Econ Journal Watch
Scholarly Comments on Academic Economics
Volume 17, Issue 1, March 2020

Editor’s Notes: Acknowledgments, 2018 Through March 2020 1–3

COMMENTS

Revisiting the Bracero Guest Worker Reforms: A Comment on Clemens, Lewis, and Postel
Robert Kaestner 4–17

Comment on Kaestner, “Revisiting the Bracero Guest Worker Reforms”
Michael A. Clemens, Ethan G. Lewis, and Hannah M. Postel 18–27

Brought Into the Open: How the U.S. Compares to Other Countries in the Rate of Public Mass Shooters
John R. Lott, Jr., and Carlisle E. Moody 28–39

The Importance of Analyzing Public Mass Shooters Separately from Other Attackers When Estimating the Prevalence of Their Behavior Worldwide
Adam Lankford 40–55

Do Film Incentive Programs Promote Economic Activity? A Comment on O’Brien and Lane
John Charles Bradbury 56–65

Reply to Bradbury: Effects of Economic Incentives in the American Film Industry
Nina F. O’Brien and Christianne J. Lane 66–70
Farley Grubb 71–89

Science on FDA Liberalization: A Response to the Status Quo Process for Medical Treatments
Bartley J. Madden 90–97

**INTELLECTUAL TYRANNY OF THE STATUS QUO**

Government-Cheerleading Bias in Money and Banking Textbooks
Nicholas A. Curott, Tyler Watts, and Benjamin R. Thrasher 98–151

**CHARACTER ISSUES**

The Stewart Retractions: A Quantitative and Qualitative Analysis
Justin T. Pickett 152–190

Captive of One’s Own Theory: Joan Robinson and Maoist China
Evan W. Osborne 191–227

**WATCHPAD**

Edward Leamer Deserves a Nobel Prize for Improving Argumentation That Uses Statistics
Arnold Kling 228–241

It Will Soon Be 1984…
Ingemar Ståhl 242–250

Afterword to “It Will Soon Be 1984…”
Lars Jonung 251–255

Bentham Versus Blackstone
Gertrude Himmelfarb 256–269
Editor’s Notes: Acknowledgments, 2018 through March 2020

We are grateful to our institutional home and friend the Fraser Institute, Canada’s leading think tank, in particular Jason Clemens, Chris Howey, and Cheryl Rutledge. For generous support we thank the Charles Koch Foundation, the John William Pope Foundation, the Richard Seth Staley Educational Foundation, and Gerry Ohrstrom (through Donors Trust). I am very grateful to Rich and Mary Fink and the Fink family for the generous Mercatus Center JIN chair, which I hold and that helps to support EJW.

I would like to extend thanks to co-editors Brendan Beare, George Selgin, and Larry White, editorial advisor Jane Shaw Stroup, web designer and master John Stephens, and my chief partner in the project, managing editor Jason Briggeman. Also, Kurt Schuler, for serving as guest lead editor on a major article.

We are grateful to our authors for generously contributing their creativity, craftsmanship, and industriousness, as well as for their patience and cooperation in our editorial process.

We thank the following individuals for generously providing intellectual accountability to EJW:

Refereeing February 2018 through March 2020

<table>
<thead>
<tr>
<th>Fernando Arteaga</th>
<th>University of Pennsylvania</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendra H. Asher</td>
<td>George Mason University</td>
</tr>
<tr>
<td>Sunday Azagba</td>
<td>University of Utah</td>
</tr>
<tr>
<td>Markus Bjoerkheim</td>
<td>George Mason University</td>
</tr>
<tr>
<td>Patrick Button</td>
<td>Tulane University</td>
</tr>
<tr>
<td>Bruce Caldwell</td>
<td>Duke University</td>
</tr>
<tr>
<td>Noah Carl</td>
<td>Independent research</td>
</tr>
<tr>
<td>Young Back Choi</td>
<td>St. Johns University, New York</td>
</tr>
<tr>
<td>Gregory Conko</td>
<td>Competitive Enterprise Institute</td>
</tr>
<tr>
<td>Tyler Cowen</td>
<td>George Mason University</td>
</tr>
<tr>
<td>Clayton Cramer</td>
<td>College of Western Idaho</td>
</tr>
<tr>
<td>Jakob de Haan</td>
<td>University of Groningen</td>
</tr>
<tr>
<td>David Henderson</td>
<td>Naval Postgraduate School, Monterey</td>
</tr>
</tbody>
</table>
Erwin Dekker
George DeMartino
Arthur Diamond
Michael Eames
Lanny Ebenstein
Joseph E. Gagnon
Susan Howson
Jeff Hummel
Douglas Irwin
Jonathan Imber
Matthew Jaremski
Garett Jones (3x)
John Kane-Berman
Pavel Kuchar (2x)
George Lady
Mitchell Langbert
Marc Lavoie
Pierre Lemieux
Guy Madison
Michael Marlow
Cesar Martinelli
Bruce McCullough (2x)
John McCusker (2x)
Ronald Michener
David Mitch
Juan Carlos Odar
Lawrence Officer
Pia M. Orrenius
Morten Ougaard
Adrian Pagan
Sam Pelizzo
Carlos Ramirez
Jon Ratner
Hugh Rockoff (2x)
Raymond D. Sauer
John J. Siegfried
Michael Thorn
Christopher Torr
Veli-Matti Törmälähto
Marian Tupy
Jorge Ugaz
Richard Woodward

Erasmus University Rotterdam
University of Denver
University of Nebraska, Omaha
Santa Clara University
University of California, Santa Barbara
Peterson Institute for International Economics
University of Toronto
California State University, San Jose
Dartmouth University
Wellesley College
Utah State University
George Mason University
South African Institute of Race Relations
University of Bristol
Temple University
Brooklyn College
University of Ottawa
University of Québec in Outaouais
Umeå University
California Polytechnic State University, San Luis Obispo
George Mason University
Drexel University
Trinity University, San Antonio
University of Virginia
University of Maryland, Baltimore County
Phase Consultores
University of Illinois, Chicago
Federal Reserve Bank of Dallas
Copenhagen Business School
University of Sydney
University of Chicago
George Mason University
Westat
Rutgers University
Clemson University
Vanderbilt University
University of Southern California
University of the Witwatersrand
European Central Bank
Cato Institute
Mathemaica
Coventry University
Authors replying to comments in EJW,
published January 2018–September 2019

Nejat Anbarci
Durham University
(article link)

Abhay Aneja
University of California, Berkeley
(article link)

K. Peren Aren
Australian National University
(article link)

Andrew Bird
Carnegie Mellon University
(article link)

Andrew Call
Arizona State University
(article link)

John J. Donohue
Stanford University
(article link)

John J. Donohue
Stanford University
(article link)

Farley Grubb
University of Delaware
(article link)

Stephen A. Karolyi
Carnegie Mellon University
(article link)

Adam Lankford
University of Alabama
(article link)

Nathan Y. Sharp
Texas A&M University
(article link)

Kyle D. Weber
Columbia University
(article link)

Jaron H. Wilde
University of Iowa
(article link)

Christina Zenker
University of St. Gallen
(article link)

Other individuals who published commentary on EJW material in EJW,
published January 2018–September 2019

Hannes Gissurarson
University of Iceland
(article link)
Revisiting the Bracero Guest Worker Reforms: A Comment on Clemens, Lewis, and Postel

Robert Kaestner

Michael A. Clemens, Ethan G. Lewis, and Hannah M. Postel (2018) report results of an impressive analysis of the effect of ending the bracero guest worker program on domestic farm workers’ wages and employment. The authors are to be praised for the thoroughness with which they execute their analysis. Clemens, Lewis, and Postel (hereafter CLP) conclude that the ending of the bracero program had little effect on wages and employment of domestic farm workers. The finding is inconsistent with the idea that a decrease in guest-worker labor supply would raise wages and increase employment of comparable domestic workers. The authors suggest that the explanation for their finding is that employers responded to the end of the bracero program and the decrease in labor supply by altering the technology of production. If true, the conclusions of CLP (2018) might suggest directions for reform in immigration policy.

In this comment, I make three points. First, I discuss the quality of the data itself, as well as how that data were used, finding significant problems, such that, ex ante, it is unlikely that the empirical analysis would be able to identify reliably a labor market effect of ending the bracero program. Second, I assess the likelihood of whether the beginning or end of the bracero program had a significant effect on farm labor supply. Based on the evidence I review, I conclude that there was likely significant substitution between legal and illegal Mexican labor and a much greater increase in the pool of domestic farm labor that likely muted, if not completely

1. University of Chicago, Chicago, IL 60637.
offset, any changes in labor supply associated with the bracero program. As a result, there is little likelihood that changes to the bracero program would have had an easily measurable effect on the wages and employment of domestic farm labor. Third, I assess the plausibility of CLP’s argument that endogenous technological change, caused by ending the program, explains the absence of an effect of ending the bracero program. I report evidence that casts some doubt on this explanation. Overall, I find that the analyses of CLP (2018) are not sufficient to draw the conclusion that ending the bracero program had no effect on the wages and employment of domestic farm labor.

**Clemens, Lewis, and Postel (2018)**

**Research design and results**

The research design used by CLP to estimate the effect of bracero guest workers on domestic farm laborers’ wages was the well-known difference-in-differences (DiD) approach. CLP compared the change (difference) in domestic farm workers’ wages before and after the end of the bracero program among farm workers affected by the bracero program to the analogous change (difference) in wages among farm workers unaffected by the bracero program. The bracero program had brought a few hundred thousand Mexican farm workers into the U.S. to work in the agricultural sector. The prediction of basic economics is that ending the program would result in an increase in domestic farm laborers’ wages because of the decreased supply of farm workers. Instead of this prediction being borne out, the comparison of wages conducted by CLP suggests that the increase in bracero guest workers had no effect on domestic farm laborers’ wages. The result might seem surprising because it suggests that foreign workers may not drive down wages of counterpart domestic workers.

**Problems with CLP’s analysis**

There are several problems with CLP’s DiD analysis, and all the problems make it unlikely that they would find an effect of ending the bracero guest worker program even if there was a true impact.

**Mis-measured treatment**

One of the most important problems relates to how CLP identified domestic
farm workers more or less likely to be affected by the end of the bracero program. Getting this classification wrong will seriously bias estimates of the effect of ending the bracero program. If workers classified as more likely to be affected were not, or less, likely to be affected by the end of the bracero program, and workers classified as less likely to be affected were, or were more, likely to be affected, then the comparison of domestic farm workers’ wages pre-to-post the end of the bracero program is likely to find no effect. As I now discuss, CLP’s classification of workers was almost surely inaccurate and seriously so.

CLP defined risk of being affected (i.e., exposure) by the end of the bracero program as the share of bracero labor of all seasonal farm labor during the peak harvest season in a state in 1955. This state-specific share ranged from zero to 60 percent. In 1955, there were 23 states with any bracero labor; 17 states with a share of 1 percent or more; 9 states with a share of 5 percent or more; and 6 states with a share of 20 percent or more. Remember that this 1955 bracero share of peak season labor is supposed to represent the change in labor supply brought about by the end of the bracero program in 1965. If the 1955 share does not measure accurately the change in labor supply before and after the end of the bracero program, then the comparison at the heart of CLP’s DiD analysis will be correspondingly inaccurate. Classifying a state by its 1955 share of bracero labor during the peak harvest season is likely problematic. The share of bracero labor in a state changed markedly over the calendar year and during the years that the bracero program was in effect (Bureau of Employment Security, Farm Labor Market Developments 1960; 1964). For example, in 1963, the year just before the end of the bracero program, the amount of bracero labor in states was substantially different than in 1955. In 1963 there were only 13 states with any bracero labor. In addition, the ranking of these states was much different than the ranking based on the 1955 share. The state with the highest share of bracero labor in 1963 was Arkansas with 46 percent. Arkansas’s 1955 share was 8 percent and its 1960 share was 18 percent. The second highest share in 1963 was Colorado at 40 percent. Colorado’s 1955 share was 4.5 percent and its 1960 share was 11 percent. South Dakota had no bracero labor in 1963 or 1960, but in 1955, the bracero share was 21 percent. The share of bracero labor in Texas, which often used either the most or second most amount of bracero labor, was 24 percent in 1955, 18 percent in 1960 and 7 percent in 1963.

To provide a summary of the variation in the bracero share of labor over time and by state, I calculated the correlation between the 1955 bracero share of labor and the actual bracero share of labor in a state for several years. In 1953, which is the first year of data provided by CLP, the correlation was 0.75 among all states

---

2. In 1955, New Mexico is clearly an outlier at 60 percent share. The next closest state was Nebraska at 32 percent.
and 0.69 among states with a positive bracero share of labor in 1955. Analogous correlations in 1957 (1960) were 0.93 (0.92) and 0.92 (0.90), respectively. By 1963 the correlations were 0.77 for all states and 0.70 for states with a positive share of bracero labor in 1955.3 These changes in the level of the share of bracero labor by year suggest that the use of the 1955 share of bracero labor would result in misclassification of states and attenuate DiD estimates. At its most basic level, the DiD approach used by CLP measured the change in the average wage from 1942 to 1964 and the average wage from 1965 to 1971 (or 1975 for a different wage measure) and correlated it with the 1955 share of bracero labor, which, because bracero labor was zero post-1965, measures the change in labor supply. As I described above, the 1955 bracero share of labor does not reflect accurately the share of bracero labor in the period from 1953 to 1964, and as I describe below, is a poor measure of the share of bracero labor pre-1953. These issues make it likely that the DiD analysis would find no effect. It would be surprising to find anything but no effect because the comparison of farm workers’ wages pre-to-post the end of the bracero program was not between workers more or less affected by the end of the program.

Also, there is significant variability in bracero labor over the year. For example, in May of 1955, there were 74,059 bracero laborers in the United States. In October of 1955, there were 135,600 bracero laborers in the United States (Bureau of Employment Security, Farm Labor Developments 1955–1956). Note how both of these numbers are much lower than the total number of bracero laborers admitted in 1955, which was 398,650. Similar figures from 1961 are: 78,400 in May and 208,500 in October. The October to May ratio of bracero labor was approximately two in 1955, but approximately three in 1961. This May-October difference shows the variability of bracero labor over the calendar year. This is important because the peak-month share of bracero labor in 1955 was used to indicate treatment—the variable to correlate with wages that are measured on a quarterly basis. Thus, even within the year 1955, the peak-share of bracero labor was not an accurate measure of the amount of bracero labor in other quarters in 1955. The variability over the calendar year and the previously documented variability over years mean that CLP’s approach is likely to lead to significant measurement error in the treatment indicator and seriously biased estimates of the effect of ending the bracero program. Again, one might be surprised to find anything but no effect given the classification problems.

3. Note that CLP also estimate a regression using wage data from 1960 to 1970. The low correlations between the bracero share in 1955 and 1963 are particularly problematic for this analysis.
Definition of before and after period

While 1965 was the official end of the bracero program, there had been large changes in the number of bracero workers admitted in years prior to 1965 that were as, or more, significant than the change that occurred in 1965 (Bureau of Employment Security, Farm Labor Market Developments 1964). From 1948 to 1955, the number of bracero laborers admitted increased from 35,345 to 398,650, or by a factor of 11.4 Between 1950 and 1951, the number of bracero laborers admitted approximately tripled from 67,500 to 192,000. The period from 1956 to 1959 experienced the largest number of bracero workers admitted to the United States with approximately 440,000 Mexicans admitted each of these years. Between 1959 and 1960, the number of braceros admitted declined by 121,797, or 28 percent. Similarly, between 1961 and 1962, the number of braceros admitted declined by 96,442, or 33 percent. From 1962 to 1964, the number of bracero laborers admitted remained relatively constant at approximately 180,000 (177,736 in 1964). So, while the end of the program in 1965 brought about a significant decline in the number of braceros, the decrease was not that much larger than in other periods. In addition, the increase in bracero labor from 1948 to 1955 was as large as the decrease in bracero labor from 1959 to 1965.

The great rise and fall of bracero labor suggest that the end of the bracero program in 1965 was not particularly important in terms of changing farm worker labor supply. There was no singular ‘event’ that clearly marked the pre- and post-periods, although the end of the bracero program in 1965 did cause bracero labor to go to, and remain at, zero. Instead, as the data reveal, there were many large changes in bracero labor that may be considered ‘events.’ These events were mostly ignored. As already noted, treatment was defined by the bracero share of 1955 peak-season labor. However, wages were measured quarterly and in the DiD analysis the average wage in the pre-1965 (1962) and post-1965 (1962) are the key values. There is no reason to expect these average wages, particularly the wage during the period from 1942 to 1964 (1962), which includes huge changes in bracero labor, to be correlated with the 1955 bracero share of peak season farm labor.

Moreover, if changes in bracero labor during the pre-period had any effects, such as altering firm production technology as argued by CLP, then these effects would likely persist and affect demand and supply of labor in other years. For example, the large increase in bracero labor between 1948 and 1955 may have

---

4. Much of the increase in bracero labor is likely to have been replacement of undocumented Mexican labor (Bureau of Employment Security 1961). Thus, there may have been little change in labor supply between 1948 and 1955. The large substitution of legal for undocumented labor also suggests that the end of the bracero program would increase illegal immigration again. See below for evidence.
delayed technological adoption. If so, then the wages in the pre-period were already affected and the difference in wages between the pre- and post-period do not represent solely the effect of the end of the bracero program, or the effect of a change in foreign worker labor supply.

**Mis-measured wages**

The wage used in CLP is problematic for several reasons. The distribution of bracero labor was not evenly distributed across geography of farms within a state. Bracero labor worked at only a small fraction of farms—mostly large farms. In 1960, approximately 70 percent of the farms that used bracero labor were in California and Texas. Only 11 percent of farms in the four top states for bracero labor, California, Texas, Arizona and New Mexico, used bracero labor (Bureau of Economic Security, Hired Farm Workers 1961). So, most farms, even in states with a high share of bracero labor such as California, used no bracero labor. And the farms that used bracero labor were different than the farms that did not. In 1960, Texas employed the largest number of bracero laborers and almost all of them were in cotton. California was the second greatest user of bracero labor and most bracero labor worked on farms growing lettuce, cucumbers, tomatoes and citrus.

This variation in the presence of bracero labor over the year and the variation in the types (e.g., crops) and locations of farms that used bracero labor seriously affects CLP’s analysis because they used wages reported on all farms and for all hired labor. To measure wages, CLP used an average of wages across all types of hired labor and across all types of farms (crops, size) in a state in a quarter.

Several problems arise with this choice. First, seasonal farm labor is only part of all hired farm labor. While various sources list different numbers, seasonal (<150 days per year) farm labor made up approximately one-half of all hired farm labor in 1955 and in 1960. It is likely that regular (>150 days) farm laborers earned a different wage than seasonal farm workers, particularly because approximately half of all seasonal farm labor worked 25 or fewer days per year, and seasonal farm labor was more likely to be paid a piece rate. Second, bracero labor in a state during the non-peak season was often a fraction (e.g., one-fifth) of its share in the peak season. Thus, there is little relationship between 1955 share of bracero labor in a state, which is the measure of treatment, and the amount of bracero labor in a year-quarter in which wages were measured. Third, as noted, bracero labor was concentrated on a small share of farms that were larger, produced specific crops (e.g., cotton in Texas) and that employed more hired labor. Wages at these farms/

---

5. These figures were calculated from figures in Bureau of Employment Security, Farm Labor Market Developments 1955; 1960; Bureau of Agricultural Economics, Farm Labor Report 1955; 1960.
crops differ from wages at other farms/crops.

A few figures provide some of the variation within state for the same crop and within state between crops. In October of 1960, the piece rate for picking tomatoes in Alameda County, California was between $0.17 and $0.20 per 50-lb. box. The wage rate for picking tomatoes in the San Joaquin Valley, California in October of the same year was $0.15 per 50-lb. box. In the San Joaquin Valley, the piece rate for fig harvest was $0.25 per 40-lb box. In November of 1960 in Texas, the piece rate for picking cotton in East Texas was $2.50 per cwt., whereas in the same month and year, the piece rate for cotton in the High Rolling Plains region of Texas was between $1.50 and $1.75 per cwt. Cabbage pickers in Texas in October of 1960 received $0.50 per hour. As these illustrative figures suggest, wages varied significantly within a state for the same crop and within a state between crops. There was also variation in whether a laborer was paid an hourly wage or piece rate by crop and state. Given this variation in wages, the widespread use of piece rates, and the fact that bracero laborers worked on a small fraction of farms, the average wage for all hired farm laborers in a state in a quarter is apt to be a quite inaccurate measure of the wage for seasonal farm laborers in the same labor market as braceros.

Finally, it is also the case that the wage data is quite incomplete and crude as the following statement indicates:

State agencies affiliated with the Bureau of Employment Security are required to delineate crop-wage areas and survey a sample of farms and workers to find the prevailing wages paid to domestic workers in each of the activities at which Mexican nationals are employed. The findings—in the form of single rates, ranges, or schedules—are considered by the Department of Labor in making determinations. Since they relate only to prevailing wages, they do not necessarily include all rates found to be paid in the area. (Bureau of Economic Security, Farm Labor Market Developments July 1960, 17, my emphasis)

Overall, the average, composite wage measure used by CLP (2018) as a dependent variable is unlikely to reflect accurately wages paid to seasonal, farm laborers that compete with bracero labor. Here too, the implication is that it would have been surprising to find anything but no effect of ending the bracero program on domestic farm workers’ wages.

---

6. The wage figures are from various issues of Bureau of Economic Security, Farm Labor Market Developments.
Do CLP’s robustness analyses address these points?

CLP conducted several sensitivity analyses. However, none uses a different measure of treatment/exposure and only one uses a different measure of wages, but the same measure of exposure. In addition, the different measure of wages is still not what is conceptually correct—the wage earned by farm workers competing (crop, geography, firm type) with bracero workers. Similarly, adding state-specific trends, as CLP (2018) do in some sensitivity analyses, and using the same measure of exposure is not, in my view, a compelling response to the problems in the measurement of exposure. Controlling for potential state-specific trends accounts for the correlation between those trends and the (mis-measured) exposure variable, and those trends and (mis-measured) wages. These changes do not address the issues raised above. Finally, the analyses that regress measured wages on the natural logarithm of the stock of bracero workers are problematic in three ways: first, as noted wages are poorly measured; second, including a measure of labor supply on the right side of a wage (price) regression is econometrically problematic ( simultaneity bias); and third, there is no justification for using the natural logarithm of bracero labor. Overall, while the robustness analyses seem thorough, they do not address the underlying problems I have noted and mostly proceed along the lines of the original analysis. Not surprisingly, the sensitivity analyses find a null effect.

