own country, both in the 1930s and in the postwar period, and it influenced his research. Frisch exercised his right and duty as a citizen to take active part in public debate both on economic and other issues.

Frisch’s willingness to marshal all resources and when necessary to work day and night was legendary among his assistants. It was an attitude like that expressed by Henrik Ibsen in the national epos *Peer Gynt*:

> To set one’s own goal is uplifting, I feel,  
> and to carry it through, hard as flint or steel!  
> (with quiet emotion)  
> Break all connections and ties, all that tends  
> to bind one with bonds to home and to friends, —  
> blow one’s treasure sky-high to heaven above, —  
> bid a goodnight to the pleasures of love, —  
> all to uncover the truth’s hidden mystery —  
> (wipes the tear from his eye)  
> there you’ve the true researcher in history! —  
> I feel there’s no limit now to my pleasure.  
> Now I have taken my destiny’s measure.  
> Now, simply hold out, thick and thin, that’s my stint!  
> (Ibsen 2007/1867, 78)

A number of statements in Sæther and Eriksen (2014) are incorrect and some of them severely so. They write as follows:

> During the second half of the 1930s, Frisch and his disciples became increasingly skeptical of the use of market forces to obtain an efficient allocation of resources and distribution of goods. … The solution was a state macroeconomic planning system and state governance with detailed regulations and selective policies for all branches of industries. The extent of such controls would depend on the economic situation.

---

3. On Frisch in this regard, see Thonstad (2005) and Bjerve (1998).
There was no place for private investors or entrepreneurs in the system. Economists should make the important investment decisions. (Sæther and Eriksen 2014, 53-54)

That Frisch’s vision of economics in the late 1930s had “no place for private investors or entrepreneurs” is blatant untruth. I am not aware of any document by Frisch—published or unpublished—giving any kind of support to these curious assertions. Politically and economically the ideas expressed are totally remote from Frisch’s thinking. Also the implied assertion that Frisch’s students and assistants would go along with such nonsense is absurd.

This kind of programmatic politics wrapped up as part of the teaching of economics is entirely out of character for Frisch, who imprinted on every cohort of his students since 1937 the distinction between a scientific statement and a value judgment. The distinction is far from trivial, as descriptions of economic problems often hide tacit value assumptions:

In tackling scientific problems…the aim is to find solutions that can be subjected to the criterion of correct or false. In the case of practical rules of conduct, on the other hand, the question at issue is one of desirability and expediency—or belief in expediency—and this is based on a number of value judgment elements which must be considered in relation to the practical or political object in view, and on which it would naturally be impossible to expect unanimous agreement. (Frisch 1965, 5)

This does not mean that the economist, like the scientist, should fail to preoccupy himself with problems involving value judgments. … It merely means that he should always endeavour to draw the line of demarcation as clearly as possible. He should bear in mind the extent to which the results of his analysis are dependent on the non-scientific presuppositions from which his analysis proceeds. To the extent that he succeeds in making a clear and definite distinction between what is a scientific statement in the proper sense and what is a value judgment, he will be able to preserve his objectivity. And this is, in fact, the only way in which it can be preserved. It was the German economist and sociologist Max Weber who first clearly pointed this out. (Frisch 1965, 7-8)

In discussions Frisch was often a boisterous participant who tended to dominate the discussion. In the heat of the battle he could be opinionated and even forget his own prescriptive advice of Weberian objectivity.
The postwar Norwegian economy

The Labor Party government which came to power at the end of 1945 had to cope with economic problems quite similar to those facing other war-devastated countries in Western Europe. The problems included reconstruction, inflationary pressure, and the balance of payments. The political stability and the postwar mood of broad consensus regarding the reconstruction provided the foundation for the Norwegian government in adopting goals for its economic policy more ambitious than those of most other Western governments.

The government put in place a system of preparation of plans and programs for economic policy, often referred to as “national budgeting” after the key policy document: the annual white paper called National Budget setting out the policy, the assumptions on which it was based, and the expected results. These plans were not directives but rather expectations and intentions. The fulfillment of plans and programs depended upon the external environment and the use of economic control measures. In the early years the monitoring of the economy took place at a detailed level and so did the use of control measures.

During the early post-war period, when excess demand existed in almost all fields, a very comprehensive system of direct controls was used to prevent prices from rising, to prevent imports from exceeding exports by more than a given amount, and to allocate resources and redistribute incomes consistent with particular goals. Direct quantity controls were used primarily in a negative manner, i.e., by prohibitions and permits. As excess demand subsided, direct quantity controls became correspondingly less essential for the prevention of price increases and less effective for the allocation of resources. However, ambitious allocative and distributive goals continued to be maintained, and instruments other than direct controls had to be more intensively applied for these purposes, e.g., indirect taxes and subsidies for redistributive purposes plus various forms of credit control for allocative purposes.

The postwar economic policy, planning, and control are well documented in official documents, studies by participants such as Petter Jakob Bjerve (1959), and postwar history accounts. Such studies are often more oriented towards political history, with little emphasis on the intricacies of economic policy. The sources relied on in Sæther and Eriksen (2014) seem somewhat inadequate.

The European postwar development was given much attention in the United States. Norway was given more attention than other countries. The high planning ambition combined with political democracy caught the attention of the Columbia University sociologist Paul Lazarsfeld, who wanted to conduct a large-scale study.
of planning in Norway as an observed social and economic experiment. It comprised a number of projects, and while the overall study never got off ground, a student of Lazarsfeld completed some of the projects (Barton 1957).

Lawrence Klein visited Oslo in 1947 and wrote a remarkable and concise account of Norwegian planning in the modern economist’s language, giving background, structure, numbers, and an appraisal (Klein 1948). There were no less than three major studies of the Norwegian postwar economy made at Harvard University by Walter Galenson (1949), Alice Bourneuf (1958) and Mark W. Leiserson (1959). Galenson (1949) and Leiserson (1959) paid particular attention to the labor relations in the postwar economy. Bourneuf (1958) had a broader coverage and is the most cited of these authors. Below I quote some passages from these studies to elucidate the character of the postwar economic policy regime.

Bourneuf (1958) pinpointed the key strategic choice at the start of the postwar years:

When World War II finally ended, Norway, like most other European countries, was faced with the difficult job of rebuilding an economy both devastated and dislocated by long years of war and occupation. …

Extreme dependence on imports combined with wartime destruction of much of its foreign-exchange earning capacity made Norway’s situation exceptionally difficult. The investment needed to restore export earnings would be a serious drain on foreign-exchange earnings and would compete with imports of food and raw materials which were sorely needed. Productive capacity had to be expanded, especially foreign-exchange earning capacity, to support a growing population at prewar living standards. This would require even more investment. Norway had to choose between rapid reconstruction and development and rapid restoration of the prewar standard of living. (Bourneuf 1958, 198)

Galenson spent time in Norway in 1946–47 and observed at close range the postwar economy in the period of the strictest planning measures. He made the following assessment:

It is one thing to plan the orderly disposition of resources, but quite another to accomplish the plan without impinging upon democratic

---

4. It is noteworthy that these studies took place in the wake of the Taft-Hartley Act, which was adopted over Truman’s veto and pretty much destroyed Truman’s efforts at providing the consensus needed for building a welfare state in the United States.
freedoms. The Soviet experience has made it clear that scientific tech-
niques of control render a comprehensive national plan feasible. But
Soviet plans are executed by fiat; the barrier to centralized allocation
of labor and capital tends to be physical rather than institutional.
Norwegian planners must face the fact that the great bulk of national
output is produced by private entrepreneurs seeking their individual
profit, and that the labor force possesses not only the theoretical right
to strike against unpalatable planning decisions, but also the organi-
zational and political capability of translating that right into reality.

The system of controls devised to channel individual initiative
into socially desirable activities is neither as complete nor as authori-
tative as a planner might desire, although it is drastic by prewar stan-
dards. It reflects the weakness as well as the strength of democracy.
Constructed by compromise between conflicting interest groups, it is
in some respects a patchwork. The process by which agreement was
achieved, however, was essential to the growth of mores that make
it possible for the system to operate with a minimum of compulsion.
(Galenson 1949, 261-262)

Leiserson, who covered the medium-to-long horizon of postwar planning,
paid particular attention to the instrumental character of the economic planning in
Norway:

It is important to emphasize the aggregative character of the Nor-
wegian economic plans, since it had a decisive influence on the means
employed in achieving planned objectives. … Such policies may be
distinguished in the first instance by their generality and secondarily
by their tendency to operate indirectly. … Within this framework of
macro-economic policies and controls, there was a whole range of
more detailed regulations and controls of varying degrees of specificity
which can be distinguished by their particular or differential character
… (Leiserson 1959, 18)

This more or less purely “instrumental” attitude toward economic
policy pervaded the whole structure of Norwegian postwar economic
planning, giving it a uniquely “enlightened,” “unorthodox,” or
“radical” character…. The resulting process of national planning and
control can be schematically described as an attempt to formulate
economic objectives (as quantitatively as possible) in real terms and
then to adapt and employ from the whole range of direct and indirect,
physical and financial controls available, those which a pragmatic and objective evaluation indicated to be the most effective in achieving the basic economic objectives.