Substitution between illegal and legal Mexican labor

The effect of changes in the amount of bracero labor admitted to the United States on the domestic, farm laborer market would be moderated if there was a coincident, but opposite, change in illegal Mexican farm labor. CLP suggest this is not the case and one of their pieces of evidence is border apprehensions: “And border apprehensions of Mexicans did not substantially rise in the years immediately after exclusion, while measured border enforcement effort did not fall” (CLP 2018, 14).

Data from the period do not support the assertion (Viallet and McClure 1980). The data shown in Figures 1 and 2 show that the approximate doubling of bracero labor admitted from 1953 to 1956 coincided with a huge drop in alien apprehensions. Also evident is the substantial rise in alien apprehensions in the post-bracero period of 1965 to 1970. There was a 26 percent increase in alien apprehensions in 1965, and alien apprehensions rose significantly in the years fol-

---

7. Figure A4 in CLP’s appendix (2018, A-26) is similar to Figures 1 and 2 with one important difference. CLP use different Y axis scales, and those scales are misleading suggesting that alien apprehensions were much lower vis-à-vis the quantity of bracero labor.
Figure 1. Alien apprehensions and braceros admitted, 1948–1970

![Alien apprehensions and braceros admitted, 1948–1970](image1)


Figure 2. Alien apprehensions and braceros admitted, 1960–1970

![Alien apprehensions and braceros admitted, 1960–1970](image2)


ollowing the end of bracero program. The average annual increase in apprehensions was 22 percent between 1964 and 1970. By 1970, apprehensions were the equivalent of bracero stock in 1960—full replacement by 1970. These figures suggest strongly that there was a non-trivial substitution of legal for illegal (and reverse) Mexican labor during the entire period of analysis in CLP (2018). These flows of legal and illegal Mexican farm laborers suggest that the beginning and end of the
bracero program would have little effect on wages because there was little change in actual labor supply.

**How likely is CLP’s explanation that technological progress mediated effects of the end of the bracero program?**

CLP argue that the absence of an effect of the end of the bracero program on wages is consistent with theoretical models that allow for firm responses. CLP review these models, and describe how allowing for capital, technology, and output responses to the decrease in labor supply moderates the positive effect of the labor supply decrease on wages. Most of these models still predict that a decrease in labor supply will raise wages, although by varying amounts. There is one exception, and this is the model in which there is a cone of diversification. In this case, firms use different technologies that are more or less labor-intensive, but despite this heterogeneity wages are constant across firms. A decrease in labor supply has no effect on wages in an industry that is characterized by a cone of diversification (although it would still affect the quantity of labor employed). The decrease in supply causes more firms to adopt the less labor-intensive technology. CLP’s overall point, however, is that a decrease in labor supply such as that brought about by the ending of the bracero guest worker program may have a relatively minor effect on wages of domestic workers because firms adopt labor-saving technology.

The foregoing evaluation of the data and CLP’s empirical analysis raises doubts about the validity of their conclusion that ending the bracero program had no effect on wages of farm laborers, and therefore, tends to render moot the analysis that seeks to explain the null finding. Nevertheless, I assess some of the evidence available and some of CLP’s evidence on whether the bracero program resulted in a significant change in farm technology.

The first piece of evidence relates to farm labor productivity. There is overwhelming evidence that technological change was substantial and ongoing long before the end of the bracero program, and that this technological change was drastically reducing the demand for farm labor. Between 1940 and 1960, farm output per man-hour worked nearly doubled every ten years (Bureau of Economic Security, Hired Farm Workers 1961). Changes in the nature of farms coincided with this rise in productivity. Between 1950 and 1970, the average size (acres) of a farm increased by almost 100 percent, the number of farms decreased by approximately 50 percent, and farms became more specialized producing fewer crops (Dimitri et al. 2005).
An important consequence of these changes in farming was that there was an increase in the potential pool of hired farm labor, as small, owner-operators of farms and their family moved into the hired farm labor market. Between 1950 and 1960, the number of family farm workers declined by approximately 20 percent (Bureau of Economic Security, Hired Farm Workers 1961). This is a large decrease relative to hired farm labor because family farm workers as a group were approximately three times the size of hired farm labor (Bureau of Agricultural Economics 1960). So, the 20 percent decline in family farm labor represents an increase in the potential pool of hired labor of 60 percent. This trend in family farm labor is unlikely to be the same across states and crops and is likely to be correlated with the use of bracero labor that also differs by state and crop. This large, state-specific, time-varying change in labor supply would create difficulties for an analysis of the effect of the bracero program on the farm labor market, and CLP do not address the issue. Note too, that the analyses in CLP (2018) that include state-specific trends do not address this issue because of the way treatment is measured.

The change in farm technology and decrease in demand for farm labor were particularly important in cotton and for bracero labor. Cotton was also, by far, the single largest sector using Mexican labor (Bureau of Economic Security, Hired Farm Workers 1961). For example, in November of 1960:

Practically all of the 152,000 Mexican nationals at work in mid-November were in Texas, California, Arizona, Arkansas and New Mexico. About three-fifths of them were harvesting cotton. … Mexican-worker employment declined about 22,000 from the November 1959 level largely because the increased use of cotton-harvest machinery had reduced the need for hand harvesters. (Bureau of Economic Security, Farm Labor Market Developments, July 1960, 5)

Another report from the period says:

In 1958, when 34 percent of the cotton crop was machine-harvested, 627,000 seasonal workers (455,000 domestic and 172,000 foreign) were employed at the peak of harvest. In 1963, 72 percent of the crop was machine-harvested and 366,000 workers (350,000 domestic and 16,000 foreign) were employed at the peak. This change amounted to a decline of 261,000 (105,000 domestic and 156,000 foreign) for an annual decline of about 52,000. Foreign workers used in 1963 were less than 10 percent of the number employed 5 years earlier. (McElroy and Gavett 1965, 21)

The concentration of bracero labor in cotton and the rapid technological change in cotton that occurred well before the end of the bracero program suggests that
technological change by firms was not an important explanation of the absence of an adverse wage effect from ending the bracero program. Indeed, the mechanization in cotton may have been an important cause of the decline in bracero labor. As noted earlier, the bracero share of seasonal farm labor in Texas went from 24 percent in 1955, to 18 percent in 1960 and finally to 7 percent in 1963.

CLP used the tomato crop in California as an example of how technology changed with the end of the bracero program. However, in 1960, a relatively small share or bracero labor was employed harvesting tomatoes. Cotton employed three times as much bracero labor (a figure that was already lowered by technological change; see above). Also, bracero labor made up a larger share of seasonal farm labor in the lettuce and cucumber crops combined than in tomatoes (Bureau of Employment Security, Hired Farm Workers, 1960).\(^8\) In 1963, during the peak harvest season, half of all bracero laborers in California were harvesting tomatoes. This concentration of bracero labor in one crop underscores earlier points about the inadequacy of using an average wage in a state to measure the effect of ending the bracero program (McElroy and Gavett 1965). Moreover, as noted, relatively few firms employed bracero labor even in California tomato farms. Finally, there was evidence of rapid technological change in tomato harvesting prior to the end of the bracero program:

California produces over half of the U.S. tomato crop, about 60 percent of which is used for processing. In 1963, when about 34,000 braceros were employed in the California tomato harvest, 25 machines harvested about 5 percent of the State’s processing tomato crop. In 1964, there were about 100 machines in operation, harvesting about 20 percent of the processing crop. As more machines become available and the quality of the mechanizable-processing tomato is improved, machine harvesting will expand; but harvest labor will continue to be needed for some processing tomatoes and for all fresh-market tomatoes, which cannot now be machine-harvested. (McElroy and Gavett 1965, 18–19)\(^9\)

Other than the example of the tomato crop in California, CLP provide little other evidence to support the claim that the absence of a wage effect from ending the bracero program was due to technological change.

---

8. As described earlier, the number and share of bracero labor in a state, year and crop changed significantly over time, so it is difficult to characterize the states and crops that relied on bracero labor.
9. This information on changes in harvesting technology differs from that used by CLP in Figure 5. Their data show much less pre-1965 changes in technology.
Conclusion

The conversation of whether immigration policies can be used to affect the wages and employment of domestic workers is longstanding and controversial. The controversy extends to the scientific community, as evidenced by the continued analysis and re-analysis of the effect of the Mariel Boatlift on the Miami labor market (Card 1990; Borjas 2017; Clemens and Hunt 2017; Peri and Yasenov 2019). The political and scientific prominence of this research question, and the historical prominence of the bracero guest worker program, makes CLP (2018) notable. The conclusions of their study may have an impact on both policy and theory.

I have argued that there are significant problems in the data and analyses used by CLP, and that the problems are such that it would be surprising if such analysis were to find anything but no effect of ending the bracero program. I also show that there is a strong likelihood that illegal migration via Mexico, a dramatically rising pool of domestic labor, and ongoing (exogenous) technological change were much more important influences on wages and employment in the farm labor market, and that these forces made it highly unlikely that ending the bracero program would have a measurable effect.

References


Robert Kaestner is a Research Professor at the Harris School of Public Policy of the University of Chicago. He is also a Research Associate of the National Bureau of Economic Research, an Affiliated Scholar of the Urban Institute and a Senior Fellow of the Schaeffer Center for Health Policy of USC. Prior to joining Harris, Kaestner was on the faculty of the University of Illinois, University of Illinois at Chicago, University of California, Riverside, the CUNY Graduate Center and Baruch College (CUNY). He received his Ph.D. in Economics from the City University of New York. He received his BA and MA from Binghamton University (SUNY). His research interests include health, demography, labor, and social policy evaluation. He has published over 125 articles in academic journals. Recent studies have been awarded Article of the Year by AcademyHealth in 2011 and the 2012 Frank R. Breul Memorial Prize for the best publication in Social Services Review. Dr. Kaestner has also been the Principal Investigator on several NIH grants focused on Medicare and Medicaid policy. Kaestner is an Associate Editor of the Journal of Health Economics and the American Journal of Health Economics, and on the Editorial Board of Demography and Journal of Policy Analysis & Management. His email is kaestner.robert@gmail.com.


About the Author

Clemens, Lewis, and Postel's reply to this article
Go to archive of Comments section
Go to March 2020 issue

Discuss this article at JournalTalk:
https://journaltalk.net/articles/5996/
Comment on Kaestner, “Revisiting the Bracero Guest Worker Reforms”

Michael A. Clemens¹, Ethan G. Lewis², and Hannah M. Postel³

LINK TO ABSTRACT

In Clemens, Lewis, and Postel (2018), we evaluate one of the largest active labor market policy interventions of its kind: the federal government decision to bar almost half a million Mexican bracero workers from the U.S. labor market in the 1960s. We—henceforth CLP—found no evidence that this policy intervention raised wages or employment for U.S. workers. Robert Kaestner (2020) likewise presents no evidence that this policy raised wages or employment for U.S. workers. Apparently there is consensus that no such evidence exists.

What Kaestner does instead is to state three speculations.

Robustness

First, Kaestner speculates that the results in CLP (2018) might not be robust to different definitions of the policy treatment. He conjectures that the empirical result might change if exposure to bracero exclusion were considered to begin in 1962 rather than 1965, and that the result might change if the measure of exposure used a different pre-exclusion year or month. He does not actually conduct any such tests.

These claims are incomprehensible, given that the original CLP paper already

¹. Center for Global Development, Washington, DC 20036.
². Dartmouth College, Hanover, NH 03755.
³. Graduate student, Princeton University, Princeton, NJ 08544.
tests and rejects them. The original paper discusses how “[t]he Kennedy administration began the process of bracero exclusion in March 1962” (CLP 2018, 1469). It discusses how that caused a pre-1963 decline in bracero exposure, which is prominent and extremely clear in the original paper’s Figure 2 Panel A (ibid., 1476). That is why the original paper exhaustively shows that the assumption of treatment in 1965 does not matter:

1. It re-runs the difference-in-differences regressions with 1962 as the treatment year (CLP 2018, A-24–A-25 Tables A7 and A8).
2. It reports full event-study regressions, with a separate coefficient on each year so that the interested reader can choose any year as their preferred treatment year (ibid., A-20–A-21 Figures A1 and A2).
3. It reports fixed-effects regressions with no assumption of a before-or-after period (ibid., 1480 Figure 4, A-18 Tables A3 and A4).

For even greater transparency, the original paper’s Figure A3 (CLP 2018, A-22) graphs the raw data, making it as clear as can be that major shocks happened in 1962 and 1965, not before that, and that the choice of peak month is not relevant. Parts of that existing figure, for the most important states, are reproduced here in Figure 1.

Figure 1. Detail of Figure A3 in CLP (2018, A-22)

(a) Texas

(b) California

These raw data show a large, sudden disappearance of braceros in 1962 and another in 1965. They show no substantial change in the hiring of U.S. workers—in any month of the year, relative to any pre-exclusion base year. This result is obviously not sensitive to the precise choice of pre-treatment year or month as the measure of exposure. These and dozens of other robustness checks have been publicly available in the CLP appendix since April 2017, over a year before the

paper was published.

CLP (2018) not only posted all of their data and replication code online but went further to post scans of all of their primary archival sources online (link). If one has concerns about the robustness of the CLP results to changes of specification, those concerns can be tested with trivial effort. Kaestner presents no reanalysis of any kind.

**Mechanism**

Second, Kaestner conjectures that the mechanism of CLP’s result might not be induced technical advance. He states that he does not find that mechanism plausible (Kaestner 2020, 5). The fruitful course for a researcher with concerns about the “plausibility” of this mechanism would be to gather data to test their hypothesis and conduct such a test.

That is what Shmuel San (2019) does in striking new research. Using bibliometric analysis of patent filings before and after bracero exclusion, San shows that immediately after exclusion there was a large surge of innovation in production technology for crops that had depended heavily on bracero labor—but not for other crops. This strongly rejects the hypothesis that bracero exclusion had no effect on labor-saving innovation, corroborating the mechanism for which CLP offered only, as they described it, “suggestive” evidence (2018, 1483).

The hard work of gathering historical data advances economists’ knowledge of historical events. Speculation about “plausibility” does not.

**Unauthorized migration**

Finally, Kaestner speculates that if the excluded bracero workers had been immediately replaced by a vast wave of unauthorized migrant workers from Mexico, that would offer a different mechanism for the null effects on U.S. workers’ wages and employment. This is hypothetically true but would require a massive historical event that did not occur and for which Kaestner offers no evidence.

As CLP (2018) discuss at length, 99.87 percent of the last bracero workers in 1964 were recorded returning to Mexico, at a time when unauthorized migration from Mexico was close to zero. The pre-exclusion braceros did not simply stay on illegally in the United States; they went home. That means that if the excluded braceros were replaced by unauthorized workers in 1965, they would have to cross the border anew.
Evidence against immediate replacement of the braceros

In other words, immediate replacement of the braceros by unauthorized migrants would require a new, sudden flood of approximately 100,000 new unauthorized migrants crossing the border immediately in 1962 and every year thereafter—plus an additional flood of 200,000 more, immediately in 1965 and every year thereafter. This is the event that Kaestner would need to document to support his claim.

That event is imaginary. Researchers who wish to posit such a historical event would need to provide an explanation for all of the following:

- Stanford historian Ana Raquel Minian (2018, 25) documents from primary sources in Mexico that most of the former braceros remained in Mexico for several years: “A quarter million braceros who had been in the United States flooded into Mexican border cities,” she writes of 1965. “Many of them elected to remain in the northern borderlands instead of returning to their hometowns in the interior of the country because they believed that the program would be renewed.” In those areas, she shows, “Unemployment rates mushroomed…reaching almost 50 percent of the population.” This is inexplicable if unauthorized migrants immediately replaced the braceros.

- John McBride (1963) narrates, firsthand and in detail, the elimination of the bracero workers from cotton farms in the Lower Rio Grande Valley of Texas—directly on the Mexican border. He does not mention unauthorized Mexican workers replacing the braceros at all. This is inexplicable if hundreds of thousands had suddenly began pouring over the border.

- The U.S. Employment Service published contemporary investigations of how farms were adapting to bracero exclusion in Texas (Hood 1966) and Michigan (Mitchell 1966). Neither of them mentions illegal migration in any form.

- The U.S. Secretary of Labor (1966) reports a detailed contemporaneous investigation into how farmers nationwide were adjusting to bracero exclusion. It does not mention illegal migration at all. It does report widespread crop losses, especially in California where bracero employment was highest.

- The U.S. Department of Agriculture conducted a nationwide, detailed review of the response of farms to bracero exclusion, state by state and crop by crop (Metzler et al. 1967). It reports systematic labor shortages and massive in-state and nationwide efforts to recruit teenagers or
indigenous people to replace the braceros. It reports that growers typically were not able “to recruit a labor force which would take over the jobs formerly performed by the braceros.” It does not mention illegal migration at all, except to note that there was little of it: In Texas, it says, “Both legal and illegal immigration from Mexico has been reduced to a minimum” (Metzler et al. 1967, 14). A typical note reports: “When no braceros were available in 1965, the cucumber growers had great difficulty in obtaining a labor supply” (ibid., 15). None of these statements would make sense if growers had quickly replaced the bracero workers.

• Massey and Pren’s (2012) bibliometric analysis of U.S. newspapers reveals no substantial rise in mentions of illegal migration until just before 1970. This accords with our Figure 2 here. When illegal migration became substantial later on, in the 1970s, it was widely discussed in newspapers.

In other words, if the major historical event posited by Kaestner occurred, it went somehow unnoticed by officials of the USDA, Employment Service, and Department of Labor charged with studying the issue at the time, and unnoticed by journalists at the time, and unnoticed by Texas cotton farmers at the time, and it has been somehow overlooked by historians since. This possibility is remote. All of this is already discussed in CLP (2018).

Kaestner presents no new evidence for his claim that this vast change occurred. He simply reproduces the same graph that is already in the original paper’s appendix (CLP 2018, A-26 Figure A4). That graph from the CLP paper already shows that illegal migration rose after bracero exclusion. What is at issue is whether it rose so quickly and so greatly that labor markets could not react. CLP write that “border apprehensions of Mexicans did not substantially rise in the years immediately after exclusion” (2018, 1481, emphasis added).

The slow rise in illegal migration in the 1960s

As the CLP appendix also already shows, the rise in illegal migration remained low for several years after the 1962 and 1965 exclusions of braceros. Labor markets had plenty of time to react.

This is apparent in Figure 2 below. It uses the same data in the CLP appendix and in the comment. The vertical axis is the rise in apprehensions of Mexicans as a fraction of the 1953 level (835,311), and this rise is shown starting from 1960. Three years after the early-1962 exclusion, at the end of 1964, apprehensions at the border had risen by just 2.6 percent of their 1953 level. In 1969, seven years
after exclusion began in 1962—and four years after exclusion was completed in 1965—apprehensions had risen by just 10.8 percent of their 1953 level.

**Figure 2.** Analysis of the existing Figure A4 in CLP (2018, A-26): The rise of illegal migration was small and slow

![Graph showing the rise in apprehensions](image)

1953 is the best comparison year because it is the last full year before the bracero program expanded to almost completely replace the market for unauthorized Mexican migrants (Hernández 2010, 187–191). Hypothetically, the same number of observed apprehensions could understate a larger number of unobserved unauthorized migrants if there were a reason to believe that enforcement effort fell greatly after 1962 or 1965, relative to 1953. But there is no such evidence. As the CLP appendix already discusses, there was no decline in the government’s own metrics of border enforcement staffing and enforcement effort in 1965 or several years thereafter (North and Houstoun 1976, 53). Any given migrant’s probability of apprehension was no lower after exclusion than in 1953: In 1967 it was 52 percent, compared to 49 percent in 1953 (Roberts et al. 2013, Appendix 1 p. 7).

The intent of the exclusion policy was to raise wages immediately, as the simplest economic theory predicts it would, not a decade later when the magnitude

---

5. Enforcement effort was clearly lower in 1965 than in the final year before bracero visa expansion, 1954—because 1954 was the year of the massive crackdown officially named *Operation Wetback*. This is why we choose 1953 as the comparison year. 1953 is the ideal comparison because, although a similar border-enforcement crackdown had been planned for that year (to be called *Operation Cloud Burst*), it was never actually carried out (Hernández 2010, 183).
of illegal migration became large enough to be an important confounder. When the U.S. Senate (1966, 16–17) Committee on Labor claimed that bracero exclusion had successfully raised farm wages, it used data extending only through October of 1965—just one season after exclusion was complete.

Implications for CLP’s analysis

Beyond ignoring the magnitude of illegal migration, Kaestner ignores major anomalies—also already discussed in CLP—that would need to be explained if the braceros were substantially and immediately replaced by unobserved workers (CLP 2018, A-23–A-26). One is that estimates of “total hired workers” that include unauthorized workers, available for most states, fall proportionately with the decline in braceros (ibid., A-25 Table A9). There would be no such change if the braceros were quickly replaced.

Another anomaly is that there is no correlation between bracero exposure intensity and U.S. workers’ wages or employment controlling for state fixed effects, prior to exclusion, during the bracero program (CLP 2018, 1480 Figure 4, A-18 Tables A3 and A4). Suppose that we accept the comment’s assertion that bracero scarcity after 1965 was uncorrelated with wages because unobserved unauthorized workers had instantaneously replaced the braceros. How, then, can we explain the fact that very high variance in bracero scarcity was uncorrelated with wages in the 1950s as well, when (as the comment accepts) unauthorized immigration was nearly eliminated? CLP already ask this. Kaestner is silent.

The role of policy evaluation

We conclude with a broader observation about this comment. When the government intervenes in the labor market to raise wages or employment (such as with a job-training program), economists do not presume that the intervention worked until someone can prove a precisely estimated zero effect. They require evidence of a nonzero effect (Card et al. 2017). When the government intervened to exclude the braceros, its explicit and principal goal was to raise wages and employment for U.S. farm workers (Borjas and Katz 2007). Such an effect must be shown with evidence. Kaestner provides none.

CLP use the government’s own data on wages and employment, precisely the data that the government claimed would show the positive effects of its policy intervention (U.S. Senate 1966, 16–17). Using those data is an appropriate test of the government’s claims about that policy. Kaestner dismisses the data as containing no useful information, but does not suggest any other way of testing
the hypothesis. This amounts to the claim that no such tests exist—that no one can ever know the economic effects of this policy. We look forward to seeing more fruitful approaches in this literature that, in contrast, advance economists’ understanding of historical events.