Needless to say, the actual planning process only approximated any such austere logical ideal. Numerous other considerations did enter into policy formation as is inevitable in a democratic country with healthy political institutions. (Leiserson 1959, 20)

Alice Bourneuf made these concluding observations on how the planning process was carried through:

Norway’s experience provides a case study of comprehensive planning designed to achieve a rapid rate of economic growth in the face of serious limiting factors. Before the war, Norway had been less fully industrialized than some of the other Western European countries. Her greatest resource, hydroelectric power potential, had only begun to be exploited. The country was not underdeveloped in the sense of having a very low level of per capita income or of having political, institutional, or educational barriers to economic growth; but there were important possibilities for expanding productive capacity and raising the level of income. …

Economic planning and direct controls did not lead to interference with political and civil liberties. The character of the Norwegian people suggests that such interference would not be tolerated. Nor were recovery and development achieved at the expense of the lower income groups. The heavy investment program went hand in hand with expanded welfare and social security programs, and also with a considerable redistribution of income within the wage earning group. (Bourneuf 1958, 2)

While Galenson (1949) and Leiserson (1959) did not mention Ragnar Frisch at all in their comprehensive studies, Bourneuf (1958) mentioned him once in the following passage:

The fact that a large number of well-trained economists were available for the job was due largely to the work of Professor Ragnar Frisch of the economics department at the University of Oslo. Of the considerable number of economists who were trained at the University, only a few were absorbed by the University; most of the others, whether Conservative, Liberals, or Labor Party members were drawn into government service. In many countries, the scarcity of highly
trained economists and administrators is a serious bottleneck in attempts at over-all economic planning. The success of any economic plan is bound to depend on the individuals available to carry it out. (Bourneuf 1958, 206)

Patrick Salmon (1990) discussed the roots of the postwar consensus that emerged in the Nordic countries, not least due to the early postwar experience. He reports the growth rate of GDP 1946–56 as 5.1% for Norway against 3.2% and 3.9% for Denmark and Sweden, respectively (ibid., 156). Salmon noted that “one of the most distinctive features of Norwegian postwar development [was] a very high level of capital formation…. The reasons had to do, on the one hand, with the government’s commitment to capital and energy-intensive industries which could utilize Norway’s hydro-electric reserves, while at the same time contributing to the economic development of the more northerly regions (as well as, debatably, demonstrating the superiority of state over private enterprise); and on the other hand, with the social priority of improving living standards, in this case housing” (1990, 162-163).

Another work of interest is Francis Sejersted’s comprehensive and impressive study of Norway and Sweden in the twentieth century (Sejersted 2011). The study pays attention to the rise of the social democratic welfare state—also referred to as the Scandinavian or Nordic model—and its historical roots, including the social and economic experience of the early postwar period.

In European countries, measures of comprehensive planning in the early postwar period were the rule rather than the exception. In some countries new institutions were erected, such as Commissariat général du Plan in France, established in 1946 to be in charge of economic planning, particularly through Five Year Plans, and Centraal Planbureau in the Netherlands, founded in 1945 by Jan Tinbergen. No such institution was erected in Norway. Such difference may be interpreted as illuminating the instrumental and temporary character of the economic planning in Norway, in line with the studies discussed above, as opposed to an emphasis on a planning ideology favoring a regulatory regime, as implied by Sæther and Eriksen (2014). The absence of large-scale nationalization measures in

5. The “large number of well-trained economists” mentioned by Bourneuf calls for a brief comment. Through 1945, the number of graduates from the master-level program opened in the mid-1930s was only 64. There were however well-equipped economists of earlier vintages with shorter education.

6. Salmon (1990) noted that investment in Norway comprised, by the accounting rules, also repair and maintenance of capital goods, and the definition of the investment rate thus differed somewhat from that of other countries.

the Norwegian postwar policy, such as took place in some other countries, again illuminates the non-ideological, pragmatic character of the Norwegian experience.

We have shown by drawing on foreign studies that the Norwegian postwar planning was something quite different from what is asserted by Sæther and Eriksen (2014). They apparently believe that the market economy was abolished in Norway, even for decades after the war. I have cited the evidence for the wrongness of such an assertion.

Sæther and Eriksen do not distinguish between “planning” and “controls” (or “regulations”). All modern governments do a lot of planning and programming; how could they otherwise manage their economies and fulfill their international obligations? The European Recovery Program (better known as the Marshall Plan) required by agreement plans to be prepared and commitments to execute them from the participating countries. The authors also have difficulties with aspects of how policy is conducted, e.g., the role of models, but that will not be pursued here.

Just as Sæther and Eriksen have opinions different from mine about Norway’s cultural insularity and the meaning of “plan” versus “regulation,” I note also other differences. Although the authors do use the term “market economy” they seem to prefer “free market economy,” or even “free enterprise economy.” I do not know exactly what they mean by the latter terms—perhaps laissez-faire? Obviously they are much opposed both to the plans prepared and to the controls put into effect in the early postwar years, such as the import licensing and the consumer rationing. But how do they think about circumstances at that time? What were the policy options in 1946? I find no clue to this, except vague suggestion that a free enterprise economy would have managed better. Would it?

The National Budget, the key policy document, was presented to the Norwegian Parliament and the public for the first time in 1946. Since then, a national budget has been presented every year for the same purpose, namely to present the short-term program for the government and its view on major policy issues. The content naturally has changed over the years. Sæther and Eriksen (2014) provide no explanation of what the national budget is, as they ought to have done as this term is not generally used in other countries. What they say about Frisch and “national budgets” before and during the war (ibid., 53, 48) is meaningless, as the national budget was not yet invented. An interesting and relevant fact in this connection, but also unmentioned, is that the econometrician and future Nobel laureate Trygve Haavelmo was in charge of the National Budget 1948. He returned

8. See the Foreign Assistance Act of 1948 (P.L. 80-472 (link)), particularly Sec. 115(b). A recent Norwegian white paper recapitulated this requirement: “A condition for receiving American support for reconstruction was that each country presented an economic four-year plan as a basis for assessing the need for support” (St.meld. nr. 9, 2008–2009 (link), p. 128, my translation).
from the USA after eight years in 1947 and served as the chief coordinator of the
economic policy for one year. Lawrence Klein, who visited, reviewed the national
budgets for 1947 and 1948 and noted the improvements in national budgeting
achieved by Haavelmo resulting in a much slimmer and more transparent
publication in 1948 and with some elements of welfare economics added to it
(Klein 1948, 812). Bourneuf (1958) also commented upon this improvement.

The studies cited above remarked on the healthy and resilient political
democracy in Norway. Sæther and Eriksen have the guts to assert that when, after
twenty years, the Labor Party was voted out of power in 1965 and replaced with
a majority coalition government “the power of the Oslo School economists in
the ministries was so strong that there was no change in the main features of the
economic planning” (2014, 62). The suggestion that a government which came into
power after 20 years in opposition, not least about economic policy, refrained from
action due to the “power of…economists in the ministries” is neither believable
nor true. There is an air of unreality over this and some of the other of the authors’
statements.

Sæther and Eriksen have difficulty explaining when the regime they have
conjured up ended. They write in one place about “the central economic planning
system from the mid-1940s to the mid-1970s” (2014, 73) and in another place
that “at the end of the 1970s the inefficiency of the planned economy impelled a
change” (abs.). They seem to suggest that the Norwegian economy had come to
the brink of collapse, which is what centrally planned economies do when they
are abolished, but there wasn’t any collapse, and rationing and the strict control
measures had ended 20 years earlier. The austerity policy of the early years
continued in a relatively smooth way in the build-up of the Norwegian welfare
state. Internationally the 1970s were marked by oil price hikes and stagflation. It
was a difficult period in many countries, including Norway but not due to “central
economic planning.” The difficulties included the readjustment of the industrial
structure, as pointed to by the authors. They let it pass unmentioned, though, that
the GDP growth rate through the 1970s was higher in Norway than in any other
OECD country, while the inflation rate was below the average.

Also notable is the blame put on economists. Sæther and Eriksen assert, with
reference to Espen Soilen (2002), that in the 1970s “economists at the Ministry of
Finance were unable to level self-criticism” (Sæther and Eriksen 2014, 70), rubbing
the point in by adding that “the top management of the Ministry…for years had
indulged a system in defiance of economic reality” (ibid.). This is unusual criticism
in a Norwegian context and hard to take seriously. Although the authors state, as
mentioned above, that the central economic planning lasted until the 1970s, they
seem to revise this towards the end of the paper as they state that first “during
the 1990s the elaborate system of detailed economic planning and control came
to an end” (ibid., 73). In the very last sentence they say further that “Norway has overcome the worst excesses of the tide that rose and partially receded in the twentieth century, but there still remains much scope for improvement” (ibid., 74). Is “improvement” perhaps suggesting eradication of irresponsible economists?