References


Michael Clemens is an economist at the Center for Global Development and IZA Institute of Labor Economics. He has published on migration, development, economic history, and impact evaluation in journals including the American Economic Review, and his research has been awarded the Royal Economic Society Prize. He has served as an Associate Editor of the Journal of Population Economics and World Development. He received his Ph.D. from the Department of Economics at Harvard University, specializing in economic development, public finance, and economic history. His email address is mclemens@cgdev.org.

Ethan Lewis is a Professor of Economics at Dartmouth College and a Research Associate with the National Bureau of Economic Research. His research investigates how U.S. labor markets adjust to immigration and to technological change, including how employers adapt their production technology to the local availability of immigrant workers. His newest work investigates the impact of U.S. immigration policies, including the Reagan amnesty. His findings have appeared in top economics journals, including the American Economic Review, the Quarterly Journal of Economics, the Journal of Political Economy, the American Economic Journal, and the Review of Economics and Statistics. His work has been covered in the New York Times, Wall Street Journal, The Economist, Newsweek and other media outlets. Ethan received his Ph.D. in economics from UC Berkeley in 2003. His email address is ethan.g.lewis@dartmouth.edu.

About the Authors
Hannah Postel is a Ph.D. candidate in Demography and Social Policy at Princeton University under a Graduate Fellowship from the National Science Foundation, and a Research Associate of the Overseas Development Institute. Prior to entering Princeton she built the first quantitative study of Chinese migration to Zambia under a Fulbright research grant. Her email address is hpostel@princeton.edu.

Discuss this article at Journaltalk: https://journaltalk.net/articles/5997/
Brought Into the Open: How the U.S. Compares to Other Countries in the Rate of Public Mass Shooters

John R. Lott, Jr.\(^1\) and Carlisle E. Moody\(^2\)

Adam Lankford (2016) asserted that the United States accounted for 31 percent of the world's public mass shooters over the 47 years from 1966 to 2012. The news media around the globe widely publicized Lankford’s claim as soon as he started circulating his unpublished paper in 2015. Yet, despite numerous requests from researchers and the news media over four years, Lankford refused to provide a list of his cases or explain how he compiled them (see Lott 2018b). In responding to our research (Lott and Moody 2019), Lankford (2019) finally provided an appendix listing the 292 cases upon which he says he based his 2016 article.\(^3\) The extreme difference between his findings and ours, we now know, is driven by Lankford not following the definitions that he says that he was using. While we are still missing the data for the regressions that he ran for his 2016 paper, we at least now know what cases his sample included and excluded.

Lankford (2016, 190–191) claimed that he followed the FBI, Department of Homeland Security, and NYPD traditional definition of public mass shootings, but we discover otherwise. He included cases for the United States that do not fit those definitions, and he excluded hundreds of cases from around the world that

---

1. Crime Prevention Research Center, Swarthmore, PA 19081.
2. College of William and Mary, Williamsburg, VA 23187.
3. Lott (2018a) had criticized Lankford, finding that Lankford must have very vastly undercounted public mass shootings in other countries. The difference was extreme. Carl Moody joined Lott to author Lott and Moody (2019).
Both errors greatly exaggerate the United States’ share of these attackers.

We also discovered that Lankford only included cases with just one shooter, except when he includes cases with two shooters. The only case involving two shooters in the United States that he counted was the 1999 Columbine attack, and the only such case outside the United States was from Russia.

Unlike Lankford, we immediately provided as part of Lott and Moody (2019) our entire list of public mass shooters as well as the news stories and sources that we relied on to put the list together. Even if Lankford thought there was a justification for studying only attacks with one or two shooters, it would have been easy to go through our list and explain why his list of such cases differed from ours. For example, Lankford excludes, without explanation, 37 foreign public mass shootings involving just one shooter and another 40 foreign cases involving two shooters. Furthermore, he does not justify the additional cases for the United States that he included that do not fit the FBI, Department of Homeland Security, and NYPD definition of public mass shootings. Both errors greatly exaggerate the United States’ share of these attackers.

We took care to exhibit, at length, official definitions (Lott and Moody 2019, 41–42). Nowhere do any of those sources confine the definition of public mass shootings under examination to cases with just one shooter. Indeed, the NYPD included cases involving up to ten shooters. Lankford’s response to our extensive demonstration of official definitions is to ignore that demonstration.

In his original paper, Lankford states, “For this study, attackers who struck outdoors were included; attackers who committed sponsored acts of genocide or terrorism were not. This is consistent with the criteria by the Federal Bureau of Investigation (FBI) in its 2014 active shooter report” (2016, 190–191). Nowhere in Lankford’s original paper does Lankford mention that he limited the attacks to one or sometimes two shooters. We invite readers to obtain a PDF of Lankford (2016) and search on ‘one,’ ‘alone,’ ‘lone,’ ‘solo,’ and ‘single’ to confirm that he nowhere reveals that he has confined his definition to cases with one shooter (except when he includes two). The first example that he provides of a public mass shooting on the first page of his paper is an exception—the Columbine attack, which had multiple shooters—so that example especially obscures that, aside from Columbine and a Russian case, his list is confined to cases of lone shooters. A more complete discussion of Lankford’s decision to include attacks with one or two shooters is provided below.

Finally, we discuss whether Lankford excluded terrorism cases as Glenn Kessler (2018) guessed he did, and we point out that even if all terrorism cases from outside the United States are excluded while those in the United States are included, the United States would account for less than 6 percent of the world’s public mass shooters—less than one-fifth of the rate that Lankford claims.
Lankford’s dataset

The dataset that Lankford provided in 2019 does not fit statements in his original paper. He says the NYPD dataset “may be nearly comprehensive in its coverage of recent decades, [though] it may be missing some older cases” (Lankford 2016, 191). Given that the NYPD found only 16 attacks with 27 killers outside the United States over the period that we examined, 1998 to 2012, it is clear that Lankford’s list comes nowhere near to being comprehensive. If foreign shooter cases where more than one shooter was involved are excluded from the NYPD dataset, there would be only 15 killers left, far too few for Lankford’s now-public “nearly comprehensive” dataset of 98 mass shooting incidents. Lankford writes: “All efforts were made to ensure that the same data collection methodology employed by the NYPD was used to obtain this information” (2016, 191). Yet, despite claiming that his dataset is consistent with the NYPD methodology, Lankford now claims for the first time that his dataset was limited to cases with just one and sometimes two shooters.

Perusing his dataset, we find that he does not limit the cases consistently. Lankford includes the 1999 Columbine case where two shooters killed 13 people. This deviation from the one-shooter requirement again raises the issue of how exactly the category is defined by Lankford.

Meanwhile, Lankford excludes the 1998 Jonesboro, Arkansas shooting where two shooters opened fire on their classmates with stolen firearms. There is no explanation for this exclusion. He also has one case from Russia with two shooters, but he excludes 40 other foreign incidents with two shooters that fit the FBI and NYPD definitions. Why are those 40 other incidents excluded?

There is no discussion whatever concerning why he includes cases with one or two shooters but excludes examples with three shooters. What is the reason for including a case of two killers working together but not three?

Again, Lankford excludes 37 foreign public mass shootings involving just one shooter, despite us providing him an entire list of our cases. Lankford does not discuss why he excluded these single-shooter cases. In seven of these cases the killer was a member of some group, but even in those cases there is no evidence that any of them were somehow not self-initiated.

Lankford also inflates the number of U.S. cases during the 1998–2012 period by including nine cases that don’t meet the FBI or NYPD definitions of public mass shootings. He includes cases that involve another crime such as a robbery, or that occur in a non-public place such as a residence, or that have fewer than four killed in a single place (see Table 1).
TABLE 1. United States cases that Lankford included that don’t meet his definition

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
<th>Reason it should be excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 20, 2000</td>
<td>Harris</td>
<td>Robbery related</td>
</tr>
<tr>
<td>January 9, 2001</td>
<td>Park</td>
<td>Fewer than four killed in the same place excluding the attacker: Three killed at the business, and the other one later found slain at the shooter’s store</td>
</tr>
<tr>
<td>September 9, 2001</td>
<td>Ferguson</td>
<td>Fewer than four killed in the same place: Two bodies found at a Sacramento city equipment lot, two bodies found at a city marina, the last one killed at the victim’s home</td>
</tr>
<tr>
<td>December 9, 2007</td>
<td>Murray</td>
<td>Fewer than four killed in the same place: Two killed at the Youth With A Mission training center, and two killed at the New Life Church</td>
</tr>
<tr>
<td>September 2, 2008</td>
<td>Zamora</td>
<td>Fewer than four killed in public</td>
</tr>
<tr>
<td>March 10, 2009</td>
<td>McLendon</td>
<td>Fewer than four killed in public</td>
</tr>
<tr>
<td>March 21, 2009</td>
<td>Mixon</td>
<td>Fewer than four killed in the same place: Two killed during a routine traffic stop and two killed near apartment of his sister</td>
</tr>
<tr>
<td>August 14, 2010</td>
<td>McCray</td>
<td>The killer is an ex-gang member. The case could be gang-related.</td>
</tr>
<tr>
<td>August 8, 2011</td>
<td>Hance</td>
<td>Fewer than four killed in public</td>
</tr>
</tbody>
</table>

On whether the United States is an outlier

Definitions are worth fighting over. But let us now consider an array of definitions and see what the numbers say.

We know that the U.S. is not an outlier in the number of public mass shootings using the FBI/NYPD definition, which does not limit the incidents to any particular number of shooters. In fact, despite having 4.5 percent of the world’s population, the U.S. share of the number of public mass shootings is 2.88 percent (Lott and Moody 2019, 53, Table 1).

Is the U.S. an outlier if we use the NYPD methodology that Lankford claimed to have used? Is the United States an outlier if we use Lankford’s definition of one or two shooters?

While the NYPD definition does not limit the number of shooters in any way, the NYPD dataset has a maximum of ten shooters. Therefore, we also investigate a definition that uses ten as a limit on the number of shooters. For the same reason, we limit Lankford-style incidents to the maximum number of shooters in his now-public dataset, namely two, since he includes the Columbine attack and a case from Russia with two shooters.

Lankford excludes all incidents involving an unknown number of shooters while chiding us for having missing values: “In fact, [Lott and Moody] admit that
they do not even know the number of shooters for all incidents they are counting” (Lankford 2019, 73). But the incidents happened, and people were killed even if the published reports did not reveal the number of perpetrators. By using only cases for which the number of perpetrators is known, Lankford substantially undercounts the number of incidents in foreign countries. This is not a problem if we are counting public mass shootings using the FBI/NYPD definitions which make no reference to the number of shooters. If we adopt the \( \leq 10 \) definition, we have 289 cases. However, adding those cases for which the number of shooters is missing, we have 1,341 cases. Lankford undercounts the number of foreign shootings by 1,052 for just the years from 1998–2012.

To satisfy the \( \leq 10 \) definition, we need to estimate the number of cases with missing values that would be likely to have more than 10 shooters. Of all the incidents where the number of shooters is known, 73 percent have 10 shooters or fewer. Therefore, we estimate that the number of cases with missing values that have 10 or fewer shooters is 0.73 \times 1,052 = 768. We assume that the distribution of cases with an unknown number of shooters is the same as the distribution of known shooters. We know the number of cases for which the number of shooters is 1, 2, \ldots, 10. We divide each one of these by 289 to get the percentage of cases with that number of shooters. Assuming the cases with missing values have the same percentages, we can estimate the total number of cases with 1, 2, \ldots, 10 shooters. For example, the number of non-U.S. single shooters in our sample is 98. The corresponding percentage of single shooters is 98 / 289 = 0.34. Therefore 0.34 \times 768 = 261 is our estimate of the number of single shooters from the cases with missing values. The estimate of all non-U.S. shooters who acted alone is 98 + 261 = 359. The number of cases with two shooters is 58, 58 / 289 = 0.2, 0.2 \times 768 = 154. The number of non-U.S. shooters who attacked in groups of one or two is therefore 359 + 58 + 2 \times 154 = 725. The number of shooters in groups of 3, 4, \ldots, 10 are calculated the same way. The results are presented in Table 2.

Using the \( \leq 10 \) definition, the United States is not an outlier in public mass shooters. We estimate that the United States has 1.25 percent of the number of all such shooters, while having 4.5 percent of the world’s population. Using only the cases where the number of shooters is known, the U.S. has 4.4 percent of the number of incidents where the known number of shooters is 10 or fewer (even though we are excluding many foreign cases with missing values, 73 percent of which can be expected to have ten or fewer shooters). Using a \( \leq 2 \) definition, the United States has less than six percent of the world’s shooters who attacked alone or with one other person. Using a slightly expanded version of Lankford’s definition, three or fewer shooters, yields an estimate of 4.45 percent of the world’s public mass shooters are American, less than America’s share of the world’s population.
### TABLE 2. Share of the U.S. in world public mass shootings

<table>
<thead>
<tr>
<th>Measure</th>
<th>U.S.</th>
<th>Rest of the world</th>
<th>Total</th>
<th>Percent U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1966–2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lankford (2016) findings</td>
<td>90</td>
<td>202</td>
<td>292</td>
<td>30.82%</td>
</tr>
<tr>
<td>1998–2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYPD definition, ≤ 10 shooters, estimated</td>
<td>45</td>
<td>3,565</td>
<td>3,610</td>
<td>1.25%</td>
</tr>
<tr>
<td>NYPD definition, ≤ 10 shooters, known</td>
<td>45</td>
<td>975</td>
<td>1,020</td>
<td>4.41%</td>
</tr>
<tr>
<td>NYPD definition, ≤ 3 shooters, estimated</td>
<td>45</td>
<td>967</td>
<td>1,020</td>
<td>4.45%</td>
</tr>
<tr>
<td>Lankford definition, ≤ 2 shooters, estimated</td>
<td>45</td>
<td>725</td>
<td>770</td>
<td>5.84%</td>
</tr>
<tr>
<td>Lankford definition, = 1 shooter, estimated</td>
<td>41</td>
<td>359</td>
<td>400</td>
<td>10.25%</td>
</tr>
<tr>
<td>Population (millions)</td>
<td>295.5</td>
<td>6,235.8</td>
<td>6,531.3</td>
<td>4.52%</td>
</tr>
</tbody>
</table>

**Note:** By ‘estimated’ we mean that the distribution of cases with an unknown number of shooters is assumed to be the same as the distribution of known shooters. We adjusted the number of incidents with missing values by the proportion of all cases where the number of shooters is known that have 10 or fewer shooters (73 percent).

All programs, data and results may be downloaded from the journal website ([link](#)).

---

### Limiting cases to where only one shooter was involved

According to Lankford,

Lott and Moody’s own data show that 29.7 percent of the entire world’s public mass shootings by single perpetrators were committed in the United States, and that America had more than six times its share of the world’s public mass shooters who attacked alone. This is remarkably similar to my original study’s published result: I found that 30.8 percent of public mass shooters attacked in the United States (Lankford 2016), which would also be more than six times our share of the world’s public mass shooters (30.8 / 4.5 = 6.8). (Lankford 2019, 75–76)

But Lankford omits all cases where the number of shooters is not known. If the incidents where the number of shooters is known is a good sample of all incidents, the U.S. has 10 percent of the world’s lone-wolf shooters, far short of Lankford’s 30 percent.

We don’t know why the U.S. has relatively more lone-wolf shooters, or why the rest of the world has fewer. We offered speculations on the matter (Lott and Moody 2019, 39, 46–49). The difference in the prevalence of lone-wolf and group shootings might be due to the fact that terrorist groups are more prevalent outside the United States. Lankford blames firearms. Our view is simply that the prevalence of lone-wolf shootings in the U.S. is a complex issue of culture and social alienation,
and it is irresponsible to jump to policy conclusions, especially because guns are also used for defense and protection.

Lott and Moody (2019), in line with the official definitions, included cases that involve multiple shooters, which are far more frequent in many other countries than in the United States. In the title of his response, Lankford (2019) claims “Confirmation That the United States Has Six Times Its Global Share of Public Mass Shooters.” But that is not what he claims in the body of the paper where he says: “the United States had more than six times its global share of public mass shooters who attacked alone” (2019, 73, our emphasis).

It is worth examining how Lankford, once pressed, responds to the challenge to his definition. In the first paragraph he speaks of “public mass shooters,” without defining the term (Lankford 2019, 69). In the next paragraph he addresses the meaning of the term. We quote the paragraph in full:

“They almost always attack alone. This is such common knowledge that I am surprised it requires any comment. Most laypeople already know this without my needing to say so, and certainly all researchers with experience in this area recognize this simple fact. It is one of the things that makes public mass shootings so terrifying: they are one of the most vivid demonstrations of just how much death and destruction a single person can cause on his own.” (Lankford 2019, 69)

Notice that Lankford still has not defined “public mass shooter.” In the next paragraph he refers to listings of cases in documents by the FBI, NYPD, and two other U.S.-based sources, saying that they show that “95–98% of these crimes are committed by solo perpetrators acting alone” (Lankford 2019, 70, our emphasis). And in the next paragraph he cites U.S. media stories to the same effect. Lankford has not in fact stated his definition of “public mass shooter.” He has worked from an idea of the kind of public mass shooter usually seen in the United States, and then imposed that idea in investigating “public mass shooters” globally.

So, when Lankford says on the first page of his response, “you have to know only one thing about this specific type of criminal…. They almost always attack alone. This is such common knowledge that I am surprised it requires any comment” (2019, 69), we reply: Well, Professor Lankford, public mass shooters almost always attack alone in the United States. In the rest of the world they attack both alone and in groups, sometimes in large groups.

Lankford says,

Lott and Moody (2019) lump seemingly everything into their list of incidents from other countries: attacks by militia groups, paramilitary fighters, terrorist cells, and more. They include the aforementioned 2004 Lord’s Resistance
Army attacks in Uganda, as well as hundreds of other acts of group violence. (Lankford 2019, 72)

Nowhere in the FBI definition of public mass shooters does it limit the number of shooters. If a large group of shooters committed a public mass shooting in the United States, would the FBI ignore it? Besides adhering to official definitions, we appeal to plain language: ‘shooter/shooting’ signifies people getting shot by gunfire; ‘mass’ signifies many victims; ‘public’ signifies in a public place. No part of the expression signifies a shooter working alone. Both plain language and official definitions support our definition of public mass shootings. If one wishes to confine investigation to cases with just one shooter, one should say so, but that was never done by Lankford (2016).

Lankford writes:

[M]ost other researchers have not included gang violence or other group violence in their studies: group behavior is so profoundly different from that of individuals. I tried to follow their lead by similarly applying consistent criteria to all cases worldwide, and therefore excluded gang violence, along with sponsored acts of terrorism or genocide that did not appear self-initiated by the perpetrator, because group behavior plays such an important causal role in those other types of crimes. (Lankford 2019, 71–72)

We also excluded gang violence—because the FBI excludes it—and other criminal activity such as bank robberies. Such actions are primarily motivated by the profits associated with illegal activity. As for “sponsored acts of terrorism or genocide,” we exclude incidents involving state actors. As we argued previously (Lott and Moody 2019, 43–44), to operationalize “sponsored acts of terrorism” would require defining sponsored—which Lankford never does—and having sufficient information to decide case by case, and that will often be very difficult for foreign cases and older cases, as reporting is often very scanty and only in foreign languages. Lankford needs to face up to the fact that events globally are far too challenging, informationally, to be handled with vague, impressionistic criteria about ascribed ‘self-initiation’ and the like.

We do not exclude incidents of public mass shooting just because we think we know the motivation of the shooter or shooters. The motivation of shooters in large groups may be different from the motivation of shooters we have seen in the United States so far, but Lankford never says in his 2016 paper that he is studying only single shooters who are exclusively inner-directed. We know of no way to determine whether the perpetrators in groups were ‘self-initiated’ or not. Presumably they joined the group knowing that they might be involved in a multiple-victim public shooting, because that is what the group does. Is that self-
or group-initiation? Lankford includes the 2009 Fort Hood case where a shooter killed 13 people. The shooter had been in extensive email contact with Anwar al-Awlaki, the al-Qaeda imam. Is that incident self-initiated?

We included incidents involving groups of shooters because we thought that his claim that the U.S. has 31 percent of “public mass shooters” included all public mass shooters. We hope that Lankford will agree that, counting all incidents that satisfy the FBI definition of public mass shootings, the U.S. has a very small share of public mass shooters as we showed in our previous analysis, i.e., less than three percent of the number of shooters, incidents, or people killed (Lott and Moody 2019, 53).

We now know that Lankford’s definition of public mass shootings does not conform to the FBI, Department of Homeland Security, or NYPD definitions. Nowhere do these organizations limit their public mass shooting cases to just one shooter who is sufficiently ‘inner-directed,’ nor do they exclude ‘sponsored’ attacks (see Lott and Moody 2019, 41–42.) The largest number of shooters in the NYPD dataset is ten, but this limit is never explicitly stated as one of the conditions required for inclusion.

Excluding terrorism cases?

Lankford originally claimed:

these public mass shootings—which are also sometimes referred to as active shootings or rampage shootings—stand out as particularly concerning because they are typically premeditated attacks that strike random, innocent victims (Newman, Fox, Roth, Mehta, & Harding, 2004). This makes them functionally similar to terrorism (Lankford, 2013). (Lankford 2016, 188)

No mention was made of the number of terrorists involved in an attack, though he did say: “attackers who committed sponsored acts of genocide or terrorism were not” included (Lankford 2016, 191). In his later paper, Lankford explained that he had included terrorist cases in instances where only one terrorist carried out the attack: “I did include shooters with terrorist motives (like the 2009 Fort Hood shooter) as long as their behavior appeared self-initiated” (2019, 75).

Despite repeated requests for clarification on what Lankford (2016, 191) meant by “sponsored” attacks, Lankford has still never answered us. Should all terrorist attacks be excluded? We think not, both because the NYPD and FBI reports include terrorist attacks and because terrorist and non-terrorist attacks often are, as Lankford himself says, “functionally similar” (2016, 188). If the San
Bernardino killers got training in the Middle East, were they sponsored? Was the first Fort Hood shooter sponsored because he was in communication with one of the influential clerics associated with ISIS? Was the Pulse nightclub shooter sponsored because he was inspired by information put out over the Internet by ISIS? Is funding required to list attacks as sponsored?