The demonization of Frisch

The most problematic—and somewhat unpleasant—aspect of Sæther and Eriksen (2014) is what for lack of a better word can be called the attempted demonization of Ragnar Frisch by means of false assertions, tainted terms, made-up ‘facts’, and odd imputation of intentions. The authors foreshadow their infamous treatment by noting that “as often is the case for famous people, there is also another story to be told” (p. 48), like a suggestion of hidden truth.

Frisch was, needless to say, very well known, highly respected, and admired by most Norwegian economists. This is not inconsistent with the fact that many—perhaps most—of them disagreed with Frisch on many issues, particularly on economic policy. Except during the war years, discussions among economists were indeed frequent, conducted openly in newspapers and other media, especially about the government’s economic policy. Sæther and Eriksen (2014) make a point of denoting some economists, whether Frisch’s university colleagues, assistants, or others, as Frisch’s “disciples” and “supporters” as if he was a religious sect leader or a political agitator. It is unseemly and uncalled for.

But it is worse than that. Sæther and Eriksen make up a story that Frisch in the 1930s built up the “Oslo School of economic research and teaching” (2014, 53). The term itself is innocent enough. But what is the purpose? The authors use this label to speak about, for example, “leading economists of the Oslo School” (ibid., 72) or “the fundamental principles of the Oslo School” (ibid.). According to the authors Frisch gathered his “disciples” in the “Oslo School” and indoctrinated them in the 1930s to believe in “state governance with detailed regulations and selective policies for all branches of industries” with “no place for private investors or entrepreneurs” and where “economists should make the important investment decisions” (ibid., 53-54). It is false through and through, although we cannot rule out that this is what these authors believe. They also depict Frisch as a scheming Machiavellian in the 1930s: “A crude theory would be that a yearning to see himself in with a governing set led Frisch to bend his thinking to make himself viable with

---

9. The expression “Oslo School” could be heard used occasionally in the 1960s and 1970s, e.g., by Copenhagen economists who generally had great admiration for the economics that emerged from Frisch’s Institute and used the expression as a counterpart to the (defunct) “Stockholm School.”
the one power faction that seemed open to him” (ibid., 51). Again a derogatory insult without a shred of evidence.

In the world of Sæther and Eriksen, Frisch was in complete control of his “disciples” and he directed them to put in place all the evil deeds of the postwar economic policy, such as when they assert that Minister of Finance Erik Brofoss, the strong man in the postwar cabinet, merely “implemented Frisch’s economic thinking in government” (ibid., 59). The foreign researchers cited above did not have Sæther and Eriksen’s imaginative power and apparently did not fathom that Frisch was behind it all! The demonization ends with “Frisch’s grand vision being dissolved,” the “overcoming of Frisch’s influence and legacy,” and it becoming “brighter” (ibid., 71, abs.). This comes amusingly close to evoking Frisch as the Prince of Darkness! This kind of denigration and calumniation of Frisch is quite deplorable, and it is astonishing that it has found its way into the columns of a respectable journal.10

This odd treatment of Ragnar Frisch evokes Henrik Ibsen’s statement in Peer Gynt that “no man’s a prophet in his native land” (Ibsen 2007/1867, 68).11 Sæther and Eriksen may have seen themselves in the role of the somewhat dubious character of the “Button molder” who bestows the final verdict on Peer Gynt:

You were ordained as a button that shone
on the coat of the world—but your shank has gone;
and so for you it’s the reject dump,
there to be rendered (we say) in the lump.
(Ibsen 2007/1867, 107)

**Additional comments**

Sæther and Eriksen (2014) is fraught with numerous infelicities; just for the record a few of them are commented below.

Leif Johansen was the only student of Frisch and Haavelmo who measured up to his teachers’ excellence. Those who knew Johansen will hardly agree to Sæther and Eriksen’s characterization of him as “aggressive” (2014, 56). Johansen’s doctoral dissertation of 1960 is characterized by the authors as “the foundation for

---

10. Another grudge from Sæther and Eriksen is that due to Frisch’s influence the entrepreneur “remains absent from economics education in Norway,” and it is worse than that as this is “representative of more general problems in much of mainstream professional economics” (2014, 74)!

11. Ibsen had borrowed it from the New Testament: “But Jesus said unto them, A prophet is not without honour, save in his own country” (Matthew 13:57).
long-term economic planning by the Ministry of Finance” (ibid.). It is correct that it was adapted to fit the Ministry of Finance’s model needs in the late 1960s. In fact it is still used, in newer versions, by the Ministry for Finance for the same or similar purposes. But the authors suppress that Johansen’s dissertation originated the computable general equilibrium model tradition, a versatile tool used in hundreds of applications by the World Bank alone (Bjerkholt 2009). Some of the early practitioners in this line of modeling, who used Johansen’s dissertation as a manual, are convinced that Johansen would have been awarded the Nobel Prize for this work if he hadn’t died prematurely in 1982 at the age of 52.

For no obvious reason the authors provide the following information: “Clifford Hildreth (1917–1995) at the University of Minnesota was advisor for Arild Sæther from 1966 to 1968. He was at the Cowles Commission at the same time as Haavelmo. He claimed that Haavelmo had a tremendous influence on the research environment. At lunch and coffee breaks Haavelmo distributed new ideas and research proposals freely to his colleagues” (Sæther and Eriksen 2014, 56 n. 19). But Haavelmo had left Chicago two years before Hildreth arrived.12

Haavelmo was not visiting professor in Aarhus in 1938 (he was a lecturer), and he was not a research fellow at Harvard (he studied at Harvard while a Rockefeller fellow), nor was he a contemporary at the Cowles Commission with Gerard Debreu and Herbert Simon. Sæther and Eriksen describe Haavelmo’s famous 1944 dissertation as showing “that the results of many of the methods used to that time had been misleading,” but Haavelmo’s 1944 dissertation did not discuss statistical methods in use or the work of any practitioners. The authors assert that Haavelmo “did not believe in neoclassical equilibrium theory,” but Haavelmo did not consider theories as objects of belief (cf. Sæther and Eriksen 2014, 55-56).

Ragnar Frisch graduated from the University of Kristiania, also known as the Royal Frederick University. Frisch did not work in 1925 as assistant on a “research program in production theory led by professor Petter Thorvald Aarum”, as no such research program existed at that time, and neither did professor Aarum work on production theory in 1925–26. Frisch’s doctoral dissertation from 1926 was not on time series. Frisch’s chair was not created by the university by funding from the Parliament (an impossible procedure) but granted in the fiscal budget after approval by the Parliament (cf. Sæther and Eriksen 2014, 47).

12. See the Cowles Commission reports for 1947 (link) and 1948–49 (link).
Conclusion

Sæther and Eriksen (2014) tell their story of the role of economists in an important period in Norwegian economic history. One of their assertions, although not about economists, is that Norway is (was?) a “relatively culturally insular country” (p. 46). In fact, after the independence of the State of Norway in 1814, authors, playwrights, and poets were joined by painters, sculptors, composers, philosophers, mathematicians, explorers, and scientists in an active exchange and interaction with other nations. The assertion adds to the impression noted above that Sæther and Eriksen’s perception of reality is sometimes peculiar. They have difficulty in setting out a structured argument and in distinguishing between proper evidence and casual and irrelevant statements. A further weakness of the article is a slight element of prejudice and ideological bias, not sobered by adhering to facts. They have obviously worked very hard to complete the paper.

References

Olav Bjerkholt is professor emeritus in economics, University of Oslo, Norway. He worked in the Research Department of Statistics Norway for close to 30 years on input-output models, electricity studies, oil and gas analyses, disarmament studies, et al. His teaching has focused on natural resource economics and the history of economics. He has edited several volumes of work by Ragnar Frisch. He has been a member of the UN’s Committee for Development Policy for 12 years. His email address is olav.bjerkholt@econ.uio.no.

About the Author


Discuss this article at Journaltalk:
http://journaltalk.net/articles/5855

Go to archive of Character Issues section
Go to September 2014 issue
A Reply to Olav Bjerkholt on the Postwar Norwegian Economy

Arild Sæther¹ and Ib E. Eriksen²

LINK TO ABSTRACT

In spite of Professor Olav Bjerkholt’s belittling language we are grateful for his spirited and highly critical commentary on our paper titled “Ragnar Frisch and the Postwar Norwegian Economy” (Sæther and Eriksen 2014). We very much hope that Professor Bjerkholt’s critique finds a large audience. Although it makes a few criticisms that are legitimate, its main tendency is to bolster our interpretation of postwar Norway.

Bjerkholt (2014) quotes the work of numerous scholars of the 1950s about Norway. It seems reasonable to suppose that these scholars were themselves admirers of economic planning under democratic socialism or social democracy. They give high marks to Norway for prosecuting planning and controls in exemplary fashion. If the reader reads the quotations from our viewpoint, it is easy to see that most of them support our interpretation. That these writers celebrate Norway’s supposed ability to prosecute its planning “without impinging upon democratic freedoms” (Galenson 1949, 261; quoted in Bjerkholt 2014, 302-303) and without “interference with political and civil liberties” (Bourneuf 1958, 2; quoted in Bjerkholt, p. 304) does not undermine our interpretation. Economic planning and controls restricted individual liberty, what Adam Smith called “natural liberty” (1767/1776, 687). We never contended that economic planning upset the country’s institutions of democratic governance.

After quoting the studies from the 1950s, Bjerkholt concludes: “The studies cited above remarked on the healthy and resilient political democracy in Norway” (2014, 307). It seems to us that Professor Bjerkholt does not understand the

¹. Agder Academy of Sciences and Letters, 4604 Kristiansand, Norway.
². University of Agder, 4604 Kristiansand, Norway.
distinction between democratic institutions and “free market economy” or “free enterprise economy.” Bjerkholt notes that we use the latter terms and remarks: “I do not know exactly what [Sæther and Eriksen] mean by the latter terms” (p. 306). Such confusion over terminology shows how deeply one’s ideological views are intertwined with one’s analysis. In his abstract, Bjerkholt accuses us of “ideological and moralistic bias,” as though he, by contrast, represents “Weberian objectivity,” which he then praises in the paper (p. 300). Isn’t it obvious that it is better to understand the disagreement as two contending interpretations of postwar Norway, neither with any scientific privilege?

Regarding the central importance of Ragnar Frisch, Bjerkholt provides a wonderful quotation from Alice Bourneuf (1958), a quotation we wish we had provided. It bears full reproduction here:

The fact that a large number of well-trained economists were available for the job was due largely to the work of Professor Ragnar Frisch of the economics department at the University of Oslo. Of the considerable number of economists who were trained at the University, only a few were absorbed by the University; most of the others, whether Conservative, Liberals, or Labor Party members were drawn into government service. In many countries, the scarcity of highly trained economists and administrators is a serious bottleneck in attempts at over-all economic planning. The success of any economic plan is bound to depend on the individuals available to carry it out. (Bourneuf 1958, 206)

Bjerkholt (2014, 300) says it is “blatant untruth” that Frisch put forward a vision of a planned economy with “no place for private investors or entrepreneurs in the system,” as we had written (Sæther and Eriksen 2014, 54). Bjerkholt adds, “I am not aware of any document by Frisch—published or unpublished—giving any kind of support to these curious assertions. Politically and economically the ideas expressed are totally remote from Frisch’s thinking.” In our reading of Frisch’s postwar writings we have found no references to the entrepreneur. As for the supposed remoteness to Frisch of the idea of an economic system where private investors play no role, the interested reader should contemplate the economic planning scenarios that Frisch imagines in his “Rational Price Fixing in a Socialistic Society” (Frisch 1966, 120-121).

It is interesting how Bjerkholt (2014, 305) draws on Patrick Salmon (1990). Firstly, Bjerkholt refrained from sharing what Salmon says about the objectives and instruments of the economic policies in Norway in the 1950s and 1960s:
Labour Party politicians and economists looked first and foremost to Britain as a model (Pharo 1984) but subjected the economy to controls over prices, consumption, and production which were ‘more stringent than in other democratic countries’ with the explicit aim of ‘transforming society into a socialist order’, as prime minister Gerhardsen put it (Hodne 1983: 143; Bourneuf 1958). (Salmon 1990, 162)

Second, Bjerkholt does show that Salmon reports that the Norwegian growth rate in the 1946–56 period was much higher than those of Denmark and Sweden, and that Salmon noted that Norway had a very high level of investment (1990, 156, 162). But Bjerkholt does not mention Salmon’s reference to Finland, which was harder hit by the war but achieved a higher growth rate of 5.5 percent. There has been considerable discussion about the reason for different growth rates in the initial postwar years. Odd Aukrust (1965, 64, our translation) notes that “The countries that were hardest hit by the war, have almost without exception had the strongest growth after 1946.” The Norwegian growth rate dropped considerably after 1950.

Finally, Bjerkholt omits part of Salmon’s explanation for the high level of investment:

But the government retained a large measure of control over the scale and direction of investments, thus contributing to one of the most distinctive features of Norwegian postwar development: a very high level of capital formation… (Salmon 1990, 162; cf. Bjerkholt 2014, 305)

The matter of Norway’s capital formation, or investment ratio, brings us to the most significant aspect of Bjerkholt’s critique. In our paper we make a major point that the record of Norwegian growth rates in the postwar years has to be viewed in light of Norway’s exceptionally high investment ratios, because that means that Norwegian consumers were in fact consuming less to provide for heightened growth rates. Professor Bjerkholt does not respond to the point at all.

Bjerkholt (2014, 306) questions what Norway’s policy options were in 1946, as though the challenges left nowhere else to turn. But most OEEC countries opted for a different policy, less controls, than Norway, and many of them performed better with lower investment ratios (Salmon 1990, 156). Bjerkholt (p. 307) then claims that we “suggest that the Norwegian economy had come to the brink of collapse” during the 1970s, but this statement has no foundation in our paper.
We conclude by saying that we are grateful to Professor Bjerkholt for engaging our interpretation of postwar Norway. We hope that readers learn from the clash of perspectives.

References


About the Authors

Arild Sæther is professor emeritus in economics, Agder Academy of Sciences and Letters, Kristiansand, Norway. He has held administrative and academic positions at the University of Agder in Kristiansand since 1970. He has for two and a half years been professor of European economic integration at the European Institute of Public Administration, Maastricht, the Netherlands; for three years, he has been Director of EuroFaculty Tartu-Riga-Vilnius. His teaching and research areas include micro and macroeconomics, welfare economics, European economic integration, international economics, regional economics, economic history and the history of economic thought. His email address is arild@saether.no.

Ib E. Eriksen is professor in economics at the University of Agder, Kristiansand, Norway. He has also been a research scholar and assistant professor at the Institute for Shipping Research at the Norwegian School of Economics, Bergen. He has been a management consultant. His teaching background is in international economics, shipping economics and finance. His research fields are in shipping economics and financial economics. In recent years he has worked on projects related to the history of economic thought. His email address is ib.e.eriksen@uia.no.

Go to archive of Character Issues section
Go to September 2014 issue

Discuss this article at Journaltalk:
http://journaltalk.net/articles/5856
Does Occupational Licensing Deserve Our Approval? A Review of Work by Morris Kleiner

Uwe E. Reinhardt¹

1. Princeton University, Princeton, NJ 08544.

Textbooks for first-year courses in economics, which for the most part are copies of one another and have progressed in content only imperceptibly over the past several decades, trot out the benchmark model of perfectly competitive markets without warning students sufficiently on how stringent the conditions are for markets to qualify for the label “perfectly competitive.” Most important among these conditions are:

1. Both buyers and sellers must be technically and intellectually competent to judge the merits and shortcomings of the goods and services being offered for exchange, which means that they must fully understand the various quality attributes of these goods and services and what benefits potential buyers personally might derive from them.

2. Aside from being technically and intellectually competent to judge the merits of what is being offered to them, buyers and sellers must be fully and accurately informed on the quality dimensions of what is being offered for trade.

3. Before making a decision to trade, the potential buyers must know the full price they have to pay per unit of the good and service and sellers must know the full price they will receive.
4. Buyers should bear the full price charged by the producer of the good or service being sold to the buyers, and producers should bear the full cost of producing them.

5. No seller or buyer should have the slightest degree of monopoly power in the market for traded goods or services. There should be many buyers and sellers trading in the market.

6. Neither the buy-side nor the sell-side of the market should be able collude to fix prices or other terms of trade.

7. There must be completely free entry for buyers and sellers to interact in the marketplace.

8. Exit from the market should be easy and relatively low-cost for both buyers and sellers.

Many exchanges of goods and services do not take place in markets that meet these criteria; indeed, such perfectly competitive markets are rare. In many instances, entry into a market or exit from it, or both, are difficult, which tends to bestow monopolistic power on incumbent sellers. Sellers in some markets enjoy full natural monopolies or artificial monopolies—e.g., producers with patent protection for their products, or unions which, as sellers of labor, represent pure monopolies, albeit ones with diffuse objectives. Finally, and most importantly, many markets are characterized by an asymmetry of information, be it on the quality dimensions of the good or service being traded or on the full prices ultimately charged for them. Usually buyers possess less accurate information than do sellers. Such asymmetries of information could be present even in markets for relatively simple commodities, such as foodstuffs, in which buyers may not know the ingredients of the food or whether or not the food (e.g., fish or vegetables) has been exposed to and may incorporate dangerous substances. Asymmetry of information is particularly prevalent also in markets for technically complex products, such as electronic products and, notably, health care.

It is well known that in the presence of asymmetry of information, the side with relatively less accurate information (usually the buyer) can be exploited by the side with relatively more information (usually the seller). That problem is amplified in the presence of third-party payment, as under private or public health insurance, which can diminish the buyer’s incentive to exert themselves in efforts to acquire accurate information on the quality of the product being traded.