Lott (2018a) used the University of Maryland Global Terrorism Database (GTD) to exclude all foreign cases that it labeled as a terrorist attack. Given that we included U.S. terrorist cases, just as the FBI and other organizations included them, but excluded foreign cases, our measure is biased towards making the U.S. share of world public mass shooters larger than it is. Lott concluded:

> Even if one were to eliminate all foreign terrorist attacks on top of all the insurgency ones (and the NYPD dataset clearly includes terrorist cases for both the U.S. and foreign countries), that leaves 709 foreign mass public shooters. That estimate of the number of shooters is still 26 times greater than the NYPD count and 42 times greater if the Mumbai case is cut. (Lott 2018a, 12)

That would imply that the U.S. would make up 5.97 percent of the world’s public mass shooters, counting only cases where the number of shooters is known. Assuming the distribution for cases with an unknown number of shooters is the same as known ones, the U.S. share would be less than 2 percent. The claim that “once these cases were removed from the analysis, Lott’s results more closely resembled Lankford’s” (Booty et al. 2019, 2) is therefore not remotely close to being correct.

## Conclusion

Lankford’s study makes it extremely clear how important it is for researchers to provide their data to others or at least tell people their data sources and how their data were collected. For four years, Lankford refused to do either, and his misleading and error-filled research received much attention worldwide. It shows that the press ought to be very skeptical of studies from scholars who refuse to provide others with their data. Even a quick look at Lankford’s list of cases would have made it very clear that there were significant errors in both the list of United States and foreign cases.

As it is, despite repeated requests, we are still missing the rest of the data that Lankford used to run his regressions, and we have been unable to replicate anything close to his estimates even when using his flawed list of public mass shooters. Lankford has also declined to even answer any questions about how that other data...
set was put together.

Lankford uses neither the FBI nor NYPD definitions that he continually said that he used. We wonder whether the New York Times, Washington Post, and USA Today would have been somewhat reticent to use Lankford’s results if they had known that he had not in fact conformed to those definitions. Now we know that his definition excluded all but a few of the non-U.S. public mass shootings. Further, his dataset contains many errors and he doesn’t use any definition consistently.

We would have no quarrel with Lankford studying lone-wolf shooters, though he didn’t do that, but if he makes an international comparison concerning the number of ‘shooters’ without the qualification ‘who acted alone,’ then he must take them as they come, often in groups. The United States has less than three percent of the world’s public mass shootings or people killed in those incidents (Lott and Moody 2019, 53, Table 1). Using the NYPD definition, the U.S. has much less than its share of public mass shooters.

Allowing other researchers to examine his data would help provide answers to the remaining questions concerning the mysterious Lankford datasets.

Data and code

All data, programs, and results used in the writing of this paper may be downloaded from the journal website (link).

References

Booty, Marisa, Jayne O’Dwyer, Dan Webster, Alex McCourt, and Cassandra Crifasi. 2019. Describing a “Mass Shooting”: The Role of Databases in Understanding Burden. Injury Epidemiology 6: 47. Link


About the Authors

John R. Lott, Jr., is the founder and president of the Crime Prevention Research Center, a research and education organization. Lott is an economist who has held research and/or teaching positions at the University of Chicago, Wharton, Yale, Stanford, UCLA, and the University of Maryland. He has published over 100 articles in refereed journals on a range of topics from law and economics, crime, finance, education, and industrial organization. His email address is johnrlott@crimeresearch.org.

Carlisle E. Moody is Professor of Economics at the College of William & Mary, where he teaches mathematical economics, econometrics, and time series analysis. His research is primarily in the economics of crime, especially the relationship between guns and crime. His email address is cemood@wm.edu.
The Importance of Analyzing Public Mass Shooters Separately from Other Attackers When Estimating the Prevalence of Their Behavior Worldwide

Adam Lankford

LINK TO ABSTRACT

Public mass shootings have traumatized Americans for more than fifty years. Notable examples include the 1966 University of Texas tower shooting, the 1984 San Ysidro McDonald’s shooting, the 1986 Edmond post office shooting, and the 1991 Luby’s Cafeteria shooting. Other horrific incidents include Columbine in 1999, Virginia Tech in 2007, the Aurora movie theater shooting in 2012, and the Sandy Hook shooting that same year. More recent tragedies include the mass shootings at a concert in Las Vegas in 2017, at a high school in Parkland, Florida in 2018, and at a Walmart in El Paso, Texas in 2019.

Over the same period, similar incidents seem to have been extremely rare in other countries, and many Americans have demanded to know why. I was also curious to find out, so several years ago I conducted a cross-national study of public mass shooters (Lankford 2016). The goal was simple: to measure how often public mass shooters attack in different countries, and identify variables that help explain why some countries have more than others.

To make the study’s focus clear, I cited the Department of Homeland Security’s definition of active shooter—“an individual actively engaged in killing
or attempting to kill people in a confined and populated area”—and noted that “their attacks must have (a) involved a firearm, (b) appeared to have struck random strangers or bystanders and not only specific targets, and (c) not occurred solely in domestic settings or have been primarily gang-related, drive-by shootings, hostage-taking incidents, or robberies” (Lankford 2016, 190). I also emphasized that “attackers who struck outdoors were included; attackers who committed sponsored acts of genocide or terrorism were not” and that “only offenders who killed four or more victims were included in this study” (ibid., 190–191).

The results indicated that the United States had 30.8 percent of all public mass shooters from 1966–2012, despite having less than five percent of the world’s population (Lankford 2016). The United States’ status as leader in this unfortunate category was consistent with findings from previous research on rampage school shooters (Böckler et al. 2013) and public mass shooters (Lemieux 2014). My results also showed a statistically significant association between national firearm ownership rates and the number of public mass shooters per country, which was also consistent with prior research (Lemieux 2014), and persisted whether the United States was included in the analysis or not (Lankford 2016).

Unfortunately, John Lott and Carlisle Moody (2019; 2020) have created a great deal of confusion with their recent claims, which grossly underestimate the United States’ global share of public mass shootings. Here I explain:

1. why analyzing public mass shootings and other types of attacks as a single form of violence is as flawed as claiming that tornadoes and hurricanes are a single type of storm;
2. how readers can sort Lott and Moody’s dataset to more accurately estimate the United States’ global share of public mass shootings;
3. how Lott and Moody misrepresent approximately 1,000 foreign cases from their own dataset, and what the corrected figures actually show;
4. why readers should think twice about trusting Lott and Moody’s claims.

The importance of analyzing public mass shootings separately from other violence: A parallel with tornadoes and hurricanes

There are many things that kill people, including viruses, diseases, accidents, natural disasters, armed conflict, and violent crime. And overarching studies of causes of death would include them all. However, it is far more common for re-
searchers to study them separately, because the explanations for each one differ, as do the solutions. If scientists tried to identify variables that jointly explain both natural disasters and violent crime, their analysis would be too broad and likely to fail.

For a moment, let’s pretend I had studied deadly tornadoes. What if after I published findings on the global distribution of tornadoes, a researcher proclaimed that my study was “bogus research,” “fake,” and “fraud” (Lott 2019), and for supporting evidence, presented a dataset which included a small number of tornadoes and a large number of hurricanes.

That is essentially what Lott and Moody have done. I studied public mass shootings, but their dataset includes two fundamentally different types of violence: a small number of public mass shootings (‘tornadoes’) and a large number of group attacks by paramilitary fighters, armed rebels, militia group members, and terrorist strike teams (‘hurricanes’).

In Table 1, I summarize Lott and Moody’s main arguments about the joint analysis of public mass shootings and other types of attacks, and then provide analogous arguments for tornadoes and hurricanes.

| TABLE 1. Lott and Moody’s arguments about the joint analysis of public mass shootings and other types of attacks, and analogous arguments for deadly tornadoes and hurricanes |
|-------------------------------------------------|-------------------------------------------------|
| Lott and Moody’s arguments that public mass shootings and other types of attacks are a single type of violence | Analogous arguments that deadly tornadoes and hurricanes are a single type of storm |
| They have the same basic ingredients: people shooting other people in public. | They have the same basic ingredients: rapidly rotating air. |
| They have the same consequences: innocent victims are killed. | They have the same consequences: innocent victims are killed. |
| These are the same type of shooting, they just appear different, depending on the location. | These are the same type of storm, they just appear different, depending on the location. |
| In the United States, they almost always involve perpetrators who attack alone—in other countries, they usually involve perpetrators who attack in groups. But this does not mean they are fundamentally different. | In some places, they form over landlocked areas and touch down for minutes or hours—in other places, they form over water and last for days or weeks. But this does not mean they are fundamentally different. |
| Group attacks by terrorist organizations, genocidal militias, paramilitary fighters, and armed rebels are just what happen when individual public mass shooters cluster together. | Hurricanes are just what happen when individual tornadoes cluster together. |
| Lankford should have been clearer that group attacks by terrorist organizations, genocidal militias, paramilitary fighters, and armed rebels were not included in his study of public mass shooters. | Researchers should be more clear that hurricanes are not included in their studies of tornadoes. |
Superficially, tornadoes and hurricanes may appear similar, because they both contain rapidly rotating air and can kill innocent people. However, scientists agree they are distinct phenomena. Tornadoes differ from hurricanes based on their warning signs, frequency, shape, size, number of convective storms, life span, and more. In fact, a Live Science investigation could not find a single study in which a shared explanation was sought for both types of storms, because experts regard them as completely different (Wolchover 2011).

Public mass shootings and the other types of attacks in Lott and Moody’s dataset also have superficial similarities: they all involve people shooting and killing other people outside of their homes. But they are fundamentally different phenomena. As I have noted previously, “There are major differences in the psychology, behavior, weapons acquisition, underlying causes, and prevention strategies that apply to these distinct types of violence” (Lankford 2019, 70).2

For one thing, public mass shooters almost always attack alone, as shown by both research on perpetrators in the United States (Berkowitz et al. 2019; Bjelopera et al. 2013; Blair and Schweit 2014; Capellan et al. 2019; Duwe 2016; FBI 2019; Peterson and Densley 2019; Schildkraut et al. 2018), and by research on perpetrators in other countries (Böckler et al. 2013; Lankford 2019; Larkin 2009; Lemieux 2014; Malkki 2014; Mullen 2004). And Lott and Moody’s own data show that more than 95 percent of incidents they include from the United States involved a single perpetrator. However, almost all of the other types of violence in Lott and Moody’s dataset involved groups. This is important because extensive evidence has shown that group dynamics powerfully influence behavior (Cialdini and Goldstein 2004; Kerr and Tindale 2004; Krech, Crutchfield, and Ballachey 1962). In addition, although violent organizations may occasionally attract people who would otherwise kill on their own, it would be absurd to suggest that terrorist groups, armed militias, rebel movements, or paramilitary organizations are simply the combination of individual mass shooters who have joined together. That would be like claiming that hurricanes are composed of individual tornadoes.

Another obvious psychological difference is that public mass shooters are often self-destructive and suicidal (Blair and Schweit 2014; Duwe 2016; Mullen, 2004; Peterson and Densley 2019; Silver, Simons, and Craun 2018), while the other types of attackers are not. Lott and Moody’s own data show that more than 60 percent of the perpetrators they included from the United States committed suicide. In fact, public mass shooters often make suicidal threats or attempts prior to their attacks (Silver, Simons, and Craun 2018), and a significant portion kill themselves before law enforcement even confronts them at the scene (Blair and

---

2. On these differences see, e.g., Duquet 2018; Eichstaedt 2009; Hoffman 1998; Lankford 2015; Moghadelam 2005; Silver, Simons, and Craun 2018; Stein 2017.
Schweit 2014). By contrast, terrorist groups, armed militias, rebel movements, and paramilitary organizations are almost entirely composed of people who want to fight and survive (Hoffman 1998; Lankford 2013). According to START (2020), less than 1.3 percent of all terrorist attacks involving firearms are committed by perpetrators who intended or expected to die. And in Lott and Moody’s dataset, only 0.3 percent of the incidents they included from outside the United States were coded affirmatively for suicide. Even when I add information on suicide for some cases they missed, this raises the total proportion of suicides in their foreign dataset to less than three percent.

There is also a major difference in weapons acquisition. Public mass shooters usually buy their firearms over the counter from gun stores, or through other legal means (Silver, Simons, and Craun 2018). On the other hand, paramilitary fighters, armed rebels, militia groups, and terrorist organizations appear much more likely to obtain weapons through illegal means, such as gun smuggling (Duquet 2018; Eichstaedt 2009).

Furthermore, much like tornadoes and hurricanes are almost never studied as a single type of storm, public mass shootings and the other types of attacks in Lott and Moody’s dataset are almost never studied as a single type of violence. Lott and Moody have taken an unprecedented and unreasonable leap in claiming they should be analyzed together. This is completely inconsistent with previous studies of these shootings in an international context (Böckler et al. 2013; Lankford 2016; 2019; Larkin 2009; Lemieux 2016; Malkki 2014; Mullen 2004). Despite an extensive review, I could not find a single peer-reviewed study of public mass shootings that categorized them with attacks by terrorist organizations, genocidal militias, armed rebel groups, and paramilitary fighters, which is what Lott and Moody have done—and then falsely claimed as the scholarly norm. They are way out of bounds on this.

The only possible exception is not peer-reviewed, but rather the New York City Police Department’s (2012) active shooter report, which included a few cases of group terrorism or genocide in its appendix, along with hundreds of active shootings as they have been traditionally defined. One of the NYPD report’s strengths was its list of many cases, which allows researchers to identify which are relevant and which are not. However, the report was focused on enhancing the security of buildings—not on analyzing causal factors or preventative solutions. Therefore, it made sense to consider multiple types of threats, even though they are fundamentally distinct phenomena, much like it makes sense for the Federal

---

As of March 15, 2020, the Global Terrorism Database showed that only 841 of 67,501 attacks involving firearms were suicide attacks (i.e., committed by perpetrators who intended or expected to die), and 841 / 67,501 = 1.25 percent (START 2020).
Emergency Management Agency (FEMA) to plan for both terrorist attacks and natural disasters. More importantly, however, the NYPD *did not include any attacks by terrorist groups or genocidal militias in their analysis*, so these incidents had no distorting effect on their conclusions.

Given that public mass shootings and the other types of attacks in Lott and Moody’s dataset are fundamentally different, why not analyze them separately? I suspect Lott and Moody knew that if they combined them into a single category, the American mass shooting problem would no longer seem so bad. They could proudly proclaim that “the U.S. has 1.25 percent of the world’s mass shooters” (Lott and Moody 2020, abs.).

However, if these are recognized as separate types of violence, then it is easy for anyone to understand that the United States could have six times its global share of one type, without having many of the other type (Lankford 2019). It is also obvious that separate explanations could exist for each type. Easy access to firearms could explain why the United States has so many public mass shooters—as both Frederic Lemieux’s empirical study (2014) and my empirical study (Lankford 2016) independently found—even if civilian firearms do not explain the prevalence of terrorist groups, armed militias, rebel movements, or paramilitary organizations.

**How can readers sort Lott and Moody’s dataset to estimate the United States’ global share of public mass shootings?**

If someone published a dataset with a small number of tornadoes and a large number of hurricanes, but did not label them by type of storm, how could readers identify the tornadoes for themselves? One approach would be to read through the entire dataset, and determine whether each storm is a tornado or not based on case descriptions. Another option would be to sort the dataset based on some distinguishing characteristic of tornadoes, exclude all cases without that characteristic, and thereby eliminate most of the hurricanes.

Similarly, anyone could read through every case in Lott and Moody’s dataset and identify which are public mass shootings and which are fundamentally different. For instance, their dataset includes:

- a massacre ordered by the President of Nigeria, in which his soldiers “killed up to 200 civilians and caused thousands of villagers to flee into the bush” (case #333)
• a group attack intended “to destroy a bridge over the Dange River to disrupt the flow of good[s] between the Zaire and Uige provinces” (case #269)
• a firefight in which “opposing political factions engaged in rocket-propelled grenade, mortar, grenade, rocket, and small arms attacks” (case #966)
• an armed robbery in which a large group of “rebels attacked a state-owned explosives depot…The group stole between 20–50 tons of explosives and detonators and 17 rifles” (case #660)
• a prolonged crisis in which “Pakistani soldiers were taken hostage by tribal militants in an ambush while on patrol” and then executed four days later (case #513)
• an incident in which suspected members of a genocidal militia killed civilians and “stole 2177 livestock animals” in Darfur, Sudan (case #689)

A massacre ordered by a country’s president; an attempt to destroy a bridge; a firefight; a robbery; a hostage crisis; and an attack in which two thousand animals were stolen. Yes, these all resulted in people being killed outside of their homes, but apart from that, they have very little in common with public mass shootings like Columbine, Sandy Hook, Las Vegas, or Parkland. I do not know a single respected criminologist who would categorize them together, nor any peer-reviewed research in which that has ever been done.

To account for these fundamental differences, another option is to sort Lott and Moody’s dataset based on a distinguishing characteristic of public mass shootings, exclude all cases without that characteristic, and thereby eliminate most of the other types of attacks.

For instance, we can focus exclusively on cases where at least one perpetrator is known to have personally killed four or more victims. This seems like a defining element of public mass shooters’ behavior, and was a criterion for my 2016 study. It is also the main reason why mass shooters are studied separately from other shooters in the first place. The rationale is that perpetrators who personally kill a large number of victims are an especially terrible form of criminal, worthy of their own categorization and scrutiny.

As shown in Table 2, if we focus exclusively on cases where at least one perpetrator is known to have personally killed four or more victims, most of the

---

4. There are two ways to know if at least one perpetrator killed four or more victims: first, if there is only one perpetrator and more than four victims were killed; second, if there are multiple perpetrators and they averaged killing more than three victims each. For instance, if two perpetrators killed seven victims, or three perpetrators killed ten victims, that suggests that at least one of them must have killed four or more.
other types of attacks in Lott and Moody’s dataset are screened out. As a result, their dataset shows that from 1998–2012, the United States had 24 percent of the world’s public mass shootings. That is more than five times its global share.

Another option is to use an approach more similar to my original study’s method (Lankford 2016). We can analyze Lott and Moody’s dataset by focusing exclusively on public mass shootings that involved at least one perpetrator known to have personally killed four or more victims, and exclude what I referred to as “sponsored acts of genocide or terrorism” (2016, 191). By this I meant that if the perpetrator attacked because he was a member of a terrorist organization or genocidal group, he was not included; otherwise he was included, regardless of his beliefs or sources of anger. No mind reading was required: post-attack investigations regularly examine whether a shooter was part of a violent organization or group. If Lott and Moody’s data are analyzed based on the cases that meet this description, they show that from 1998–2012, the United States had 30 percent of the world’s public mass shootings, which is more than six times its global share. This is similar to my original study’s result of 30.8 percent, even though I studied many more cases that met these criteria over a much larger time span (Lankford 2016).

As I demonstrated in Econ Journal Watch one year ago (Lankford 2019), another way to sort Lott and Moody’s dataset is by number of perpetrators, and focus on the cases where shooters attacked alone. Of course, a few legitimate cases would be missed, such as the 1999 Columbine attack. However, research on incidents in the United States and in other countries suggests that focusing on single perpetrators would account for 95–99 percent of total cases (Berkowitz et al. 2019; Bjelopera et al. 2013; Blair and Schweit 2014; Böckler et al. 2013; Capellan et al. 2019; Duwe 2016; FBI 2019; Follman, Aronsen, and Pan 2019; Lankford 2019; Lemieux 2014; NYPD 2012; Peterson and Densley 2019; Schildkraut et al. 2018).

By this metric, Lott and Moody’s dataset shows that from 1998–2012, the United States had 29.7 percent of the world’s public mass shootings by perpetrators who attacked alone. This is more than six times the United States’ proportionate

---

5. Some examples illustrate how misleading it is for Lott and Moody to count cases where this is unknown. For instance, they include an attack in which seven victims were killed by 300 members of the separatist Moro Islamic Liberation Front (case #1037). These perpetrators averaged killing 0.02 victims each (7 / 300 = 0.02). As another example, they include a case in which “Thousands of armed protestors stormed a jail in an effort to free 23 members of the Akramia religious group… At least nine people were killed” (case #585). Because this incident involved at least 2,000 perpetrators (it is described as multiple “thousands”) and nine victim fatalities, the perpetrators averaged killing 0.005 victims each (9 / 2000 = 0.005).

6. The determination of whether each attack was committed by member(s) of a terrorist organization or genocidal group was based solely on information in Lott and Moody’s dataset, and therefore may be incomplete. Additional foreign attacks would likely be excluded if further research were conducted on perpetrators’ group affiliations.
share, given its population. In fact, based on these data, the United States had more public mass shootings by perpetrators attacking alone than any continent except Asia, which has more than ten times the U.S. population (Lankford 2019). In this issue of *Econ Journal Watch*, Lott and Moody (2020) attempt to undermine this finding, but as I will demonstrate in the next section, their new claims provide further reason to question their credibility, because they misrepresent approximately 1,000 cases from their own dataset.

### TABLE 2. The United States’ global share of public mass shootings, according to different criteria

<table>
<thead>
<tr>
<th>Data source</th>
<th>Criteria</th>
<th>The United States’ global share of incidents worldwide</th>
<th>Number of U.S. incidents / Number of total incidents worldwide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lankford (2016)</td>
<td>all incidents (1966–2012)</td>
<td>30.8%</td>
<td>89 / 289</td>
</tr>
<tr>
<td>Lott and Moody (2019)</td>
<td>all incidents in which at least one perpetrator is known to have killed four or more victims (1998–2012)</td>
<td>24.0%</td>
<td>42 / 175</td>
</tr>
<tr>
<td>Lott and Moody (2019)</td>
<td>all incidents in which at least one perpetrator is known to have killed four or more victims, and the attack was not committed by member(s) of a terrorist organization or genocidal group* (1998–2012)</td>
<td>30.0%</td>
<td>42 / 140</td>
</tr>
<tr>
<td>Lott and Moody (2019)</td>
<td>all incidents in which the perpetrator is known to have attacked alone (1998–2012)</td>
<td>29.7%</td>
<td>41 / 138</td>
</tr>
</tbody>
</table>

*Notes: The United States has approximately 4.5 percent of the world’s population. One of Lott and Moody’s (2019) foreign cases was removed prior to these calculations because it was a duplicate of the same incident (#960, #961). Lankford’s (2016) study included 292 public mass shooters from 289 incidents. See Appendix B for coding and calculations. This is most similar to the criteria used in Lankford’s (2016) study, even though it does not screen out some of the other fundamentally different cases in Lott and Moody’s dataset.