Public policy could mitigate market imperfections in a number of ways. Antitrust policy, for example, might be used to force competition on markets that would otherwise be monopolized on one side or the other. The inefficiency and welfare loss begotten by asymmetric information can be mitigated by basically two methods.
First, public policy, abetted by modern information technology, can encourage or legislate the provision of more accurate information on the product being traded to all sides. Forcing the revealing of ingredients in foodstuffs or medicines, for example, is one such measure. Government can also help finance, or provide directly, websites with accurate information on the products offered in the market, especially in health care. The government-sponsored website “Hospital Compare” (link) is an illustration of such an attempt. Economists generally favor this approach for reducing asymmetry in information.

A second, more difficult method for overcoming the problem of asymmetric information is direct regulation of the behavior of participants in the afflicted markets. It can be done by prescribing in detail how participants in the markets must behave. Prescribing minimum nurse-to-bed ratios, for example, is an attempt to safeguard the quality of health care delivered by hospitals. Requiring non-physician health workers to be (a) licensed and (b) supervised directly by a physician is another.

Licensing occupations is a very forceful intervention in markets. Such regulations forbid anyone without a license to perform the tasks permitted in the licensed occupation. The prospect of losing that license then provides a powerful incentive to follow prescribed behaviors, e.g., to provide customers with accurate information on the quality of what is being offered for sale.

Unfortunately, like any powerful medicine, such regulatory measures can have a number of untoward side effects. They limit managerial discretion and thus hinder the efficient management of enterprises. They can also stifle innovation. Also, by raising the cost of entering the occupation, occupational licensing shifts up the supply curve of the services provided by the licensed occupation, with the result that employment in the occupation falls and the price of its services rises.

Occupational licensing usually is defended with appeal to this asymmetric-information rationale. One should think, therefore, that the demand for occupational licensing would typically originate with buyers of the goods or services provided by the occupation to be licensed, or by the legislative champions of these buyers.

In fact, however, more often than not occupational licensing has originated from those practicing the occupations to be licensed. To be sure, the legislative initiatives typically are also advanced with appeal to consumer protection, perhaps even sincerely so. But given the likely impacts of occupational licensing—reduction in the number of competitors in the occupation and increases in the prices of the occupation’s services—economists naturally suspect the motives for occupational licensing when it is requested by the occupation to be licensed. In some regards, occupational licensing functions as a substitute for other attempts to monopolize the market, e.g., the monopolization of labor markets through unionization.
Because occupational licensing is pervasive and growing in modern economies—certainly in the United States (Kleiner and Krueger 2013)—one would expect textbooks in introductory economics to cover it thoroughly. Remarkably, it is not so. The widely popular Principles of Economics by Greg Mankiw (2012), for example, does not cover the topic at all. Other textbooks may cover it in passing, in a paragraph or two (e.g., Hall and Lieberman 2013). And as Frank Stephenson and Erin Wendt (2009) have shown in this journal, even more advanced texts in labor economics tend to give scant attention to occupational licensing. Very recently, however, at least two labor texts in new editions have expanded their coverage of the topic (Stephenson 2013).

Highly welcome, therefore, has been the extensive and intensive work on this subject by Morris Kleiner, as evidenced once again in his latest book titled Stages of Occupational Regulation: Analysis of Case Studies (2013). That volume is a veritable theoretical and econometric tour de force through this lightly trodden terrain. In the introduction, Kleiner presents it as a follow-on to his earlier Licensing Occupations: Ensuring Quality or Restricting Competition?, published in 2006. Furthermore, it draws on a long list of his scholarly papers on occupational licensing, spanning more than two decades. Former U.S. Secretary of the Treasury and once again now Harvard professor of economics Lawrence Summers calls Kleiner “our foremost expert” on this important topic (Kleiner 2013, front matter).

In Stages of Occupational Regulation, Kleiner explores the progression of occupational regulation, from mere registration to certification to outright licensing—three distinct stages. Of these, the least restrictive is registration, under which individuals practicing an occupation merely register their name, address, and qualifications with a public agency but otherwise are free to practice their art. More restrictive is certification through what is called titular acts. Under certification individuals are free to practice an occupation as they see fit, but they can use a specific title—e.g., M.D. or R.N.—only if they have been certified to possess a certain set of skills, usually requiring examination. Finally, the most restrictive regulation, occupational licensing, imposed through what is called practice acts, allows only licensed individuals to practice the occupation.

As Kleiner and Alan Krueger (2013) have noted in their joint work, since World War II occupational regulation of some form has grown enormous, as unionization gradually declined, so that by 2008 almost 40 percent of the U.S. labor market were subject either to certification or licensing by some level of government. To examine how occupational regulation progresses from relatively mild registration to stringent occupational licensing, Kleiner (2013) carefully selects for his analysis a series of occupations representing the stages of regulation, devoting a chapter to each occupation. After an illuminating introduction and overview, the analysis begins with interior designers, who are only lightly regulated
and in only a few states, to mortgage brokers, providers of preschool child-care services, electricians, plumbers, and, at the more stringent end, dental hygienists and dentists, who now are licensed in every state.

For each of the several occupations analyzed, Kleiner describes first the work of the occupation and the history of its regulation. Next he explores the impact of that regulation on the relevant labor market, focusing on three main facets of the market: (1) levels of employment in the occupation, (2) wages earned in the occupation, and (3) outcome for the occupation’s customers—quality of services, errors, customer satisfaction—as best as can be defined and measured for each particular occupation. For electricians and plumbers, he includes in “outcomes” also the hazard to the practitioners themselves. Finally, Kleiner also offers analyses on economic factors in a state that might encourage occupational regulation.

To explore the impact of occupational regulation on the three facets of the relevant labor market, Kleiner uses a variety of statistical approaches to tease out, from numerous databases, what the impact of mild to heavy regulation on labor markets appears to be. Given the plurality of economic and other factors that can, in theory, drive such impacts, the limits of the available non-experimental databases and the statistical methods at hand naturally do not allow the authors to capture and properly control for all of those potential drivers. As do all econometric studies, Kleiner’s end up as essays in persuasion.

That said, the empirical findings in the volume overall do conform more or less with the theorized impact of occupational regulation on labor markets, especially as occupational regulation gets heavier, as the author reports in his summary chapter. Specifically, he shows that the relatively mild occupational regulation of interior designers, mortgage brokers and preschool teachers appears to have only small (and often not statistically significant) impact on the relevant labor markets. The evidence shows that occupational licensing does serve to increase the wages of electricians, ceteris paribus, but the evidence on plumbers was mixed and inconclusive. There was no discernible impact on the occupational safety of these two professions.

This reviewer would have liked to have seen in Kleiner (2013) also a chapter on the still hotly contested economic turf between medical doctors and sundry physician substitutes, notably nurse practitioners. It is a topic that has engaged economists ever since it was confronted boldly by Milton Friedman in his classic *Capitalism and Freedom* (1962, ch. 9). Friedman advocated permissive licensing (i.e., certification) of physicians in place of the more restrictive occupational licensing, which would have allowed nurse practitioners to hang up their own shingles in the delivery of primary health care, in independent practice from physicians. Perhaps an *ouvre* focused just on physicians is yet to come from Kleiner. It would be timely as the nation is said to be beset once again by an overall physician shortage,
especially at the level of primary care, and yet in some states, notably California, organized medicine still objects to the idea of independently practicing clinical professions, such as nurse practitioners (Reinhardt 2013).

On the other hand, Kleiner could respond that his chapter on dentists and dental hygienists is a very good proxy for a chapter on physicians and non-physician substitutes. In that chapter, Kleiner develops a model of dentistry that embodies features of a monopsonistic market in which a licensed, high-skilled profession, namely dentists, can harvest the rents from the occupational licensing of a lower-skilled profession, namely dental hygienists, through a requirement that the lower-skilled occupation may not practice independently of dentists, but must be employed and supervised by licensed dentists. Although Kleiner’s thesis does not represent a pure monopsonist model, it is an intriguing one, rendered graphically in his Figure 6.2. Kleiner’s empirical analysis of this market then shows that, as his theory predicts, occupational regulations requiring dentists to supervise dental hygienists yield dentists higher earnings and dental hygienists lower earnings and leads to lower overall employment of dental hygienists than do markets in which dental hygienists can practice independently as self-employed professionals. That finding has a direct bearing on the current, heated fight before California’s legislature over task allocation and professional independence in primary care between California’s Medical Association and nurse practitioners (Reinhardt 2013).

Overall, Kleiner’s work leads him to call for a pervasive review of occupational regulation in the United States, with a view towards replacing occupational licensure, which introduces the most inefficiency and welfare loss, with mere certification of occupations. That recommendation gains plausibility in an age where cheap computation and data mining makes it possible to protect consumers from low-quality and possibly dangerous services by providing robust, user-friendly information on the quality of services provided by competing occupations, e.g., nurse practitioners and physicians.

Economists can hammer away on this point, as they should; but under our system of governance, at all levels of government, special interest groups such as occupations seeking to be licensed to guard their economic turf can purchase the affection of legislators by helping to finance their political campaigns. Shifting the nation away from occupational licensure to certification, although not impossible, will be a hard slog.

It is remarkable, as Kleiner emphasizes, how rarely calls for occupational licensing come from clients who ostensibly are to be protected by occupational licensing. Both the occupations seeking licensure and the legislators who serve them will argue that practice acts are intended solely to protect the clients of the licensees. Many of them may even believe it. But it merely shows the capacity of adults to fool others to the point of fooling themselves.
Kleiner’s splendid volume is recommended reading for the many graduate students in economics specializing in industrial organization, labor economics, or health economics. Moreover, it calls out to doctoral students searching for research topics for their dissertations. Kleiner’s own extensive research and the literature he surveys in the concluding chapter can serve as a springboard for further research in this area.

References


About the Author

Uwe Reinhardt is the James Madison Professor of Political Economy and Professor of Economics and Public Affairs at Princeton University. His major research interests are fiscal/monetary policy and health policy. He has been a member of the Institute of Medicine of the National Academy of Sciences since 1978, and he has served on the editorial boards of the Journal of Health Economics, the New England Journal of Medicine, and the Journal of the American Medical Association, among others. His email address is reinhard@princeton.edu.
Capitalism and the Rule of Love

Preface by Daniel B. Klein

Here we reproduce an article by Clarence Philbrook (1909–1978). Philbrook received his Ph.D. in economics from the University of Chicago in 1947. He was professor of economics at the University of North Carolina from 1947 until 1975. Philbrook published a number of thoughtful, spirited articles in the Southern Economic Journal (see, e.g., the January 1953, April 1953, October 1954, April 1957, October 1957, and January 1961 issues), including the one reproduced here with permission of the Southern Economic Association, ©1953 (“Capitalism and the Rule of Love,” SEJ 19(4), April: 458-466). Also in 1953 he published an article in the American Economic Review, an article still read by some today, “‘Realism’ in Policy Espousal” (AER 43(5), December: 846-859). From 1959 to 1969 Philbrook was treasurer of the Mont Pelerin Society. I would like to suggest that the outlook of the present article is representative of that of the Mont Pelerin Society generally, then and now.

The article reproduced here seems to have been entirely forgotten. As of this writing, Google Scholar shows just two citations to “Capitalism and the Rule of Love.” I read “rule of love,” which Philbrook also denominates as “universal love,” as the ethic of universal benevolence suggested by Francis Hutcheson and Adam Smith (among others): The duty to advance the beauty of the scene of humankind beheld by a being, be it God or allegory, who is super-knowing, universal to humankind, and benevolent towards the whole of humankind. There is a long history to the invisible-hand notion that, by and large, the pursuit of honest income advances universal benevolence better the more that governments leave private enterprise free. But it was not until 1945, so far as I have discovered, that anyone discussed market phenomena—prices, residuals (profit and loss), inventories,
volumes, queues, and so on—as communication, that coming in Friedrich A. Hayek’s famous article “The Use of Knowledge in Society” (AER 35(4), September: 519-530). In that article, Hayek did not use the word signal, as in “price signals” or “market signals,” but such talk goes with the communication talk, as a signal is a piece of communication. Philbrook writes of “signals to tell individuals what they are to contribute toward the performance of some social function” (emphasis added). As one can see in Google Ngram charts (link), talk of “price signals” and “market signals” had barely begun as of 1953, the time of Philbrook’s essay. What is most striking about the essay is how it works from the rule of love and, on that framing, addresses the big questions of politics in terms of the generation and operation of “signals.” The root word signal occurs 17 times in Philbrook’s essay.

Meanwhile, pursuing honest income is only one side of the story. Philbrook also suggests that there are great opportunities to serve universal benevolence on the other side, distributing honest income and other resources of one’s own.

Philbrook likens the free enterprise system to a “mangy beast.” One will find his writing to be quite unlike anything now found in the Southern Economic Journal or the American Economic Review. A few remarks may be helpful.

A “poke,” as in the idiom pig in a poke, is a bag. Wikipedia says that the idiom refers to a “confidence trick originating in the Late Middle Ages, when meat was scarce, but cats and dogs were not. The idiom pig in a poke can also simply refer to someone buying a low-quality pig in a bag because he or she did not carefully check what was in the bag” (link).

A small rent in the poke is a tear or slit. “Rend your heart and not your garments” comes from the Bible (Joel 2:13). The preacher Charles Spurgeon, in his 1866 collection of devotionals Morning by Morning, offered this explanation: “Garment-rending and other outward signs of religious emotion, are easily manifested and are frequently hypocritical: but to feel true repentance is far more difficult, and consequently far less common” (emphases in original).

At a few places, particularly in its second paragraph, Philbrook’s essay prompted me to insert clarifying words contained in square brackets [like this]. We have also corrected three or four minor typographical errors. But otherwise the essay is unchanged and complete, including the two footnotes from the original SEJ printing.

The final words of Philbrook’s essay bring to mind these by Adam Smith: “The fatal effects of bad government arise from nothing, but that it does not sufficiently guard against the mischiefs which human wickedness gives occasion to” (The Theory of Moral Sentiments, IV.2.1).
Capitalism and the Rule of Love:

Clarence Philbrook

“...the spectre of capitalism. Little longer will it rattle its chains at those who try to make a better world.” The Communist Manifesto, if slightly revised for today, might well open with these words. Yet such a statement would be consistent with the thinking of a large group to whom Communism itself is a dread spectre—a crucially large proportion of the intellectuals of the United States.

Many [such intellectuals] are content to argue simply that the end of capitalism is now inevitable. The conditions necessary for its existence are gone, they say, and we cannot “turn the clock back”: capitalism is no longer a live option. Probably few, however, would strongly press this notion but for the fact that they believe the change to be desirable. As a means of promoting the decline of the old system, assertion of inevitability is an effective argument. In these days of vast uncertainty about principles, few [intellectuals] are willing to espouse a losing cause; the way [for opponents of capitalism] to win support is to create an appearance of power. Under questioning, those who rest upon this reasoning [that is, of an inevitable end of capitalism] usually do make clear that they welcome the fall of capitalism. Moreover, many others are avowedly glad to see the capitalistic system go down, but forthrightly consider its obliteration to depend upon our wishes and resulting actions. Thus the real issue is whether the departure from genuine capitalism—private enterprise—is good or bad. Given our currently ruling attitudes, the system is indeed probably doomed; but there can be no serious doubt that its decline is largely due to the fact that we simply do not have a consensus to the effect that the old ways are worth retaining. As to whether capitalism actually is

1. This article was written with non-economist readers in mind. With some reluctance the author released it for publication here [in the Southern Economic Journal], influenced by argument that the controversy invited by its substance, while anything but new in a calendar sense, remains unfinished business of the first order among students of political economy.

2. University of North Carolina. Mr. George M. Woodward, Major Robert L. Bunting, and Dr. Edwin J. Stringham have given most generous help toward narrowing the range of defect of this paper.
or is not desirable, the extreme difficulty of arousing properly focused discussion of this question is one of the most profoundly disturbing facts of our day.

The most desperately needed focus of discussion requires the following consideration. No social movement—be it the rise or decline of Christianity, of Catholicism, of Protestantism, of the welfare state, or of capitalism—can be supposed free of scoundrelly self-seekers. For present purposes we need not speak of the role of such men. They must be dealt with; but there is little hope of success unless those of us who search for the means to the good society can reach a consensus. We may speak until further notice, then, of and to only those who mean to take ethical considerations seriously. Now, one way or another the beliefs of men highly sensitive to ethical values have, whether “religiously” orientated or not, converged in the conception of universal brotherhood—of universal love. Social arrangements are, according to this view, to be judged by the degree to which they are in keeping with that principle. It is to opponents of capitalism who conceive their opposition to stem from such an interest that this discussion is directed. In so far as such persons have influenced events, it is fair to say that the reason for the decline of capitalism has been a growing desire to bring about conditions more in keeping with the rule of love. Now, in spite of the strong appeal of this motive, the conclusion to be offered here is emphatic: the destruction of capitalism would constitute a tragedy to the human species. It must be made equally unambiguous, however, that this conclusion in no way requires repudiation of the fundamental ethic of love. Much of the argument involving persons who stress this ethic runs in terms which suggest on their part a belief such as the following. Either men are interested in ethics or they are not. If they are, ipso facto they will repudiate capitalism. If they are not, only then may they support it. The fallacy of thinking in terms of such a simple dichotomy must be recognized. The belief expressed here rests on the only ground which ultimately can support a preference for a particular social system: that capitalism does less violence to the rule of love than would any other system so far conceived.

To perhaps most critics of the capitalistic system it will appear that the desire to give greater sway to brotherly love not only would be capable of leading to the destruction of capitalism, as we have suggested it has in fact done, but actually must do so if given any significant expression at all. To them it will appear that subjecting so evil a thing to any institutional embodiment of love would be like confronting Mephistopheles with a crucifix. The matter is not quite that simple. However, it is true that, although their realization does not require it, our finer aspirations may, and latterly do, result in practices such that in their presence that fiendishly ingenious and productive Lucifer, private enterprise, will faint and fail.