Again, by focusing on (a) incidents in which at least one perpetrator is known to have killed four or more victims, (b) incidents in which at least one perpetrator is known to have killed four or more victims and the attack was not committed by member(s) of a terrorist organization or genocidal group, or (c) incidents in which the perpetrator is known to have attacked alone, we can distinguish most public mass shootings in Lott and Moody’s dataset from the other types of attacks. However, this approach is still not as effective as reading through every case and carefully identifying which should be included and which should not, based on definitions of public mass shootings and prior research on this behavior.
Lott and Moody’s misrepresentation of approximately 1,000 cases from their own dataset

Lott and Moody (2019, 39) admitted in *Econ Journal Watch* (apparently for the first time) that the United States has “an outsized number” of public mass shootings by single perpetrators. In their 2020 rejoinder, however, they attempt to deny what their own data show: that from 1998–2012, the United States had 29.7 percent of the world’s public mass shootings by perpetrators who attacked alone.

Lott and Moody’s new claim is that 1,052 foreign attacks from their dataset are completely “unknown” (2020, 31–32) as to whether they were committed by single or multiple perpetrators, and they assign each of these cases a 24.8 percent chance of being committed by a single perpetrator. Then they add more than two hundred “estimated” cases to their known cases. This greatly reduces the U.S. proportion of total incidents, because Lott and Moody left the U.S. count unaltered, while more than tripling the number of foreign attacks with their estimate of supposedly “unknown” cases.

A close inspection of their dataset, however, reveals that Lott and Moody have misrepresented approximately 1,000 foreign attacks. As shown in Table 3 and Appendix B, only 58 cases are actually unknown as to whether they were committed by single or multiple perpetrators. Lott and Moody fail to acknowledge the difference between not knowing the total number of perpetrators in an attack, and not knowing whether the number was more than one. As a result, they misrepresent 953 incidents for which their data clearly do indicate there was more than one perpetrator.

Readers can see this by reviewing the incident summaries and details for the supposedly unknown cases in Lott and Moody’s dataset. The vast majority of attackers are described using plural nouns (e.g., “assailants,” “rebels,” “militants,” “terrorists,” or “gunmen”), not singular nouns (e.g., “an assailant,” “a rebel,” “a militant,” “a terrorist,” or “a gunman”). Although the importance of this distinction seems obvious, I verified it by asking Erin Miller, principal investigator and manager of the Global Terrorism Database (which was the source for over 90 percent of Lott and Moody’s data). She confirmed that the singular/plural distinction matters, and added that she “would recommend interpreting assailant descriptors much like you would in a news article, because the team follows what the...

7. Lott and Moody (2020) applied estimates of 73 percent and 34 percent in two stages. Overall, this means they assigned each of these cases a 24.8 percent chance of being committed by a single perpetrator (73% × 34% = 24.8%).
source articles say.”

<table>
<thead>
<tr>
<th>Table 3. Lott and Moody's (2020) misrepresentation of approximately 1,000 foreign attacks from their own dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single perpetrators</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>Lott and Moody's (2020) claims</td>
</tr>
<tr>
<td>Corrected figures, based on information in Lott and Moody's own dataset</td>
</tr>
</tbody>
</table>

Notes: For Lott and Moody's (2020) claims, see my Appendix A; for the corrected figures based on information in Lott and Moody's own dataset, see my Appendix B. Case #1095 had a missing number in the summary description (“At least _ Muslim gunmen…””) which was verified through an internet search. The status of some of the remaining “unknown” cases could also be clarified through additional research. In their dataset Lott and Moody (2019) identified 16 cases as having “>1” perpetrator, but now Lott and Moody (2020) count each of those cases as having two perpetrators in their rejoinder’s estimates.

Lott and Moody also fail to account for multiple-perpetrator cases in other important ways. For instance, some have a numeric estimate for how many attackers were captured, killed, or wounded—not counting those who escaped. Others are described as having perpetrators who attacked while simultaneously driving multiple vehicles, which is impossible for one person to do alone. And a few involve attackers striking at multiple locations at the same time.

There are also more than 40 foreign attacks that Lott and Moody falsely classify as “missing values” (2020, 31–32) in their rejoinder, despite having a numeric estimate for multiple perpetrators listed in their own dataset’s incident summary or details. As examples:

- “about 150 insurgents seized two police stations…” (case #534)
- “Around 100 Allied Democratic Forces (ADF) rebels attacked the Kirindi displaced persons camp…” (case #146)
- “Fifty to 60 unknown attackers ambushed an Algerian military convoy…” (case #561)
- “More than 50 unidentified gunmen killed 18 Shi’i employees of a brick factory…” (case #673)
- “a group of more than 50 militants attacked a village…” (case #1080)
- “Approximately 50 attackers dressed in police and army uniforms attacked Bompai police barracks…” (case #1264)

Again, these cases are all claimed by Lott and Moody to be unknown as to whether
they were committed by single or multiple perpetrators, and they estimate each one has a 24.8 percent chance of having been committed by a single attacker.

**Conclusion**

Many Americans want to know how their country’s public mass shooting problem compares to the rest of the world’s. Unfortunately, the evidence consistently shows that the United States has far more than its share of these shootings and attackers (Böckler et al. 2013; Lemieux 2014; Lankford 2016). As I demonstrated last year (Lankford 2019) and in this article, my original study’s results (Lankford 2016) have now been confirmed by multiple replications, with various approaches, using Lott and Moody’s own data.

Beyond the specific crime of public mass shootings, if Lott and Moody want to study how the United States compares to the rest of the world based on all violence, homicides, mass killings, or mass shootings, there would be nothing wrong with that. However, that is not what they have done, either. Instead, they have excluded many mass shootings that occur in public locations because they were related to gangs, robberies, or other criminal behavior. As a result, their overall dataset reflects neither the global distribution of public mass shootings as a specific and cohesive type of violent behavior, nor the distribution of all shootings that kill masses of people in public.

Lott and Moody have also made many false claims, distortions, and errors that raise serious questions about their credibility. For instance, Lott has:

- used “inappropriate statistical methods . . . to create the false impression that mass shootings are less frequent and less deadly in the United States than in European countries,” according to analysis by the fact-checking service Snopes (MacGuill 2018);
- changed how he counted public mass shootings by including nearly 500 battles over sovereignty in his dataset (Lott 2018a; Lott and Moody 2019), after previously claiming they should be excluded to make “a fair comparison” with shootings in the United States (Lott 2015);
- along with Moody, claimed the Northern Mariana Islands has a mass shooting rate 100 times higher than that of the United States, without calling attention to the fact that the Northern Mariana Islands had only one incident in the entire study period (Lott and Moody 2019, 62);
- along with Moody, insisted they found “no significant relationship” between firearms and terrorist shootings worldwide, but also suggested “more guns, less terrorism” (Lott and Moody 2019, 60, 47); and
• accused the FBI of “slicing the evidence to distort the results” about mass shootings in the United States “to promote a political agenda,” which Lott initially blamed on the FBI’s operation under the Obama administration (Lott 2014), but more recently has blamed on the FBI under the Trump administration (Lott 2018b).

In addition, Lott and Moody (2020) now:

• hold up the FBI’s definition as a shining beacon of legitimacy, while making similar accusations and conspiratorial innuendos about my work;
• claim that I included “nine cases that don’t meet the FBI or NYPD definitions of mass public shootings” (p. 30), even though all nine of those cases appear in the FBI or NYPD reports, as readers can verify for themselves (Blair and Schweit 2014; Lankford 2019; NYPD 2012);
• falsely count 16 attacks from their dataset as having only two perpetrators in their 2020 rejoinder, after labeling them “>1” in 2019 (which indicated a much larger and more accurate range);
• claim they “excluded gang violence” (p. 35) in both the United States and other countries, even though their dataset has many foreign attacks described as involving a “gang,” as readers can verify through a quick word search;
• categorize public mass shootings together with attacks by terrorist organizations, genocidal militias, armed rebel groups, and paramilitary fighters—and then falsely claim this as the scholarly norm;
• include more than 1,300 cases in their dataset (more than 90 percent of their foreign cases) for which it is unknown if any perpetrators personally killed four or more victims, based on their own coding;
• misrepresent approximately 1,000 foreign cases from their own dataset, by claiming these cases are “unknown” (pp. 31–32) as to whether they were committed by single or multiple perpetrators, even though information in their own dataset shows that more than 90 percent of these attacks involved multiple perpetrators; and
• either did not bother to read their own dataset’s incident summaries and details for foreign cases, or are intentionally deceiving people about them.

8. The following cases are described as involving a “gang”: #34, #41, #74, #195, #212, #305, #449, #502, #560, #828, #912, #922, #983, #1148, #1198, and #1355. This does not include other attacks by gangs that were described using some other term.
If you are a reader who believed Lott and Moody in the past, and now realize that was a mistake, this is a chance to make a change in who you trust. Choose wisely, and it could greatly improve your life. When the information you receive actually reflects reality, your ability to understand the world around you grows exponentially stronger.

Appendices

There are two online appendices: Appendix A (link) is the Lott and Moody (2020) dataset unedited, and Appendix B (link) is the Lott and Moody (2020) dataset with my corrections.

References


Lott, John R. Jr. 2018b. The Problem With the FBI’s ‘Active Shooter’ Data. RealClearPolitics (RealClear Media Group, Northbrook, Ill.), October 23. Link


Adam Lankford is a professor of criminology and criminal justice at The University of Alabama. He is the author of two books and many peer-reviewed journal articles on various types of criminal behavior, including mass murder, mass shootings, and terrorism. His research has examined perpetrators’ psychological tendencies, mental health problems, suicidal motives, fame-seeking tactics, copycat behavior, and weapons acquisition—along with the strategies that might be used to prevent their attacks. His email address is adam.lankford@ua.edu.
Do Film Incentive Programs Promote Economic Activity?
A Comment on O’Brien and Lane

John Charles Bradbury

LINK TO ABSTRACT

By the mid-2000s, most U.S. states had adopted some sort of film incentive program—a development that this economist is sorry to see. The incentive programs typically are issued as a percentage of in-state production expenditures, through tax credits that are government-refundable or transferable to third parties, though they are sometimes funded through direct grants. Film incentives represent a substantial subsidy that typically allow the recouping of between 15 to 30 percent of in-state production expenditures covered by some such program (Bradbury 2019a). Proponents of such programs often argue along such lines as: Film incentives attract an otherwise absent industry to the state, encourage growth in a new sector, or stimulate economic activity in complementary industries; growth in these sectors may then spill over into unrelated sectors that translates into economy-wide growth through a multiplier; etc.

The abundance of film incentive programs has also stimulated a recent small but vibrant literature on their efficacy in stimulating economic activity. Using a variety of samples and empirical methods, most studies have found little to no positive net impact of film incentives on economic activity (e.g., employment, industry establishments, and output), which indicates film incentives are ineffective as economic development policy (Bradbury 2019a; 2020; Button 2018; Swenson 2017; Thom 2018). One recent study that is representative of this literature is that of Patrick Button, who uses a panel difference-in-differences regression approach

1. Kennesaw State University, Kennesaw, GA 30144.
to examine the film industry across states and finds “the ability for tax incentives to affect business location decisions and economic development is mixed, suggesting that even with aggressive incentives, and ‘footloose’ filming, incentives can have little impact” (2019, 315). In another recent study, using a discrete choice model that accounts for competing incentives offered by all states, Mark Owens and Adam Rennhoff (2018) “fail to find strong evidence that incentives create a more permanent movie industry in a state.”

While conducting research on film incentives, I discovered the article by Nina O’Brien and Christianne Lane (2018), published in Regional Studies and titled “Effects of Economic Incentives in the American Film Industry: An Ecological Approach,” which presents empirical results for the period 1998–2010 that appear contrary to what I perceive to be an academic consensus. O’Brien and Lane’s abstract says that “the simple presence of economic incentives, organizational diversity and the presence of dominant organizational populations are associated with increases in filming, employment and new establishments” (2018, 865). The study also says that its results support “Hypothesis 1a: Incentives targeting an industry will promote greater levels of industrial activity” (ibid., 866). The study says it uses an “ecological approach,” “employing concepts from biological science” to explain “how organizations of various kinds interact as they struggle to secure resources” (ibid.). The study “offers a comprehensive, longitudinal overview of project-based incentive programmes as they developed over time and across the United States as a whole” (872). The authors say that their estimates imply “that economic development outcomes are related not only to the presence and value of financial incentives but also to the ecological factors organizational diversity and presence of dominant firms” (ibid.).

On the other hand, a careful reader of the study will also find results that reflect negatively on film incentive programs. On the whole, the paper does not present itself as providing strong support for incentive programs. I shall not attempt to convey the authors’ ecological approach and their findings. Policy implications are suggested in the article’s penultimate paragraph, which I quote in full:

This study also highlights the importance of a state’s organizational diversity. For each additional organizational type present in a state, we find an increase in filming (3.8%), employment (4.0%) and establishment (1.9%). Lack of diversity may exacerbate the tendency for ‘fly-in fly-out’ production described above. Together, the effects of the simple presence of an incentive, dominant populations and diversity suggest that rather than outspend one another through costly tax credits, states might achieve better results by tailoring existing incentives to increase organizational diversity, particularly by targeting companies associated with distribution, marketing and sales. (O’Brien and Lane 2018, 873)
All of this intrigued me, but as I read the paper more closely, I noticed some odd choices regarding data coding and sample selection that raise doubts about their empirical estimates. Here I share some comments on the article.

In order to assess the findings, I needed to validate the coding of film incentive policies and spending by state and year. I knew from experience the difficulty of identifying and quantifying incentives across states and over time. Records of film incentives are not compiled in a central database, so empirical analysis requires building a database from disparate sources to designate states as having a film incentive program in force. When exactly incentives were implemented, suspended, or ended, and how much funding each program provided, requires checking multiple government and media outlets that do not always agree. Other studies that present their data on film incentives often differ slightly from each other in their designations. Perhaps the authors’ database was more comprehensive and superior to those collected by other researchers.

I reached out to the authors and requested their database of film incentive programs and spending. They did not respond to my initial email requests. Follow-up emails over several weeks yielded cursory responses, but neither author was willing to provide the data to me. I reached out to a Regional Studies editor for assistance in acquiring the data, noting that the lack of response was inconsistent with the journal’s data availability policy: “Authors are encouraged to share or make open the data supporting the results or analyses presented in their paper where this does not violate the protection of human subjects or other valid privacy or security concerns” (link). In my request to the editor I also stated several concerns about the data that I note in this comment. The editor responded that the journal’s data sharing policy does not require authors to share their data, but that I was welcome to share my perspective through the review process. Therefore I wrote up my data concerns in a short comment (raising several of the issues I discuss in the present article) and submitted it to the journal. After several months the editor informed me that my comment received consideration at an editors’ meeting but was not accepted because “the commentary does not reflect the spirit of the journal to promote research and scholarly advancement” and “the tone does not strike the expected balance to be collegiate and constructive.”2 The decision letter mentions comments from a referee, but the editor did not provide a referee report. The editor did not address any of my concerns about the accuracy of the data.

In the remainder of this article, I detail shortcomings in the empirical analysis in O’Brien and Lane (2018). It is my hope that publishing my concerns in Econ Journal Watch will elicit a response from the authors regarding my concerns about

---

2. I have uploaded at SSRN my comment as it was submitted to Regional Studies, except for removing my anonymity (Bradbury 2019b). Readers are free to judge the appropriateness of my tone.
the estimates as they are presented in *Regional Studies*.

O’Brien and Lane (2018) use panel data of U.S. states from 1998 to 2010, a period that extends from the early days when only a few states offered film incentives to when most states employed film incentives. The analysis estimates the impact of incentives and other factors on filming and other economic activity. The study draws data on film production locations, industry diversity, and industry dominance from the Internet Movie Database (IMDb), and economic data are gathered from the U.S. Census Bureau and the Bureau of Labor Statistics.

The main variables of interest on the availability and magnitude of film incentives were determined by O’Brien and Lane, but the data are not reported in the paper or elsewhere. They report that “incentive data were collected from states’ departments of economic development, film and television commissions, and the official record of incentive legislation” (O’Brien and Lane 2018, 868). The empirical strategy uses two variables, “Incentives (dichotomous)” and “Incentives (millions),” to measure the use and magnitude of film incentives by states. The former is an indicator variable equal to one if the state offers incentives and zero otherwise. The latter is a continuous variable that measures state spending on film incentives. The paper reports the descriptive statistics for incentive spending—ranging from $0 to $300 million, with the average being $27.12 million (for all states, not just states offering incentives)—but it does not provide full summary statistics or more detailed information regarding the dichotomous existence of incentives, such as which states adopted film incentives and when. In the text, the authors report that 67.5 percent of the observations occur without incentives in place (ibid.).

The variable description indicates a potential problem with the coding of incentive spending in states without funding caps. States with uncapped incentives are not limited in the amount of funding that they can provide for film production; thus, these states have the potential to offer greater subsidies than states with funding constraints. Rather than report actual spending for these states, O’Brien and Lane value uncapped incentives as follows: “After consultation with film commission informants, uncapped incentives were valued at US$300 million, reflecting a 10% increase over the highest incentive offered” (2018, 868). While the authors justify this coding as informed speculation from insiders, available data on film incentives offered by uncapped states do not support the use of a proxy for film spending of that magnitude. In fact, states were spending far less than $300 million to fund their incentives; thus, O’Brien and Lane’s reported estimates of the association between incentive spending and economic activity might be quite inaccurate.

Collecting data for all 650 observations in the sample to replicate the authors’ estimates would be a laborious task, so I provide data from a few representative
states that are illustrative of film incentive spending in uncapped states. Table 1 reports tax credit subsidies per year for four states that operated with uncapped film incentive programs during the sample period: Georgia, Louisiana, New Mexico, and North Carolina. Though other states have operated film incentive programs without caps, the states listed in Table 1 have been recognized as offering generous and/or aggressive incentive programs to encourage filming and thus should be representative of the high end of subsidies offered by states (Button 2018; Bradbury 2019a).

<table>
<thead>
<tr>
<th>Year</th>
<th>Georgia</th>
<th>Louisiana</th>
<th>New Mexico</th>
<th>North Carolina</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2003</td>
<td>NA</td>
<td>1.5</td>
<td>1.11</td>
<td>NA</td>
</tr>
<tr>
<td>2004</td>
<td>NA</td>
<td>1.5</td>
<td>1.74</td>
<td>NA</td>
</tr>
<tr>
<td>2005</td>
<td>NR</td>
<td>47.4</td>
<td>2.1</td>
<td>0.23</td>
</tr>
<tr>
<td>2006</td>
<td>NR</td>
<td>70.4</td>
<td>5.71</td>
<td>10.67</td>
</tr>
<tr>
<td>2007</td>
<td>NR</td>
<td>134.6</td>
<td>18.5</td>
<td>17.55</td>
</tr>
<tr>
<td>2008</td>
<td>NR</td>
<td>92</td>
<td>46.03</td>
<td>11.54</td>
</tr>
<tr>
<td>2009</td>
<td>89.25</td>
<td>101</td>
<td>76.71</td>
<td>7.7</td>
</tr>
<tr>
<td>2010</td>
<td>171.87</td>
<td>165.6</td>
<td>65.39</td>
<td>9.33</td>
</tr>
<tr>
<td>Mean</td>
<td>131.56</td>
<td>76.75</td>
<td>27.16</td>
<td>9.5</td>
</tr>
<tr>
<td>Maximum</td>
<td>171.87</td>
<td>165.6</td>
<td>76.71</td>
<td>17.55</td>
</tr>
</tbody>
</table>

Notes: Nominal dollars in millions. NA: Not applicable. NR: Not reported by the state. Sources: Georgia: Governor’s Office of Planning and Budget (various years), tax credits approved. Louisiana: Mathis 2012. North Carolina: Department of Revenue (various years). New Mexico: Popp and Peach (2008); New Mexico Taxation and Revenue Department (2012).

The sources listed under Table 1 tell us the following:

- Georgia instituted its uncapped film tax credit program in 2005. The state reports that it certified an average of $131 million in tax credits annually in 2009 and 2010. Though the state does not report the value of tax credits approved in prior years, the reported tax credits were issued after Georgia increased its incentives in 2008, thus it is reasonable to assume that the state offered fewer subsidies in previous years. The value of tax credits Georgia has issued since that time has increased every year, and the state did not issue more than $300 million in tax credits until 2014, four years after 2010, the close of O’Brien and Lane’s sample period.
- After instituting its film incentive program in 2002, Louisiana’s incentive funding ranged from $1.5 million to $166 million per year through 2010, with an average value of $77 million. Louisiana would institute a
$180 million budget cap in 2015.

• New Mexico instituted its uncapped film tax credit program in 2002. The state averaged $27 million in annual subsidies through 2010, funding a maximum of $77 million in tax credits in 2009. New Mexico imposed a $50 million cap on its program in 2012.

• North Carolina instituted an uncapped tax credit program in 2005 that averaged $9.5 million in expenditures per year from 2005 to 2010. The state increased its incentives in 2010, and from 2011 to 2015, the state paid out almost $60 million per year to fund its incentives. North Carolina operated a small and capped film grant program from 2000 to 2004 and switched back to a capped grant-based film incentive program in 2015.

These examples from states with reputations for generous film incentives without program spending caps indicate that to assume states with uncapped incentives spent $300 million annually on film incentives during the sample period is to overstate greatly the funding provided. It is unclear how many observations in the O’Brien and Lane study are designated as uncapped. The high variance in incentives offered across these uncapped states indicates that assuming a simple value of incentives offered for all uncapped incentive programs is inappropriate without further justification.

Evidence that incentive spending in uncapped states was far less than $300 million is also available in other published studies. In separate estimates, Joseph Bishop-Henchman (2011) and Michael Thom (2018) report that cumulative expenditures on film subsidies for all states, not for any one state, did not exceed $300 million until 2006 and reached $1.4 billion by 2010. These data indicate that no individual state-year observation exceeded $300 million prior to 2006; yet, during this period several states operated with uncapped incentives. Even after 2006, there is not much room more for more than a few states to offer $300 million in incentives, and, after that time, a majority of states employed film incentives.

Charles Swenson (2017) documents state-specific film incentive spending for all states from 1998 to 2011—a range that includes the entire period studied by O’Brien and Lane (2018)—by listing the maximum annual funding for each state during the period. The only state that exceeded $300 million in annual film incentive expenditures during this period was New York, which had a funding cap of $420 million. Louisiana is reported to have the highest annual expenditure among uncapped states at $236 million. Neither of these extreme values support the assigning of the value $300 million for uncapped states.