Broadly speaking, our manner of bringing the private-enterprise system to its present precarious state consists of the act of governmental determination—direct
or indirect—of the prices or the quantities of commodities or services, including the services of productive resources. By indirect governmental determination is meant the deliberate placement of an individual or a collectivity in a position to fix the price or the quantity. Myriad examples of destructive action spring to mind: resale-price maintenance laws, tariff protection of industries, control of foreign exchange rates, price floors and ceilings for specific commodities and services, government organization of farmers for collusive fixing of outputs, encouragement of labor monopolies, and the like.

To most it comes as a shock to hear an assertion that such actions are destructive. What could be more reasonable, they think, than to change the various elements of the economy so as to do social good? And if indeed such actions do destroy the system, what of it? Is man made for the system, or the system for man? Such reactions suggest insufficient contemplation of the role of systems, or institutions, in society. Let the following proposition be considered: there is not one single institution but sometimes brings results which we should, if we knew all the facts, consider evil; and concerning each we agree that we must accept from it some evil for the sake of the net good it does. Who would dare maintain that justice is in every case aided by the right of habeas corpus, by freedom of the press, by jury trial, by the institution of marriage, by punishment of theft and homicide? Yet who would, on account of the certainty that these systems often yield bad results, advocate casual departure from the rule? In nearly all spheres but the economic we universally take for granted that a system, in order to supply the poor best which human devices can provide, must be allowed to operate according to a logic of its own.

Institutions which we can meaningfully speak of reforming probably always can be described as sets of signals and incentives: signals to tell individuals what they are to contribute toward the performance of some social function, and incentives to perform the actions called for. If a system (institution) is altered, the change must be made with care to see that the necessary functions are provided for by means of specific signals to specific individuals, along with incentives to action. Moreover, the manner of operation for the performance of one function imposes some kind of limitations upon the types of change in the manner of performing some other function, in exactly the same sense as the gear arrangement on an automobile rules out shifting into reverse while moving forward. Thus, for any system which is to be improved rather than eliminated, while certain types of reform are appropriate, others are inappropriate simply because they violate the essential mechanism, even though in a partial view they do good.

Now the private-enterprise system is one of our numerous and always highly imperfect sets of signals and incentives, and as such has a logic of its own. Money incomes are in effect books of ration coupons, and are to be distributed in a manner
as just as we can manage, all things considered. Each person is to choose his own combination of goods, subject to the condition that the total of the individual rations of any commodity must add up to the amount available to society. If all of us together are trying to use more than that, the fact is signaled by a rise in price which furnishes an incentive to each of us to help serve the social need by cutting corners—economizing—in the use of this relatively scarce commodity. In a similar process any productive resource (for example, labor, coal) also is “rationed.” The “ration coupons” received by a business man for his goods are evidence that he is to that extent the agent of members of the public for purposes of exerting their claims to the use of resources for producing what they want. With these “coupons” business men bid for each kind of productive resource until a price is set upon it which “rations” that resource among different commodities according to the desires of the public. The price of the resource too is both a signal and an incentive: a rise signals that products of the resource are more wanted than before in relation to the amounts available and that society therefore needs to have this resource economized more drastically. The incentive to help society economize arises from the fact that the business man can now get more product for his expenditure by substituting some of the less scarce resources and that the higher costs bring about higher prices for the products of the scarce resource. Thus, the process and the possibility of social economizing through private enterprise depend squarely upon our allowing prices to be fixed by the market.

In spite of the essential character of the private-enterprise system, recent economic reforms have to a marked degree leaned upon group interference with individual prices and quantities. Most of our intellectuals, including many economists, enthusiastically support these reforms, and many call for more. What can we make of the implied urge toward the destruction of capitalism? Let us consider a parable of pigs.

Tom Jones insisted that I should exchange my visible pig for the one in his poke. He correctly pointed out what a mangy beast was mine, and called down the laughter of assembling passers-by upon my hesitancy to trade. He now and then allowed to protrude, through a small rent, the tail or an ear or the snout, to each of which he pointed with pride; but he became quite infuriated at my insistence upon seeing the whole pig. The crowd, moreover, shared the irritation. To be sure, for all that any single member of the throng really knew, if he himself had actually seen Tom’s pig, he might have agreed that I ought to retain my own poor creature. Each, however, seized upon his private picture of the poke pig. Each did not, on the other hand, picture a whole pig, but the pig in his mind had such fine hams, or such beautiful jowls, or such excellent bristles, or what not! Thus, perhaps no member of the crowd would himself have preferred to my pig the one in the mind of anyone else. Still more likely, no one would have wanted an animal consisting of
a composite of all the features imagined by individuals. Yet, in sheer terror of the throng, I came to feel that my pig would be torn from my arms unless I showed that any pig which could possibly be in that poke must be worse than mine. That is approximately the position in which we find ourselves, we few who still really believe in the enterprise system.

It is impossible, of course, to discuss all pigs which conceivably might be in the poke. Indeed, it is shocking that the crowd should force me to try to do so—that it should deride me for hesitating to trade and not Tom Jones for refusing to show his animal. Any Jones may reasonably ask that I consider the flaws of my pig, which are indeed grave; but if he is to ask serious consideration of any exchange, a minimum requirement upon him is that he shall, first, show awareness of the parts which a pig must have in order to live and, second, show that his pig has those parts.

Fortunately we can, if we but will, derive crucially important insight simply from this reluctance to open the poke, in conjunction with the fact that much trading has already been forced. In our parable we have till now taken for granted that there was in the poke a functioning pig of some sort, which I was to receive intact. Actually, the case was somewhat different. The throng partially dismembered my pig. It then repaired the damage by allowing various persons to reach into poke after poke, snatching out individual parts to be grafted upon my animal. My cries that the parts did not fit were ridiculed: were we to treat as sacred the products of blind evolution, as if a benign hand had wrought their symmetry? My appeals for closer inspection of the parts were considered to be nothing but hypocritical delaying tactics mixed with escape into poetic contemplation of anatomy. My cry of “Tiger claw!” was thought to be largely an effort to deprive the animal of the fine hoof in which I said the claw was concealed. A perverse objection to his having an excellent tusk was supposed to account mainly for my saying it was a snake fang. As the analogy suggests, we have, by our administrative manipulation of individual prices and quantities, thrown awry our old set of signals and incentives; but we have outlined no new set, to say nothing of inspecting—or even being tolerant toward efforts to inspect—each member with care to see whether the collection we have really constitutes an internally consistent system. We have triumphantly repudiated the market principle, but have supplanted it with no other criterion-plus-means-of-administration (of a price or a quantity) which would create any presumption of a tolerable or viable social organization.

The concern for consistency is generally written off as the “hobgoblin of small minds”; but are we not even to consider, for example, whether restricting crops to help “the farmer” may decrease the demand for hired farm labor and thus worsen the lot of one of the most poverty-stricken groups? This issue, however, important as it is, fades into secondary significance in comparison with the
problems which focus in the insistence that we must possess generally-accepted criteria—signals to guide action upon prices and quantities. The stock answer to statements of this need is that our new principle calls for doing right in each instance according to the merits of the case—in accordance with the rule of love to establish fair prices, fair shares of the product; but these are weasel words. The very question is, what principle is available to tell us what is right, what are the merits of the case, what is in accordance with the rule of love? The problem is represented dramatically in the distribution of income. Each citizen ought to search his soul over this, for few questions if any are as crucial in this century: at just what stage in their relative-wage gains through organization shall we say to, for example, coal miners, “Now you have your deserts; you must go no farther”? Absolute equality of money income—for which surely we ought to wish—would at least be a perfectly clear-cut criterion; but the valid objections to forcing that outcome are so universally recognized that we must take for granted that a different principle is required—one telling exactly how much inequality is to exist. It is difficult to engender thought upon this problem in the right perspective, for many persons can conceive of patterns of relative income which they themselves would consider fair. But to regard one’s own judgment on this matter as a test is to miss the point completely. It is not sufficient that the principle carry conviction to you or me as disinterested bystanders: it must carry conviction to the persons standing to lose by its application. Otherwise men will constantly feel that they are treated in an unprincipled manner. In these circumstances any economic adjustment whatever will represent merely an uneasy truce tolerated only as an opportunity for a build-up of power—through strengthening of organizational ties, propagandizing the public, and, increasingly, amassing of political influence even to the degree of raising doubt that the law can be enforced against particular groups. “We cannot mine coal with bayonets.” This is a recipe for chaos.