While exact amounts of subsidies to film production companies are difficult to identify and differ across sources, available evidence indicates that the subsidies
awarded by uncapped states during this period were far less than the $300 million assigned by the authors. I have been unable to identify a single instance of an uncapped state offering incentives approaching this magnitude. The $300 million designation likely has a profound effect on the estimated coefficients in the analysis in terms of both magnitude and statistical significance; therefore, it is likely that the estimates are misleading.

I also wish to note that it is also possible that none of the observations in Table 1 are coded as offering $300 million in film incentives in the researchers’ data, and that the authors’ coding of those observations resembles the expenditures reported here. Perhaps the variable description in the text of the paper does not accurately describe the data used to generate the estimates. Because the authors have not met my request to share their data, I have been unable to determine how they coded their observations. I must accept the authors’ description of their coding, and as such the coding would seem to be highly inaccurate.

I have another concern regarding sample selection. O’Brien and Lane use film incentive data from 49 states and the District of Columbia, excluding Iowa, “because criminal investigations into that state’s incentive programme prohibited verification of reported data” (2018, 869). The authors cite Richard Verrier (2011) as the source for this justification. Rather than support the contention that data for Iowa were not available for verification, however, Verrier (2011) reports tax credit information from a state audit, which was released to the public on October 26, 2010. I was able to find the audit in an online archive using an internet search for the document (Office of Auditor of State 2010). The 273-page audit contains detailed documentation of all film tax credit spending in the state. In my experience researching film incentives, I have not found a more detailed account of film incentive funding than the Iowa audit. Also, I am not aware of any other cross-state empirical studies that exclude Iowa for the reason the authors give when estimating the impact of film incentives. Thus the exclusion of Iowa is odd and is not supported by the authors’ justification for doing so. Given the sensitivity of the coefficient estimates to included/excluded variables across specifications, it is reasonable to wonder if the study’s estimates are sensitive to its unique exclusion of Iowa.

A third potential problem is that the model estimates for industry employment and number of establishments are based on data from the North American Industry Classification (NAICS) code 5121: motion pictures and video industries. While seemingly appropriate, this is an aggregated category that includes both film production and exhibition. The latter classification includes establishments such as movie theaters and video rental stores. These types of businesses would not be incentivized by film production incentives, and they also reflect the population of the state, which is not included as a control variable in the regression analysis. Film incentives are designed to encourage production not
consumption of motion pictures. The more pertinent NAICS code is the sub-category 51211: motion picture and video production, which excludes activity related to consumption of movies and videos. The impact of using the broader category is likely minor, but the production sub-category is more appropriate. Button (2018; 2019) uses employment and establishment data from the 51211 classification, which is available from the Bureau of Labor Statistics. It is possible that O'Brien and Lane used the more appropriate sub-category but misdescribed what they had done (2018, 868), but we cannot know without seeing their data.

My final concern regards O'Brien and Lane’s interpretation of their estimates, even if those are taken at face value. The authors report that the estimates “show that above the impact of their dollar value, the simple presence of economic incentives,” along with a host of other factors, “are associated with increases in filming, employment and new establishments” (2018, 865). The “above the impact of their dollar value” qualification is curious phrasing, which is misleading for policy evaluation. All film incentive programs require funding, sometimes tens to hundreds of millions of dollars in spending (as noted by the authors). Even if the estimates are correct, further investigation of the opposing impacts with real-world data is needed to identify whether the overall effect of incentives tends to be positive, negative, or negligible. Furthermore, the estimates show a positive impact of incentive spending on filming activity but negative impacts on film employment and establishments. That incentive spending would induce increased filming while decreasing the number of film-industry employees and establishments is difficult to resolve. This tension is not addressed in the paper, and the discrepancy indicates that the results are in some sense not robust.

In conclusion, O’Brien and Lane (2018) conduct an empirical investigation of the impact of film industry incentives on several measures of industry performance, and their findings are mixed and, frankly, murky. The study’s estimates appear to rely on incentive spending estimates that appear to be much higher than actual spending. The authors have not met my requests for the original data that could verify the variable coding. In addition, the authors make some odd choices, such as an unjustified sample exclusion of Iowa from the sample and some estimates use data from a less-than-ideal industrial classification. And even if the estimates are accepted at face value, the reported findings do not necessarily imply that film incentives have anything like the reported impacts, because the estimates are in tension with one another and are sensitive to specification choices.

The desirability of film incentives is an important and relevant issue for policymakers. Concerns regarding the accuracy of the data and interpretation the data need to be answered before the results of O’Brien and Lane (2018) can be relied upon for evaluating film incentives as a policy tool. Researchers working in this area should be aware of the apparent deficiencies with the study.
References


Governor’s Office of Planning and Budget. Georgia. Various years. Agency Performance Measures. Governor’s Office of Planning and Budget (Atlanta). Link


New Mexico Taxation and Revenue Department. 2012. 2012 New Mexico Tax and Expenditure Report. New Mexico Taxation and Revenue Department (Santa Fe, N.M.). Link

North Carolina Department of Revenue. Various years. Film Credits and Grants. North Carolina Department of Revenue (Raleigh, N.C.). Link


About the Author

John Charles Bradbury is professor of economics at Kennesaw State University. His research interests include the economics of sports and public economics. He developed his recent research agenda on the efficacy of film incentives after learning that his home state of Georgia now issues over $800 million in tax credits to film production companies every year. His email address is jcbradbury@kennesaw.edu.

O’Brien and Lane’s reply to this article
Go to archive of Comments section
Go to March 2020 issue

Discuss this article at Journaltalk:
https://journaltalk.net/articles/6000/
Reply to Bradbury: Effects of Economic Incentives in the American Film Industry

Nina F. O’Brien¹ and Christianne J. Lane²

Our 2018 study “Effects of Economic Incentives in the American Film Industry: An Ecological Approach” examined whether and to what degree state film incentives in the United States contributed to film production, industry specific employment, and the creation of new firms serving the motion picture industry. As we wrote: “The range and diversity of film incentive programmes present challenges for programme evaluation, and makes comparison across states particularly difficult. A longitudinal examination of the national industry helps to identify trends and determine what role ecological factors like dominance and diversity may play in the effectiveness of incentives promoting project-based industries” (O’Brien and Lane 2018, 868).

The crux of our findings that are relevant to this response is that the impact of incentives are relatively small for all dependent variables. We found that simply offering an incentive of any value has significant effects on outcomes, but that offering more money did not promote better outcomes, except for filming, the shortest-term gain and one which may not promote longer-term local economic development. This is important for policy makers, who may find that small incentives are just as effective as large ones in terms of promoting filming activity, employment in filmmaking, or the establishment of film-specific firms.

These results do not run “contrary” to academic consensus, nor do we quarrel with J. C. Bradbury’s claim that our paper “does not present itself as providing

---

¹. California State University, Los Angeles, Los Angeles, CA 90032.
². University of Southern California, Los Angeles, CA 90007.
strong support for incentive programs” (Bradbury 2020, 57). Indeed, our paper recommends that states carefully reconsider and tailor their programs, paying closer attention to local diversity and ‘fly-in fly-out production.’

The measurement of incentives

The crux of Bradbury’s critique seems to be that he either disagrees with, or fails to understand, our use of legislative budgetary allocation rather than actual spend to operationalize incentives. Bradbury incorrectly and repeatedly states that our measure for incentives documents state spending on film incentives. As we discuss in the published manuscript in Regional Studies: “The size of an incentive is the allocation a state’s legislature earmarks for a programme, specified in the state budget. … the variable Incentives(millions) reflects the US dollar amount a state committed to film incentive programmes per year (in millions)” (O’Brien and Lane 2018, 868).

In a rich and robust field of inquiry, it is commonplace as well as advantageous for researchers to work through different measurement possibilities. In Bradbury’s own words, “studies that present their data on film incentives often differ slightly from each other in their designations” (Bradbury 2020, 58).

In previous responses to Bradbury,3 we assumed this confusion between spend and allocation was a simple inattention to the definitional terms of our variable specification and made the clarification. However, by this, the third iteration of the critique, we can only infer that Bradbury understands perfectly well and yet chooses to persist in his error. Happily, Bradbury’s repetition of this false conflation does not change the facts.

The measurement of uncapped incentives

Distinguishing between allocation and spend goes most of the way toward addressing Bradbury’s second concern: our decision to value uncapped incentives at $300 million. The highest incentive allocation in our dataset was $274.9 million (offered by New Mexico in 2012).

3. Contrary to the claim that Bradbury publishes this commentary in order to “elicit a response” (Bradbury 2020, 58), we have addressed these concerns twice already: first in an exchange moderated informally through the editorial board of Regional Studies, where our article is published, and a second time through the formal peer-review process of a commentary strikingly similar to that published here (rejected by Regional Studies on the grounds that it was unconstructive and un-collegial, but nonetheless subsequently self-published on SSRN and again resurrected for Econ Journal Watch).
In terms of the ways incentives signal munificence, an *uncapped incentive* signals that there is no pre-set limit on incentives offered by the state. The message this sends to filmmakers is that uncapped incentives are theoretically larger and more attractive. However, as explained in the published work, we did not wish to *overestimate* the value of an uncapped incentive and so worked with film officials and producers to determine a reasonable estimate value.

An increase of ten percent over the highest allocation addresses the necessity of valuing uncapped incentives more highly than capped ones while applying appropriate constraint. Assuming a simple value for uncapped incentives is both necessary and appropriate, keeping in mind that here, as throughout the study, the incentive variable reflects a value on uncapped *allocation, not spend.*

**The exclusion of Iowa**

Bradbury describes the exclusion of data for the state of Iowa as “odd” and “unjustified” (Bradbury 2020, 62, 63). However, during data collection, Iowa was embroiled in an audit following reports of mismanagement and potential fraud within their incentive program. As a result of those investigations, it was not possible to verify the numbers we found with the state film and economic development office. We made the choice to exclude these data, which represent 2 cases, or 0.3 percent of the sample. Given our goal of understanding how incentives operate across the United States generally, excluding Iowa remains the appropriate choice.

**The use of NAICS code 5121**

Perhaps to distinguish it from the commentary rejected by *Regional Studies,* in this latest critique Bradbury raises a new claim that our use of NAICS code 5121 is incorrect. NAICS 5121 is an aggregate category which includes a number of subcategories pertaining to motion picture and video industries. Among these subcategories are Motion Picture and Video Production (512110), Motion Picture and Video Distribution (512120), Teleproduction and Other Postproduction Services (512191), and Other Motion Picture and Video Industries (film labs, libraries, and storage facilities).

All of these subcategories describe business concerns which are vital to our arguments about diversity. Briefly, an enduring film production community consists not only of producers, but also of the many below-the-line and behind-the-scenes firms which support production through the provision of goods and
services, as well as those working in postproduction capacities such as editing, titling, animation, effects, etc. For these reasons 5121 is the appropriate NAICS category.

It is correct that 5121 also includes exhibitors, namely Motion Picture Theaters (512131) and Drive-in Motion Picture Theaters (512132), and that these organizational types are not the targets of incentive legislation. This is an interesting and helpful point and something we will keep in mind as we continue our work.

**In sum**

Overall, it seems Bradbury would simply have liked it better if we had made different choices in conducting our study. We humbly suggest that this affords him excellent avenues for further research. In closing his commentary, Bradbury cautions against relying on our work “for evaluating film incentives as a policy tool” (Bradbury 2020, 63). Indeed, we concur that any study, ours included, be considered in the larger context of an expanding body of work which collectively builds knowledge and informs policy.

We wish the author of the commentary all the best in making his own contributions to the fascinating, complex, and expanding field of inquiry around motion picture tax incentives.

**References**


About the Authors

Nina F. O’Brien, Ph.D., is Associate Professor of Management and Communication Studies at California State University, Los Angeles. Her research focuses on organizational and communication practices and systems which characterize creative industries like film, television and new media production, as well as the not-for-profit and civil society sectors. Her email address is nobrien2@calstatela.edu.

Christanne J. Lane, Ph.D., is adjunct faculty at University of Southern California and a Business Analytic Senior Advisor for Cigna. She has designed, analyzed, and disseminated over 800 clinical studies from a wide range of areas including diabetes and metabolic syndrome, stroke rehabilitation, longitudinal cognitive aging, and autism. She has an M.S. in Biostatistics and a Ph.D. in Quantitative Psychology, giving her a broad range of statistical expertise which she applies to solve specific research questions with appropriate rigorous methods. Her email address is clane@usc.edu.

Discuss this article at Journaltalk: https://journaltalk.net/articles/6001/

Farley Grubb

Ronald Michener (2019a) takes issue with my approach to and analysis of colonial New Jersey’s paper money (Grubb 2016a). He objects to how I calculate the money’s asset present value when that money is hypothesized to be zero-coupon and interest-bearing bonds. He objects to how I calculate the money’s market exchange value. He questions my exchange rate data on colonial New Jersey’s paper money, and he objects to the econometric treatment I apply to my model of that money’s performance. He, however, presents no new data constructions, nor presents any new explanations for the patterns in the data, nor does he present any new hypotheses or do any hypothesis testing to account for the level and movement in the value of colonial New Jersey’s paper money.

The back story

Readers should be aware that the exchange here between Michener and myself is not a one-off affair but has gone on for approximately 20 years. It is a one-sided war: I write original research papers, and Michener attacks them. Michener’s comment here is his sixth published comment on my research that has appeared

1. University of Delaware, Newark, DE 19716.
in a scholarly journal (the other five being: Michener 2018; 2019b; Michener and Wright 2005; 2006a; b)—with four of the six appearing in *Econ Journal Watch*. These comments were on my original research articles (Grubb 2003; 2004; 2018a; Celia and Grubb 2016), and in response to Michener’s five prior comments I have published five rejoinders (Grubb 2005; 2006a; b; 2018b; 2019c).

If I am not mistaken, Michener’s six published comments on my research represent a majority of Michener’s published papers in scholarly journals over the last 20 years. And if I’m not mistaken, there has not been an original research paper on colonial or Revolutionary era paper money that I have submitted to or published in a scholarly journal over the last 20 years that Michener has not rejected as a referee and/or written and submitted a comment on—referee reports and comment submissions that have been often disregarded by editors. An earlier incarnation of Michener’s comment here (2019a) was rejected by the editors of the journal publishing my paper on which Michener’s comment is based (Grubb 2016a).

Besides what has already been cited above, my original research papers on colonial and Revolutionary era paper money that I have submitted to or published in scholarly journals that I suspect Michener also rejected as a referee or failed to get his comments published after my paper was published include: Grubb 2006c; 2008; 2012; 2015; 2016b; c; 2017; 2019a; b; and Cutsail and Grubb 2017; 2019. I have never submitted or written a comment on any original research paper published by Michener. Michener and I were Ph.D. students in economics at approximately the same time at the University of Chicago. While readers might conclude that Michener’s 20-year obsession with me is unnatural, unseemly, and weirdly personal, I could not possibly comment.

**Preliminaries**

Michener’s (2019a) comment is over twice as long as my original publication (Grubb 2016a) on which Michener is commenting. It addresses much that is not in that article—most of which Michener has published before. Regarding my article, he repeats claims he has published before, and repeats them without taking into consideration the replies that cast doubt on the value of his claims. I will not address his review of the field, his comments on the research of other scholars, or his dislike of cliometrics and econometrics—two terms he mistakenly equates.² I

---

². Econometrics are statistical application methods for testing hypotheses in economic theory. Whereas, cliometrics is the application of explicit economic theory and quantitative data, not always statistical, to historical issues. While econometrics are often used in cliometric studies, econometric applications are
will also not repeat my published assessment of Michener’s claims regarding my research, but only summarize them and cite where the reader can read my prior published replies in full.

Variable construction

Regarding colonial New Jersey’s paper money, Michener (2019a, 192–193) claims I incorrectly constructed that money’s asset present value (APV). He says I constructed it as an average utility measure. He says I should have constructed it as a marginal utility measure, and thus I am erroneously inflating its value. He also claims I incorrectly constructed that money’s market exchange value (MEV), erroneously deflating its value by deducting the transaction, transportation, and time costs involved in executing the exchange through to consummation. He claims the result of my erroneously inflating APV and deflating MEV is the striking coincidence of levels for the two series (see Grubb 2016a, 1223, Figure 2). He, however, does no recalculation to show what he thinks the values should be or show whether such changes would alter the coincidence in APV and MEV levels.

Regarding the APV construction, nowhere do I say I constructed APV as an average utility measure. Michener made that up. In addition, Michener commits a fundamental error in microeconomic theory. He conflates average utility with the average of the marginal utilities, and conflates marginal utility with the marginal of the marginal utilities. This is explained in detail in my previously published reply to Michener (Grubb 2019c, 401–402). My construction of APV stands as correct regarding levels.

Regarding the MEV construction, Michener himself (2019a, 185–186) claims that exchange rates need to be adjusted for transaction and risk costs to get the effective exchange rate. I even use material Michener cites in his past publications to, in part, justify the average adjustment I make. For Michener to say in the same comment both that such adjustments to the exchange rate data to get my MEV measure should not be done (Michener 2019a, 193), and that such adjustments are necessary to get effective exchange rates (ibid., 185–186), should tell the reader something. My construction of MEV stands as correct regarding levels.

Finally, Michener (2019a, 210) claims I constructed APV as a function of the current paper money supply (M), thus making M and APV mechanically related. This in turn invalidates any estimated relationship between M and any variable with APV in it. He is disputing my APV analysis because he does not want anyone neither necessary nor sufficient for a study to be sound cliometrics.
to think about colonial paper monies in asset terms. Michener wants everyone to think of colonial paper monies only as pure fiat currencies.  

Michener leaps to this conclusion by looking at my formula for APV and seeing M in the denominator of that equation (Grubb 2016a, 1224). Michener, however, does not evaluate the complete formula. All M is doing in the formula is weighting the individual yearly redemption contributions (RED) across the window of redemption years (T). It turns yearly raw RED amounts into their percentage contributions to the time-discounting calculation. For any given M, APV can take on any value from zero to 100 percent of face value depending on how the legislature sets and executes the length of the redemption window (T) and where, within that window, yearly redemption amounts (RED) are lumped. Therefore, to assert that APV and M are mechanically related, you need to show that M and T are mechanically related, and that M and RED are mechanically related. Michener does not do this. Even a cursory look at how New Jersey structured its paper money redemptions over time indicates that such mechanical connections are not there (Grubb 2015; 2016c). I explain this in more detail in a prior published reply to Michener (Grubb 2019c, 402–403). Therefore, APV is not constructed in such a way that it is mechanically or serially correlated with M, and so Michener’s (2019a, 209–211) objection to my econometric specification in my

3. This view also shows up in Michener’s assessment of Federal Reserve notes as toilet paper (Michener 2019a, 192 n.10). Therein, Michener states that “Federal Reserve notes have commodity value as toilet paper, but they do not fluctuate in value as the price of toilet paper changes.” This appears to be his only effort to dismiss my model of money in equations (1) and (2) of Grubb (2016a, 1217–1218) wherein I decompose the market value of the money ‘thing’ into its real commodity or asset value when not used as money and its transaction premium. Setting aside the fact that if the price of toilet paper tracks the general price index then Michener’s claim above is problematic, and setting aside the fact that no one has seen anyone use Federal Reserve notes as toilet paper, Michener’s statement shows that he does not understand my model of money because his example actually proves the efficacy of my model. Federal Reserve notes are considered a fiat currency. If you take a Federal Reserve note to the Fed and ask for its real non-money value, they will just give you another Federal Reserve note of equal denomination. Its market value is overwhelmingly tied to its transaction premium and not its real commodity value when not a money. That a fraction of a percent of its market value is real commodity value (toilet paper) means that fluctuations in its value are driven by fluctuations in its transaction premium and imperceptibly affected by fluctuations in the market value portion that is comprised of real commodity value. This result is precisely what my model captures. It distinguishes between fiat and asset monies and measures the extent of a money’s fiat-ness.

The proper analogy to colonial paper money is not a current Federal Reserve note, but a U.S. saving bond if that bond was tradeable. The value of a saving bond at any point in time before its maturity date is not its face value but is its real asset present value in Federal Reserve notes when time-discounted from its maturity date. As such, its current market value is predominately determined by its APV, just as colonial paper monies were. If saving bonds were tradable, then some small transaction premium contribution to its current market value might exist depending on its usage in trade compared with other media of exchange. The structure of ‘money’ matters. Michener chooses to ignore that, as well as all the colonial laws dealing with the structure and timing of face-value redemption of colonial bills, but then he admits that colonial bills were “called in and burned” like bonds (Michener 2019a, 190).
Table 3 (Grubb 2016a, 1229) is erroneous.

My Table 3 (Grubb 2016a, 1229) is estimating the determinants of the transaction premium attached to colonial New Jersey’s paper money, namely the gap between MEV and APV. Michener redoes this regression by dropping APV from the construction of the transaction premium, leaving only the money’s market exchange value (MEV) as the dependent variable (Michener 2019a, 210). Michener asserts that his specification is measuring the “Determinants of the transaction premium.” Yet he no longer has a transaction premium measure as the dependent variable. Thus, his claim is erroneous. He is estimating something else, not the transaction premium. Michener (ibid.) justifies his removal of APV from the dependent variable construction because APV and M (one of the independent variables) are correlated in the raw data—claiming that they are mechanically or serially related in construction. However, his assertion that these two variables are mechanically or serially linked in construction was shown above to be erroneous. This leaves only his simple correlation coefficient between APV and M in the raw data as his objection. However, if only variables that are completely uncorrelated are allowed to be regressed against each other, then all econometrics collapses to irrelevance. As such, Michener’s argument for rejecting what I did is spurious.

Finally, Michener’s specification and results from redoing my Table 3 as his Table 8 (Michener 2019a, 210), where he shows no statistical relationship between MEV and M, is old news. It has already been estimated and published by me. MEV can be used as an inverse proxy for prices in a quantity-theory-of-money estimation. I have estimated that effect elsewhere and found little relationship between MEV and M (Cutsail and Grubb 2017; Grubb 2016b, 182; 2019a). Michener makes no new discoveries or contributions here.

**Econometrics**

Michener claims that the econometrics I report in my Table 2 (Grubb 2016a, 1227) are wrong, and when he redoes them his way he gets nonsense—wanting the reader to conclude that my whole approach is nonsense. He claims (2019a, 180, 192–193) that “The heart of his [Grubb’s] paper lies in using econometrics,” and that “Grubb…grounded his analysis in econometrics”—implying that my whole approach rises and falls with the success or failure of the econometrics. Michener also claims that his alternative econometric specifications when using uncorrected data—data and specifications that yield no statistically significant results on the APV variable—are preferable.

If I’m not mistaken, Michener has never published applied econometric estimates in any of his prior work on colonial paper money, except in his comment
on my analysis of colonial Virginia’s paper money just published (Michener 2019b), work by him that spans over 30 years of research on the topic. His objections here (2019a) are the same as in that prior (2019b) comment on my research. In his comment here, as well as in past papers, Michener has expressed a disdain for applied econometrics, holding that it is a methodology that has no value and should not be used.

Nowhere do I say the econometric applications in Table 2 (Grubb 2016a, 1227) are the “heart” of my paper—as Michener (2019a, 192) claims it is. Michener made that up. The ‘heart’ of my paper is equations (1) and (2) and the data as displayed in Figures 2 and 3 (Grubb 2016a, 1217–1218, 1223, 1225). As such, Michener is trying to distract the reader away from the core of my paper and into Michener’s econometric hash. Nothing Michener does to my econometric analysis changes the results derived from equations (1) and (2) and displayed in Figures 2 and 3. The econometric results could be entirely eliminated and the results displayed in Figures 2 and 3 would still stand. Finally, if you use Michener’s uncorrected data and redraw Figures 2 and 3 (something Michener fails to do), nothing perceptible changes. The coincidence between the levels and movement of APV and MEV for New Jersey’s paper money is unmistakable and undeniable. Nothing else, and certainly nothing else offered by Michener, tracks the level and movement of MEV as does my modeled APV.

The data for colonial New Jersey’s paper money regime span only 67 years, and the key variables, namely the market exchange value (MEV) and asset present value (APV), are measured with error. Applying time series econometrics to such a short data span using variables measured with error can only be illustrative and ancillary at best. Trustworthy standard errors are difficult to generate under these circumstances. A person would be foolish to “ground his analysis in econometrics,” or make econometrics the “heart” of his analysis, or dwell exclusively on statistical significance under these conditions. The same can be said for anyone who would make such an econometric exercise the core of their comment. The econometric exercises in Table 2 are only illustrative of and ancillary to the core results displayed in Figures 2 and 3 (Grubb 2016a, 1223, 1225, 1227).

Michener wants to redo the econometrics in my Table 2 (Grubb 2016a, 1227) in order to produce no statistically significant relationship between MEV and APV because he does not want anyone to think of colonial paper money in asset terms. Doing so is easy given the nature of time series econometrics. One way to do this is to load the right-hand side of the regression with time variables, which Michener does in his Tables 2 and 3 (Michener 2019a, 197–198), thereby diluting the influence of the non-time variables. Doing so also violates the model

4. Any univariate time series can be closely approximated by a function of continuous and discrete time
being estimated. I am estimating a decomposition model. It is an identity (see the equations in Grubb 2016a, 1217–1218; 2016b, 163; 2018a, 127). You cannot add time variables to it without violating the model. It would be like adding a time trend to purchasing power parity, which is also an identity. Doing so invalidates the purchasing power parity model.

Another simple way to produce no statistically significant relationship between MEV and APV would be to add erroneous observations to the data set. Doing so would increase the standard errors and drive the estimated coefficients toward statistical insignificance. Now doing so would be an obvious trick that most readers would see through and reject. But what Michener does by refusing to correct the exchange rate data is the same thing. By not removing errors in the data, and assuming such errors are random, the standard errors are increased thus driving the estimated coefficients toward statistical insignificance.

Michener’s primary econometric objection boils down to whether the exchange-rate data are correct or whether there are errors in it that need correcting, and then whether my corrections are justified. I will turn to that next. While Michener has been doing research on colonial paper money for over 30 years he has never indicated that he has previously looked into the exchange rate data, verified its construction and sources, or attempted to correct any errors therein.

Michener refused to track down and examine the sources in his library cited in Grubb (2016a, 1221) that I used to correct the exchange rate data. He demanded that I do that for him and show him exactly how I made the corrections. I refused and said I was not Michener’s research assistant. All the sources were published and were in Michener’s university library. I saw no reason to have to explain to Michener how to use a library, how to read English, and how to use his hand calculator to do long division—the only techniques needed to figure out what I did.

The editor of this journal then asked if I would track down the sources again in my library, re-verify them, and explain to him what I did. I did so. The editor then passed that information on to Michener—information Michener then used to pre-emptively craft counter arguments to my likely reply to his comment (Michener 2019a, 202, 205, 206). I will repeat and expand on that information in what follows.

variables. Regressing this time series on these time variables can produce a close fit, but it does not mean anything. All it does is re-describe the time series in a different mathematical way.

5. Michener (2019a, 201) claims that “nowhere” did I “specify” these errors but only provided a “huge” list of sources referencing where they might be, implying that it was too difficult for anyone to figure out and verify what I did. I, however, explicitly stated that I made corrections to the years 1739, 1741, and 1762 (Grubb 2016a, 1219). The sources I cite in Table 1 amount to three lines and six items all with explicit pages listed (Grubb 2016a, 1221). The items and pages that correspond to the three years listed for which I made corrections are even fewer. It took under half a day in the library, and only a hand calculator, to figure out what I did regarding data corrections.
Exchange rates

John J. McCusker (1978) reports exchange rates for the monies used by individual American colonies and select European counties, by month, for each year from 1600 through 1775. McCusker’s exchange rate compilation is the definitive source used by scholars for exchange rates between the paper monies used by individual American colonies and English pounds sterling. McCusker’s data compilation was a colossal undertaking for a single individual. It used a massive number of disparate original and secondary sources, sorted the data by month, year, colony, and country, and transformed irregular information as stated in original and secondary sources into a standard format. Given the nature of this data compilation exercise, it would not be surprising to discover some errors in the data reported in McCusker (1978).

Colonial New Jersey’s paper money exchange rate to pounds sterling

McCusker’s (1978, 172–174) exchange rates for colonial New Jersey’s paper money to pounds sterling are supposedly market exchange rates. He presents his rates as £NJ needed to purchase 100£S (£NJ = New Jersey paper pounds; £S = pounds sterling). This is the standard format most historians use when reporting exchange rates for colonial New Jersey. For discussion convenience, I will convert everything to £NJ needed to purchase 1£S from here on.

For colonial New Jersey’s paper pounds after 1723, par (face value) was set at 1.33£NJ = 1£S, or 1£NJ = 0.75£S. The par rate was printed on the face of each New Jersey bill and was the rate used by New Jersey’s provincial government when redeeming its paper money (Grubb 2015, 16, 18; Newman 2008, 249–259). This par rate was also Queen Anne’s 1704 Proclamation rate. New Jersey’s provincial government did not redeem its bills on demand, but only at designated future dates (Grubb 2015; 2016a).

Original sources report exchange rates in three different ways. Historians compiling exchange rates transformed this original source information into £NJ = 1£S.

(a) The most common listing of exchange rates in original sources is as 1£NJ = 1£S, such as 1.5£NJ = 1£S. No transformations are required to present these exchange rates in the way historians typically do. Except for an occasional transcription error (which are known to have occurred), few errors occur with these sources in the McCusker data.

(b) Occasionally, original sources list exchange rates as xx£NJ = yy£S. To get
the exchange rates as historians like to present them they have to do some long division, so that \((xx / yy)_{\text{NJ}} = 1_{\text{S}}\). Occasionally, the historians who do this make long division errors and so report the wrong exchange rate. Considering that most historians who compiled the exchange rates reported in the secondary sources did so in the era before hand calculators, such errors may not be surprising.

(c) Finally, original sources occasionally report exchange rates as the percentage of a £ required to purchase 1 £. Par is 1 £ = 0.75 £, or at par 1 £ trades at 75 percent of 1 £. Thus, if 1 £ trades at 70 percent, then 1 £ = 0.70 £, or 1.4286 £ = 1 £. To get the exchange rates as historians like to present them, historians have to do long division and realize they are dealing with an inverse interpretation from method (a) above. The historians who did these exchange-rate compilations, especially considering that most were done pre-1979, occasionally made errors in doing the long division and in realizing the inverse reporting in the original source, and thus sometimes reported the wrong exchange rates.

**My corrections to the McCusker data on colonial New Jersey exchange rates**

**1739**

For 1739, McCusker (1978, 172) made an error when he reported an exchange rate of 1.70 £ = 1 £ for May of that year, which he derived from Lewis Morris (1852, 49). Morris was Governor of New Jersey. In a letter written on 26 May 1739 from Perth Amboy to the Lords of Trade in England, Morris said that 1000 £ = 550 £, which is the same as 1.818 £ = 1 £. Apparently, McCusker occasionally had problems doing long division. This same error shows up in Richard Lester’s book (1970/1939, 127). McCusker may have just copied Lester’s long division error here for his entry for May 1739. I changed McCusker’s rate for May 1739 to the correct rate of 1.818 £ = 1 £.

**1741**

For 1741, McCusker (1978, 172) lists one rate for January, one rate for April, one rate for June, and one rate for August. Those rates were 1.50 £ = 1 £, 1.50 £ = 1 £, 1.25 £ = 1 £, and 1.40 £ = 1 £, respectively. The only published source McCusker (1978, 174) lists for where he took these rates for the year 1741 is Morris (1852, 133). Yet, Morris lists rates for 1741 in his August 1741 letter without giving exact months for each rate. Morris merely references a “2 to 3 months time” between the rates he lists (Morris 1852, 134). Morris also lists the rates on page 134,
not on page 133 as cited by McCusker.

McCusker also reports the average of these rates for 1741 as being $1.425\,\text{NJ} = 1\,\text{S}$. The actual average for his four reported rates is $1.4125\,\text{NJ} = 1\,\text{S}$. It looks like McCusker had a transcription error in writing down the average, mistakenly dropping the “1” in the true average so $1.4125\,\text{NJ} = 1\,\text{S}$ was erroneously recorded as $1.425\,\text{NJ} = 1\,\text{S}$. Alternatively, there may have been several rates used in a given month for the months when rates were reported and including them in the average might explain why the average reported is not the simple average of his four reported rates for 1741. This possibility, however, is inconsistent with how McCusker arrives at yearly averages for other years (see the discussion of 1762 below). Thus, it is likely that McCusker just made a transcription error here in reporting the average for 1741.

For January 1741, McCusker reports an exchange rate of $1.50\,\text{NJ} = 1\,\text{S}$. This rate does not come from any published sources he cites for where he says he took the data for 1741. So where is it from? The source of this rate appears to be Donald Kemmerer (1956, 119) where Kemmerer reports that the Boston Evening Post of 12 January 1741 listed New Jersey’s paper money exchange rate as $1.60\,\text{NJ} = 1\,\text{S}$. Thus, McCusker’s reported rate for January appears to be a recording error. I changed it to the rate as listed in Kemmerer (1956, 119) being $1.60\,\text{NJ} = 1\,\text{S}$ as opposed to McCusker’s $1.50\,\text{NJ} = 1\,\text{S}$.

For April, June, and August of 1741, McCusker reports exchange rates of $1.50\,\text{NJ} = 1\,\text{S}$, $1.25\,\text{NJ} = 1\,\text{S}$, and $1.40\,\text{NJ} = 1\,\text{S}$, respectively. These three rates do not come from any published sources he cites for where he says he took the data for 1741. So where do they come from? The source of these rates appears to be Kemmerer (1956, 120) where Kemmerer reports, “During the preparation for the West Indian expedition, while supplies were being accumulated in quantity, the supply of English bills of exchange was such, and the demand for specie so great that the New Jersey exchange rose sharply. The number of paper pounds needed to buy £100 sterling went from 170 to 150, then to 125 at the peak, after which the rate dropped to 140, 150 and finally back to 170 again.” Kemmerer also says that, “Morris’ explanation [for this pattern] was that the specie had left the province.” Kemmerer does not give the months for the rates he lists. It appears that McCusker may have just taken the middle three rates as listed by Kemmerer above, $1.50\,\text{NJ} = 1\,\text{S}$, $1.25\,\text{NJ} = 1\,\text{S}$, and $1.40\,\text{NJ} = 1\,\text{S}$, and arbitrarily placed them in April, June, and August of 1741, respectively.

Kemmerer (1956, 120) lists the source for the New Jersey exchange rates to pounds sterling that he reports for 1741 as being N.J. Archives, VI, 134. This source is Documents Relating to the Colonial History of the State of New Jersey, VI (1882, 130–137). This is the same exact source as Morris (1852, 132–137). McCusker cites Morris (1852, 133) as his source for the 1741 exchange rates. So both Kemmerer
and McCusker are citing the same source for their 1741 exchange rates. This source is a letter written by New Jersey Governor Lewis Morris from Trenton to the Lords of Trade in England and dated 16 August 1741. Therein, Morris never mentions any of the exchange rates that Kemmerer and McCusker report for 1741, at least not in the form that Kemmerer and McCusker report them. Morris makes no statement that in 1741 exchange rates were \(1.50£_{\text{NJ}} = 1£_{\text{S}}\), \(1.25£_{\text{NJ}} = 1£_{\text{S}}\), or \(1.40£_{\text{NJ}} = 1£_{\text{S}}\), or that, as Kemmerer (1956, 120) stated it, “the number of New Jersey paper pounds needed to buy £ 100 sterling went from 170 to 150, then to 125 at the peak, after which the rate dropped to 140, 150 and finally back to 170 again.” Let me repeat that—there is no statement in the original source cited by Kemmerer and McCusker for the exchange rates these two scholars report. Kemmerer and McCusker transformed what was in the original source into their reported exchange rates, a transformation that turns out to be math errors on their part. Kemmerer erroneously interpreted the statement by the original writer about these exchange ratios and McCusker appears to have just copied Kemmerer’s error.

Two things indicate that something is fishy in how Kemmerer and McCusker transformed Morris’ statements into their exchanges rates. First, Kemmerer’s reasoning for the movement in New Jersey’s exchange rates in 1741 gets the logic backwards. Kemmerer (1956, 119–120) claims that specie got scarcer, therefore, the number of New Jersey bills needed to buy specie (pounds sterling) got fewer! His reasoning amounts to a basic error in supply and demand (Kemmerer has demand curves sloping up and supply curves sloping down here). He puzzlingly refers to the low number reported, “125,” as the “peak” number of New Jersey pounds needed to purchase 100 pounds sterling. It appears that Kemmerer just mistakenly calculated up as down and down as up when speaking about the ratio of New Jersey pounds to pounds sterling here.

Second, in all the monthly exchange rates reported by McCusker (1978, 172–174) for New Jersey between 1703 and 1775 the only rate reported above the par rate of \(1.33£_{\text{NJ}} = 1£_{\text{S}}\), namely a number for £_{\text{NJ}} lower than \(1.33£_{\text{NJ}} = 1£_{\text{S}}\), is for June 1741, out of a total of 44 monthly rates reported. This makes McCusker’s \(1.25£_{\text{NJ}} = 1£_{\text{S}}\) rate for June 1741 a dubious entry. An above par (face value) rate is irrational in that New Jersey pounds did not have face value redemption or payoff dates in specie equivalents until well into the future. No one would pay more now for a bill that would pay less 12 years from now. It would be irrational non-maximizing behavior.

6. This same anomaly shows up in McCusker’s data for Pennsylvania. Out of 568 monthly exchange rates reported for the Pennsylvania paper pound from 1720 through 1775, only one is reported as above par, and that one is for June 1741. By contrast, out of a total of 99 monthly exchange rates reported for the Maryland paper pound from 1734 through 1764, no rates are reported above par, including the rate for June of 1741. In addition, out of 527 monthly exchange rates reported for the New York paper pound from
My guess is that McCusker simply copied Kemmerer’s reported exchange rates for 1741 and then cited Kemmerer’s original source without looking deeper into it. He then arbitrarily chose months in 1741 to place these individual rates in—based on Morris’s August-dated letter and Morris’s 2 to 3 month statement regarding the other rates he mentioned in his August letter.

So what does Morris (1852, 132–137) actually say about the exchange rates between New Jersey paper pounds and pounds sterling for 1741? As Governor of New Jersey, Morris was explaining to the Lords of Trade in England the status of New Jersey’s paper pounds in 1741 in part to explain New Jersey’s contribution to funding the Crown’s expedition against the Spanish in the West Indies. New Jersey had printed 2,000 £NJ in paper money to pay “for victualling and transporting the troops raised in this Colony sent against the Spaniards.” The problem to be explained to the Lords of Trade was that 2,000 £NJ did not buy 2,000 £S worth of provisions and troops, not just because par exchange was $1_F^{\$}=0.75F_S$, but because specie (pounds sterling) was scarce due to the demands of the war and that meant that $F_{NJ}<0.75F_S$ in the marketplace. Foreign expeditions were funded in specie. New Jersey paper pounds had to be exchanged for specie, and with specie acutely scarce due to wartime demands it took more $F_{NJ}$ to buy $F_S$ than normal. This was an acute crisis over the first half of 1741. Morris pointed out to the Lords of Trade that New Jersey had not altered its currency, only that the exchange rate was free to move with the scarcity of specie and the shifting trade balances between New Jersey and England.

Morris (1852, 133) said that he had discussed this with the New Jersey assembly in January and April of 1741 and that the assembly pointed out that there had been no alteration in its currency, but only “that bills of exchange [pounds sterling] had got to a higher rate than they had been, and that the Exportation being Encreas’d, the course of Exchange had fallen to 50 pr cent, & that the Increase of the Exportation was the chief cause thereof.”

In other words, bills of exchange (pounds sterling) had become more expensive to buy, namely it took more $F_{NJ}$ to buy $F_S$ than before such that exchange had fallen to 50 percent. Thus, $F_{NJ}$ now got you only 50 percent of $F_S$, i.e.,

1709 through 1775, no rates are reported above par, including the rate for June of 1741. And out of 304 monthly exchange rates reported for the Massachusetts paper pound from 1720 through 1749, no rates are reported above par, including the rate for June of 1741. Plus, out of 43 monthly exchange rates reported for the North Carolina paper pound from 1715 through 1774, no rates are reported above par, including the rates reported in 1741 (McCusker 1978, 140–141, 163–165, 184–185, 202–203, 217–219). It appears that McCusker made the same source interpretation error for Pennsylvania as he made for New Jersey regarding the June 1741 exchange rate. This error does not show up in the other colonies’ exchange rates because different original and secondary sources were used for those other colonies by McCusker.

7. The assembly said that exchange had “fallen.” There is no way to interpret the assembly’s “50 pr cent” exchange rate statement as $1.5F_{NJ}=F_S$ and be consist with their use of the word “fallen” as referring to
1\text{£}_{\text{NJ}} = 0.50\text{£}_S\) as opposed to the par amount of 75 percent of 1\text{£}_S. Doing a little long division yields, \(2.00\text{£}_{\text{NJ}} = 1\text{£}_S\). Morris (1852, 133) goes on to point out that the “owner of the bill [a New Jersey paper pound] could not have purchas’d so much silver and gold for his 20 shillings \([1\text{£}_{\text{NJ}}]\), … as when silver pass’d at 6’ 10\text{d}'\) per ounce,—which was the nominall [face] value his bill was struck at.” Morris went on to say that he thought the reason for this was not that British factors were exporting too many bills of exchange, but instead was due to an acute scarcity of specie caused by the war with Spain.

Morris then states that:

…it seems plain to me that if a guinea [an English gold coin] was at any time before that current at 30 shillings in bills of credit [New Jersey paper pounds], that, when it was current at 5 pounds in ye same bills, it required 5 pounds to purchase that guinea which 30 shillings of the same currency or bills would have done before, w’ch must make those bills (whatever nominall vallue was Impress’d upon them) of so much less reall vallue than they were before.\(^8\)

The falling of Exchange from 70 to 50 and after that so low as even to 25 p’ cent in 2 or 3 months time, and its rise again to 40, and rising, seems to be too sudden to be owing to the increase of Exports [of bills of exchange] as our Council [the New Jersey Assembly] says, or the Contrary; and is said to be chargeable to another account viz. the want of specie [specie scarcity caused by wartime demands]... (Morris 1852, 134)

So how did Kemmerer and McCusker derive their 1741 exchange rates from the above passage? It appears they simply disregarded the first paragraph and then added a “1” to the percentage statements in the second paragraph—assuming that number went in front of \(\text{£}_{\text{NJ}}\) in the exchange rate statements, thus the exchange rate in 1741 would go from 1.70\text{£}_{\text{NJ}} = 1\text{£}_S\) to 1.50\text{£}_{\text{NJ}} = 1\text{£}_S\) to 1.25\text{£}_{\text{NJ}} = 1\text{£}_S\) to 1.40\text{£}_{\text{NJ}} = 1\text{£}_S\) over a 2 to 3 months’ time period over the first half of 1741. This matches the rates reported in Kemmerer and the last three rates for 1741 reported by McCusker.

\(^8\) This paragraph is an initial ballpark estimate used by Morris to illustrate what had happened due to the war panic during the first half of 1741 that he then explains in more detail in his following paragraph. This paragraph is smoking-gun evidence showing that Kemmerer, McCusker, and Michener are wrong in their transformation of the exchange rate statements made by Morris in Morris’ following paragraph. Kemmerer, McCusker, and Michener completely ignore this paragraph in Morris’ letter. There are 20 shillings in a pounds, therefore 30 shillings in New Jersey paper money = 1.5\text{£}_{\text{NJ}}, i.e., 30 / 20. A guinea = 1.05\text{£}_S (McCusker 1978, 11). Using Morris’ actual numbers in this paragraph, Morris is saying that the exchange rate had gone from 1.5\text{£}_{\text{NJ}} = 1.05\text{£}_S\) to 5\text{£}_{\text{NJ}} = 1.05\text{£}_S\) over the relevant period. Doing a little long division yields Morris’ exchange rate going from 1.43\text{£}_{\text{NJ}} = 1\text{£}_S\) to 4.76\text{£}_{\text{NJ}} = 1\text{£}_S\). These rates are clearly within the range of exchange rate movements I use in my corrections to the McCusker data for 1741 when using the complete information Morris gives in his following paragraph, and it is outside the narrow range insisted on by Michener (2019a, 202, 204). See the material linked to footnote 9.
As such, Kemmerer and McCusker made a fundamental error in interpreting and transforming ratios. It is clear from the passages quoted above that what Morris means by “falling of exchange” from 70 percent to 50 and then to a low of 25 is that it takes more and more \( \text{£}_\text{NJ} \) to buy 1\( \text{£}_\text{S} \), not less and less as Kemmerer and McCusker’s transformations yield. Kemmerer and McCusker got it upside down.