In the absence of accepted criteria, fewer and fewer disputes over prices of any kind will be settled short of the influence of the President or his deputies. If we do not at one time grant to the chief executive the power to seize parts of the economy, we shall later prepare emergencies which require us to do so. To suppose that labor groups will remain exempt is absurd. Moreover, the triteness of a further idea must not be mistaken for insignificance. Where every price is potentially a political issue, the burdens of the chief of state become overwhelming; he cannot be expected to give more than haphazard thought to mighty issues. To a man maddened by the intolerable responsibilities thrust upon him, criticism of his honest efforts to maintain economic activity must look like sheer, perverse obstructionism. Why should we be surprised, then, if some president did successfully seize the press and radio? At some stage of chaos he would have the populace behind him; we should accept with relief the leadership principle—dictatorship.
Be it noted that all this can happen where none but good men are involved. But bad men both find and make opportunities in these conditions; and, since fewer qualms limit the effectiveness of their dealings with “obstructionists,” there is a real tendency for bad men to rise to the top. Unwise efforts to enforce the rule of love contain enormous danger of its negation.

In the international sphere comparable developments may be expected from our systemless procedure, except for possibilities of quicker and more literal destruction of civilization. There, interferences with prices constitute actions favoring one people against another. Of course national honor becomes involved, as well as the need of ruling politicians to save face with their electorates; and there is not even a dictator to referee. How many doubt that the Iranian oil settlement has been made especially difficult by the fact that a government was selling to a government-owned buyer? We might conceivably have a fair amount of peace without an international government, but we certainly cannot unless the impersonal market is allowed to make many of the disagreeable but necessary adjustments which, when settled by political discussion in the absence of criteria, necessarily become cases where each of two cohesive groups honestly believes the other is trying to “put something over on it.”

There is no use in looking in other directions for the barrier which divides “capitalistic” and “anti-capitalistic” seekers for a good society: it arises out of this problem of criteria, of signals and incentives. Huge numbers of defenders of capitalism would be won over to a system aimed at more equalitarianism—more of the rule of love—if only a satisfactory answer were forthcoming on the crucial question. Once the market is repudiated, what principles—signals and incentives—are to perform the functions of prices—that is, the rationing of goods, the allocation of resources, and the distribution of income, in the light of the fact that the total product is drastically influenced by incentives to effort? The short-cut to consensus is conceptually clear: let those who wish to depart from the market system propose carefully conceived sets of principles and encourage discussion and criticism of them. It would seem that this suggestion ought to be greeted as a most friendly and constructive one. Actually, it serves merely to irritate critics of free enterprise. Yet, until we can see signals and incentives which promise to carry us through the storms which rage upon the repudiation of the market, some of us will defend an enterprise system—not entirely to protect our vast properties, or because we delight in the sufferings of the poor, or even to curry favor with industrialists, but because free enterprise, bad as it is in comparison to our dreams, seems to offer possibilities of embodying more of the rule of love than we so far see how to embody in any different system.

That “men of good will” should join forces for social improvement is most highly to be desired. Why is it that communication between the opposing schools...
of thought is so ineffectual toward consensus? Obviously there must be crucial differences between the underlying premises which determine their respective attitudes toward the need of signals and incentives and their respective appraisals of particular sets.

First, consider the “anti-capitalists.” As to incentives, some appear to believe that, since all men ought to be saintly, we must act as if in fact they were. This seems to be the position of the more gentle and appealing reformers. More common is an implicit if not explicit idea that the ninety per cent, or the fifty-one per cent, or possibly the five per cent, comprising the “men of good will” ought to whip the sinful into acting as saints would. Often there is also, sustaining these views, an implicit belief that production possibilities are so limitless that no strong incentives are needed to maintain adequate production.

As to signals, again the idea that output has no limits blinds these reformers to the need of any highly effective indicators; for in an “economy of abundance” it matters little if considerable quantities of resources are used for purposes less important than some other, less attended needs. But beyond this there is a more fundamentally divisive idea. It is apparently a belief that the problem of life is one of satisfying to the maximum possible degree a set of known wants. Believing they know the true values in a detailed way—including how much of what goods it is desirable for each person to consume—they are confident that a good majority could be brought, after the removal of capitalism, to adopt right and definite ideas of the correct price or quantity of each good or service, who is to produce each in what manner, and how the total is to be divided among persons. Among other flaws here is a fantastic underrating of the vast number of decisions to be made and of the limits to the scope of attention of a human being. Even now, how many think they could even list five per cent of the specific decisions made governmentally, to say nothing of passing sound judgment upon them? The device intended to meet this difficulty is, of course, to make officials responsible for good decisions on details; but, if decisions are too numerous for citizens to review, how can the soundness of officials be judged? However, under the influence of this belief, in response to questions about how it is proposed to meet this or that problem quite ready replies are forthcoming in terms of “society would do thus and so.” It has been playfully suggested that all such statements should be legally required to run: “the politicians in power would” and so forth. The implication of this jest is the basis for the correct charge that such reformers in fact picture themselves (in their most idealistic frame of mind) as the effective determiners of policy in the brave new world. Since there would, to be sure, remain at least some unreasonable men, it is taken for granted that it would be ethically desirable to take decisions out of the hands of recalcitrants. Reformers with these thoughts on signals and incentives
may happily contemplate a repudiation of capitalism and feel it to be a sure step toward enlargement of the role of love in human life.

Defenders of capitalism, on the other hand, insist upon “seeing the whole pig” before exchanging. They are highly dubious of any signals and incentives which could be expected to emerge in the process of departure from the old system. The reason lies in some crucial parts of what seems to be the inherent logic of true liberalism.

As to incentives, this reasoning includes some equivalent of the proposition that man, whatever may be his ultimate perfectibility, is sinful. The relevant part of the thought is that the element of saintliness, although present in the worst of us, is insufficient in all but rare persons either to keep us striving day in and day out in the labor which life requires or to assure against our abusing others if placed in positions of power. We simply have no way, as yet, either of calling forth necessary effort or of preventing mutual mistreatment, unless in considerable degree through the motive of self-interest. If room is to be made for the employment of such brotherly love as we do have, it must be made without yet eschewing the use of the opposite, powerful force. If day-to-day reliance upon self-seeking precluded behavior according to the principle of love, the case might be different; but the common idea that it does so is sentimental and vain twaddle. If we act selfishly, the reason is that we are selfish. All are free either to donate their services or to earn by service and devote their earnings to good works. Statements to the contrary effect in reality mean, “I should be quite virtuous, if it cost me nothing.” The proper order of reform is first to attain great virtue and only then organize in a manner which requires it.

As to signals to guide action, the logic of the traditional liberal rests on the idea that life is best viewed as a quest for the good life—a seeking for the details of behavior and attitude which constitute the correct content of what some would describe as loving God, others as experiencing truth: it is this in contrast to a process of maximizing the satisfaction of a set of known values. If there is a way of advancing us all toward higher understanding and realization of values, that way consists in each spirit’s searching freely, experimenting, making errors, learning, and sharing with others its discoveries. Indeed, man is simply not man unless he participates in such a quest. Thus, the thing to be guarded, even at tremendous cost if necessary, is freedom, in the common-sense meaning of freedom from arbitrary dictation to one soul by another—a meaning, incidentally, known to all before some reformers resorted to semantic trickery and corrupted our universe of discourse. We must not, then, guide action by decisions made by uneasy (or even easy) compromise among the fifty-one per cent and forced upon the forty-nine, except where there is simply no other way available. What the rule of love calls for above all surely is non-interference with the quest. Decisions must stem...
from the tastes and ideals of men, freely developed and freely expressed. This is the fundamental purpose served by a market system—a great system of proportional representation applied where proportional representation makes sense. The same set of institutions which permits conspicuous consumption assures that an Albert Schweitzer will not be deflected from his destiny by some administrator who believes that the morale of the people calls for Schweitzer’s music.

All who believe that capitalism constitutes the least bad society now available to us must deplore the necessity of opposing themselves to the fine persons who are so numerous among the destroyers of the system. The motive of these opponents of private enterprise has an all but irresistible appeal to sympathy: it patently stems from belief that a great store of mutual understanding and good feeling awaits only liberation from an outworn set of laws to gush forth and carry more of love into life. But the difficulties are far deeper than such thought suggests. The old-fashioned type of prophet’s cry still is needed: ye people, rend your hearts—rend your hearts, and not your garments! We are not forced to assign limits to the virtue to which it [that is, the cry to rend our hearts] may lead us; but we are forced to doubt that its work has yet been done. We can scarcely be impressed by the spectacle of men rending their garments and crying, “The System made me the imperfect thing I am!” We may well be appalled by the spectacle of men rending the garments of others and crying, “We are the good! Give us power to rend the hearts of all who fail to conform to our precepts!” For each to rend his own heart there is much reason. It should serve at least to convince us that we are not yet fit for a society of saints. It might lead to removal of the political and intellectual barriers to significant improvement of the results of a free-enterprise system.

Capitalism is capable of giving us a much better society than we have known. Even apart from its fabulous tendency toward increased production, immense change expressive of the rule of love is available in that depression can be largely eliminated and inequality of income mitigated, both of these by methods quite in keeping with the logic of the system. Moreover, fantastically more brotherly love than has ever been exercised can be given expression through individual attitude, decision, and action in a capitalistic society. But if we repudiate that system by making changes which conflict with its essential mechanism, we give up one of the few protections we have against the evil that is in us.