So what are the correct exchange rates to be derived from Morris (1852, 134) for 1741? In August 1741, Morris said that there had been an acute exchange rate crisis over the prior 2 to 3 months caused by an acute scarcity of specie money brought on by the war with Spain. He says that a New Jersey paper pound fell from 70 percent of a pound sterling to 50 percent of a pound sterling, to a low of 25 percent of a pound sterling, and then rose back to 40 percent of a pound sterling. Note that Morris never says that a New Jersey paper pound was worth 75 percent or more of a pound sterling. Par was 75 percent and any percentage above 75 percent would be an above-par exchange rate. Morris’s rates are all below-par rates. New Jersey paper pounds were depreciating substantially in the first half of 1741, only to recover some by August. They were still, however, trading at a depreciated rate in August 1741 relative to par.

Assuming that Morris’s last rate that he mentions is for August (the date of his letter), implies that the prior three rates occurred between April and August of 1741, given his “2 to 3 months time” statement. He does not say exactly which months these other rates are from. While Morris gives four rates in this April to August window, McCusker only lists three. Why McCusker did not list all four is unknown. McCusker also arbitrarily places his two pre-August rates in the months he did between April and August. I am only interested in a yearly average, and so the exact month each rate is for is not relevant to my data correction.

Morris’s sequence of four rates from April to August 1741 corresponds to \( 1\text{£}_\text{NJ} = 0.70\text{£}_\text{S}, \ 1\text{£}_\text{NJ} = 0.50\text{£}_\text{S}, \ 1\text{£}_\text{NJ} = 0.25\text{£}_\text{S}, \ and \ 1\text{£}_\text{NJ} = 0.40\text{£}_\text{S} \), respectively. Doing a little long division transforms these rates into how McCusker reports rates, namely into \( 1.43\text{£}_\text{NJ} = 1\text{£}_\text{S}, 2.00\text{£}_\text{NJ} = 1\text{£}_\text{S}, 4.00\text{£}_\text{NJ} = 1\text{£}_\text{S}, \ and \ 2.50\text{£}_\text{NJ} = 1\text{£}_\text{S} \), respectively. I replaced McCusker’s three exchange rates listed for 1741 in April, June, and August, respectively, with these four rates derived from Morris. Adding the rate of \( 1.60\text{£}_\text{NJ} = 1\text{£}_\text{S} \) for January 1741 from the Boston Evening Post yields five rates for 1741, the average of which is \( 2.306\text{£}_\text{NJ} = 1\text{£}_\text{S} \)—which is the rate I used for 1741.

---

McCusker (1978, 173–174) lists two exchange rates for 1762, one for March and one for September, 1.7625£NJ = 1£S and 1.775£NJ = 1£S, respectively. He calculates the average exchange rate for 1762 as 1.7688£NJ = 1£S [(1.7625 + 1.775) / 2]. He cites two published sources for where he took the rates for 1762, namely Kemmerer (1956, 131) and Joseph Sherwood (1851, 147).

Sherwood (1851, 147) lists two rates for all of 1762, namely 1.75£NJ = 1£S and 1.775£NJ = 1£S, without reference to their month of observation. Kemmerer (1956, 131) lists two rates for 1762 and indicates that both are from March of 1762, namely 1.75£NJ = 1£S and 1.775£NJ = 1£S. McCusker appears to have taken Kemmerer’s two rates for March of 1762 and reported their average for March of 1762, namely 1.7625£NJ = 1£S [(1.75 + 1.775) / 2]. Then McCusker appears to have taken Sherwood’s second rate (1.775£NJ = 1£S), considered it an independent observation from that reported in Kemmerer, and arbitrarily placed that rate in September of 1762. As such, McCusker appears to be double counting the 1.775£NJ = 1£S rate, counting this single observation in both March and September of 1762.

This outcome in McCusker also shows that McCusker is reporting the average rate per year as the simple average of the monthly rates he reports and not as the average of the individual rate observations. His average for 1762 is 1.7688 = [(1.7625 + 1.775) / 2], with the first number being the average of the two rates reported in March [(1.75 + 1.775) / 2]. If McCusker had reported the yearly average as the average of all three rates he took from his sources that yearly average would have been 1.7667 = [(1.75 + 1.775 + 1.775) / 3].

I removed the 1.775£NJ = 1£S from September of 1762 from the McCusker data as being an erroneous double counting of that one rate, being already counted in March of 1762. That leaves McCusker reporting only two rates for 1762, 1.75£NJ = 1£S and 1.775£NJ = 1£S, which averages to 1.7625£NJ = 1£S for the year 1762 for these two independent observations.

Sherwood (1851, 137) reports one more exchange rate unnoted either by McCusker or Kemmerer, or anyone else that I have found. In a letter dated 17 August 1762, Sherwood mentions an exchange of 287.34£NJ for 152.46£S. Doing a little long division renders this to be 1.88£NJ = 1£S. However, Sherwood also states that a deduction of 66£ had to be made to execute this exchange. He does not state whether this 66£ was in pounds sterling or in New Jersey paper pounds. A biased-low deduction would be to assume that it is in New Jersey paper pounds, i.e., 66£NJ. In other words, you had to pay an additional 66£NJ to exchange 287.34£NJ.

---

10. There is a source citation error in my prior work, i.e., Grubb (2014, 17; 2016a, 1221), where I only cited page 147 rather than both pages 137 and 147 in Sherwood (1851).
for 152.46£. If this cost is deducted from the exchange transaction then the actual exchange rate is 2.3176£ = 1£, namely [(287.34 + 66) / 152.46].

Thus, I have three exchange rates for 1762, 1.75£ = 1£, 1.775£ = 1£, and 2.3176£ = 1£. The average of these exchange rates is 1.9475£ = 1£ for 1762 which is what I used.¹¹

**Michener’s reasons for rejecting my data corrections**

Michener’s (2019a, 198–205) arguments for rejecting my data corrections include: (a) that my data corrections place exchange rates off trend, (b) that the exchange rate spike I generate with my correction for 1741 is not observed in the exchange rate series for other colonies’ paper monies, (c) that the exchange rates for Virginia’s paper money were above par between 1769 and 1772, and (d) that Governor Morris in 1743 equated exchange at 60 percent with 1.6£ = 1£.

Regarding (a): Throwing out data because it is off-trend means acute monetary shocks, booms, and recessions no longer exist in the data and so cannot be explained. Michener (2019a, 183) himself claims that “wild” swings in exchange rates occurred in colonial America. But here he wants to claim that any off-trend or wild exchange rate observations should be eliminated. Doing such also means that Michener is artificially increasing the error variance in the data and so erroneously driving estimated coefficients toward statistical insignificance (which appears to be his goal).

Regarding (b): New Jersey’s Governor, Lewis Morris, makes it clear that the exchange-rate panic in early 1741 is a localized New Jersey paper money phenomenon. It was directly linked to the 2,000£ New Jersey paper pounds newly emitted to fund New Jersey’s participation in the West Indies expedition against the Spanish—paper pounds that had to be quickly converted to specie in New Jersey to be used for that funding.

Regarding (c): The Virginia rates listed above par are for raw rates unadjusted for the transaction, transportation, and time costs required to consummate the exchange. Once adjustments are made for these costs, the effective exchange rates for Virginia’s paper money are all below par (Grubb 2018a, 130).

Regarding (d): This 1743 statement by Morris does not appear in Morris (1852), but does appear in Morris (1993, 262).¹² Morris’s 1743 statement, however, does not prove much. If exchange at 60 percent means 1£ = 0.6£, then 1.67£ = 1£, which by Michener’s standards of closeness cannot be distinguished from

---

¹¹. See footnote 8 regarding the range that exchange rates can take.
¹². Michener (2019a, 216) mistakenly lists Eugene R. Simmons as the editor of the 1993 Morris volume; the editor’s name is Eugene R. Sheridan.
exchange at 60 percent meaning $1.6\,\text{£}_{\text{NJ}} = 1\,\text{£}_{\text{S}}$. Morris also provides no sequence of exchange rates in his 1743 letter to determine the direction exchange rates are moving relative to specie scarcity. Finally, the smoking-gun evidence provided by Morris in his 1741 letter, evaluated in footnote 8 above, shows that Michener is dead wrong on this issue.

By Michener insisting that no corrections to McCusker’s data are ever needed and that McCusker’s data should be used only as McCusker originally reported it, Michener is in effect repudiating basic mathematical theory and basic economic theory. For Michener, demand curves slope up, financial panics do not exist, and long division no longer holds (so that 1000 divided by 550 now equals 1.70). Michener’s desperate desire to inflate standard errors (by not correcting erroneous data) in order to drive coefficient estimates on APV toward statistical insignificance led him to this sad state.

**References**


Michener, Ronald W., and Robert E. Wright. 2005. State “Currencies” and the Transi-
Farley Grubb is professor of economics at the University of Delaware, NBER research associate, and Financial History Series editor for Routledge, Taylor & Francis Group. He earned his Ph.D. in economics at the University of Chicago. He took multiple graduate courses in economics from each of the following: Gary S. Becker, Robert W. Fogel, Robert E. Lucas, Douglass C. North, and George Stigler. He has published numerous articles in refereed journals and edited volumes on the economic history of colonial and Revolutionary-era America, in particular on contract labor (indentured servitude and convict labor) and on monetary institutions and their performance. His email address is grubbf@udel.edu.
Science on FDA Liberalization: A Response to the Status Quo Process for Medical Treatments

Bartley J. Madden

LINK TO ABSTRACT

In Japan and elsewhere there have been some moves toward liberalization of the approval of new drugs and treatments. An article in the 16 August 2019 issue of Science sounds an alarm against such developments. The authors Douglas Sipp and Margaret Sleeboom-Faulkner deprecate at length the reform proposal of my Free To Choose Medicine (hereafter FTCM; see Madden 2018), which served as a model for Japan’s legislation authorizing conditional approval for medical treatments focused on regenerative medicine (e.g., stem cells). FTCM enables patients, advised by their doctors, to make informed decisions about early access to new drugs after successful completion of initial safety and efficacy trials. Although Sipp and Sleeboom-Faulkner are on solid ground in recognizing some unscientific use of stem cells via direct-to-consumer marketing, they are, I think, wrong in supposing that Japan’s conditional approval will hurt patients by enabling them to gain early access to regenerative medicine treatments. Their judgments are not substantiated by evidence or argumentation about the merits and demerits of the reform; they simply assume that the liberalization is a bad thing. Their article is about 2,900 words in length and appears in Science, one of the top-five most cited journals, so it merits a response. I submitted an earlier version of the present article to Science, but it was dismissively turned away.

Sipp and Sleeboom-Faulkner give a fair amount of attention to my work and its influence, writing:

the key principles adopted in Japan’s deregulation of regenerative medicine were previously outlined by a free-market policy institute, the Illinois-based
Heartland Institute, in the form of a book-length proposal titled *Free to Choose Medicine* (FTCM)… it is an important illustration of how attempts by private policy groups in one country may influence lawmaking in another, with consequences that may be disadvantageous to the publics they are intended to serve. (Sipp and Sleeboom-Faulkner 2019, 645)

In Japan, FTCM found more fertile ground. An early version of the proposal was translated into Japanese by the president of the free-market organization Japanese for Tax Reform, who proceeded to lobby it to members of the Japanese government. (Ibid.)

In 2012, the Japanese Society for Regenerative Medicine began to call for regulatory reforms aimed at accelerating approvals through revisiting clinical testing standards. By 2013, mentions of FTCM began to appear in presentations made by staff in Japan’s drug regulatory agency, the Pharmaceuticals and Medical Devices Agency. The same year, the conditional approvals pathway for regenerative medicine products was introduced. The author of FTCM has since thanked the translator for helping to make his ideas into law in Japan. (Ibid., 646)

Sipp and Sleeboom-Faulkner do not treat liberal economic reasonings with respect, speaking of such reasoning as subservience to “economic agendas, cloaked in the language of serving patients” (2019, 646), as though I do not sincerely believe that liberalization would significantly benefit humankind. They suggest that the idea of a drug lag is used as “a cudgel in the hands of free-market policy organizations” (Ibid., 645).

### System goal

Sipp and Sleeboom-Faulkner (2019, 644) believe the goal of government regulators should be ensuring that any drug accessible to patients is safe and effective. Who wants unsafe and ineffective drugs? However, with such a mindset a government regulatory body tends to steadfastly demand ever more testing while giving low priority to opposing views concerned with associated costs. Such costs include the continuation of delays in accessing beneficial new treatments, and higher prices of new drugs, to meet regulatory testing requirements. Our goal should be *better drugs sooner at lower cost*—and that is the subtitle of my book—where the word ‘drugs’ represents all medical treatments requiring regulatory approval. The fundamental issues are not only statistical problems of determining drug efficacy, but also the following: understanding the overall drugs-to-patients system; avoiding procedures that yield unintended and deleterious consequences; and
identifying and removing constraints that impede better drugs sooner at lower cost.

The Food and Drug Administration (FDA) knows that when it approves a drug that subsequently results in unanticipated adverse side effects, especially deaths, it will face negative media attention and, if many people die, Congressional hearings. Such consequences incentivize the FDA toward preserving or even enhancing their clinical testing requirements. What does not make the nightly news, however, is the invisible graveyard of patients dying from not having had access to very expensive, or delayed, or simply non-existent possible new treatments. The FDA has implemented various programs to accelerate the testing of promising drugs by incrementally changing a single regulatory process (Woodcock and LaVange 2007). Notably absent is competition from an alternative access mechanism that may significantly improve the conventional process (Conko and Madden 2013).

**Optimal regulatory load**

What is the optimal regulatory load for clinical tests and analysis for potential approval of new drugs? Sipp and Sleeboom-Faulkner do not know. No one knows. Nevertheless, Sipp and Sleeboom-Faulkner oppose a FTCM approach that combines informed choice (retaining prescription requirements) and rapid data dissemination and adaptation. Such opposition to liberalization has deep roots. Regulators prefer having simple binary yes/no approval decisions, which exclude the complexities of patient populations comprised of individuals with unique health conditions and unique risk preferences. Sipp and Sleeboom-Faulkner believe that moving away from the so-called gold standard of randomized control trials must necessarily lead to lessening of a firm commitment to product efficacy, but they ignore that sticking to this gold standard is fraught with its own unique ethical concerns (Deaton and Cartwright 2018). Keep in mind that the costs to companies of randomized control trials can have undesirable consequences for how companies select new drug candidates. For example, since cancer survival rates (a key readout for randomized control trials) are far less costly to measure for late-stage cancer patients compared to early-stage cancer patients, all else equal, this motivates companies to allocate resources to late-stage cancer drugs.

To achieve statistical rigor, government regulators responsible for randomized control trials strive for homogeneity of clinical trial patients with minimal concern for the cost of this testing. Such methods do not address the wide diversity of real-world patients as to health characteristics and risk preferences. Moreover, the elimination of choice is justified for today’s patients by assuming either that it is necessary for today’s patients to join clinical trials in which many do not get
a promising new drug (due to having been assigned to a control group), or that these patients are incapable of making health decisions in their own best interest and therefore choice is not a viable option. Should we not revisit these outdated assumptions, which were the centerpiece of legislation empowering the FDA in 1962 (Grove 2011)?

In contrast, Free To Choose Medicine embraces heterogeneous, real-world patients and utilizes rapid technological advancements that continually improve the identification of subsets of patients most likely to favorably respond to a new drug (Khozin et al. 2017). As personalized drugs become increasingly more effective, one can envision ever smaller subsets of patients identified as highly likely to achieve a favorable treatment outcome so that randomized control trials are no longer feasible (Lillie et al. 2011). Such ramifications of personalized drugs spotlight an increasingly significant ethical concern for randomized control trials wherein those patients who are randomly assigned to control groups do not receive the promising new treatment (Stewart et al. 2010), an especially important concern for patients with life-threatening diseases.

A self-adjusting, dynamic system

Sipp and Sleeboom-Faulkner note that the initial stem cell treatments receiving conditional approval and early access in Japan were of questionable value to patients. This is not surprising since conditional approval was in its early startup stage and missing feedback data that helps patients and doctors make informed decisions about the use of early access treatments. Sipp and Sleeboom-Faulkner ignore the long-term benefits to patients and biopharmaceutical researchers from the FTCM focus on rapid posting of treatment results from early access, including patients’ health data, genetic data, and relevant biomarkers—all maintained with patient identity kept confidential. The posting of treatment results is the function of FTCM’s Tradeoff Evaluation Drug Database (TEDD) which provides the biopharmaceutical industry with a treasure trove of data to spur innovation. Contrast TEDD’s open access and real-time availability of data with the status quo process that keeps detailed clinical trial data confidential with only summary data available years after they were generated.

Japan is in the process of implementing their version of TEDD to provide the needed feedback data and move Japan’s conditional approval closer to the comprehensive FTCM proposal. Meanwhile, Athersys, a leading company in stem cell science, has partnered with the Japanese company Healios in order to generate clinical trial data in Japan that may lead to conditional approval for an innovative stem cell treatment for heart attacks. Two relevant questions are: For Japanese
citizens, how important is early access to an innovative alternative to standard treatments for heart attacks? Assuming this stem cell treatment is granted conditional approval, what will likely be the experience for Japanese patients who voluntarily choose early access?

As to importance, ischemic stroke is a leading cause of disability and mortality worldwide, especially so in Japan with its aging population. The approved treatments, tissue plasminogen activator and mechanical thrombectomy, need to be administered quickly, unlike the 36-hour window for Athersys’s Multistem therapy.

Japan’s conditional (FTCM) process involves a fundamental tradeoff that impacts the patient experience. The patient, in consultation with doctors, may opt to forgo the standard assessment of safety and efficacy based on randomized control trial data. The patient has the freedom to choose new drugs five to seven years earlier than waiting for the standard approval process. This becomes more important in an environment of fast-paced innovation. The FTCM focus is on observational (real-world) data and the freedom of patients, advised by their doctors, to decide on one of three choices: (1) standard approval drug, (2) conditional approval drug based on TEDD data currently available, or (3) waiting for additional TEDD data before making a decision. Such liberalization promotes greater drug development. For any possible drug, the issue is not merely one of how long it takes to get it to patients but whether it even comes into existence.

No one knows the optimal level of regulation, especially in a world of fast-paced innovation. Is it not advantageous to allow patients, in consultation with doctors, to make an informed decision? If the Multistem treatment gains conditional approval, expect usage of the treatment to accelerate if early treatment results are superior to standard treatments and vice versa. This constitutes a dynamic, self-adjusting system. With favorable results, thousands of patients will generate both treatment results and patient-specific data enabling subsets of patients to be identified who either do exceptionally well or experience an unfavorable outcome. This better equips patients to make an informed choice based on real-world data whose utility increases with thousands of observations, far greater than the number of patients in a typical randomized control trial.

Sipp and Sleeboom-Faulkner assert that “sacrificing efficacy requirements for speed is unwise” (2019, 645). How do they know that? Apparently, this assertion is due to their skepticism about the ability of patients and doctors, even with access to TEDD information and likely private-sector products to assist the evaluation process, to discern the ‘good’ not-yet-fully-tested drugs from the ‘bad.’ This is an empirical issue. Japan’s forthcoming implementation of their version of TEDD will enable a test of whether the Japanese experience with conditional approval more closely resembles the dynamic, self-adjusting system that benefits
patients, as outlined above, or a chaotic environment with patients and doctors making decisions that fail to provide patient benefits and thus supporting Sipp and Sleeboom-Faulkner’s skepticism (Hudgins 2018).

**Innovation and resource allocation**

Sipp and Sleeboom-Faulkner ignore the role of drug development in the overall drugs-to-patients system. The ‘better drugs’ part of the goal of *better drugs sooner at lower cost* is driven by the speed and effectiveness of a society’s innovation process for developing new drugs. One key driver is rapid dissemination of new data so that scientists throughout the biopharmaceutical industry and other research organizations quickly gain insights leading to new and fruitful hypotheses. Again, the publicly available TEDD data will be a treasure trove for scientists seeking a better cause-and-effect understanding to undergird the development of more and better drugs.

Another key driver is resource allocation. Imagine a world where resources automatically flowed to the most highly skilled scientists, including those with ideas that substantially differ from the existing paradigm of cause-and-effect logic for a particular target disease. Ideal, yes; but this is not today’s world. For biopharmaceutical companies, including startup companies, capital is allocated based on risk-adjusted return on investment. The larger the regulatory costs, delays, and uncertainties, the lower is in large part the anticipated return on investment. With FTCM, new drug revenues can begin five to seven years earlier versus the standard approval process, and with far less expenditures for regulatory costs. Expect substantially more capital invested in drug development due to FTCM, plus heightened competition among companies participating in early access programs. Consider a startup company with exceptionally skilled scientists and a potential breakthrough drug that entails fundamental new thinking, but that is in need of capital funding. Because of its unconventional approach relative to existing FDA-tested drugs for treating the targeted disease, venture capitalists will view future FDA clinical evaluation criteria for late-stage randomized control trials to be difficult to forecast and possibly excessively stringent. Hence the risk for providers of capital will be high. In contrast, such risk is reduced in a FTCM world where drug effectiveness is more swiftly ascertained with real-world data. In this environment, expect more capital to flow to startup companies with new thinking and high scientific skill.

In opposing liberalization, Sipp and Sleeboom-Faulkner suggest that the conventional randomized control trial approach coupled to incremental changes is beyond criticism. It is not. Their article is neither based on logical argument
nor evidence. I think their judgments are irresponsible, and when such judgments appear in an influential journal like *Science*, we must do what we can to bring greater accountability and responsibility to the discussion.

**References**


About the Author

Bartley J. Madden retired as a managing director of Credit Suisse Holt after a career in money management and investment research that included the founding of Callard Madden & Associates. His early research was instrumental in the development of the cash-flow return on investment (CFROI) valuation framework that is used today by money management firms worldwide. He is now an independent researcher and his new book, *Value Creation Principles*, will be published by Wiley in May 2020. His work in public policy has resulted in the Free To Choose Medicine plan, which was developed in journal articles published in *Regulation, Cancer Biotherapy & Radiopharmaceuticals, Medical Hypotheses*, and *Engage*. His email address is bartmadden@yahoo.com and his website is LearningWhatWorks.com.

Go to archive of Comments section
Go to March 2020 issue

Discuss this article at Journaltalk: https://journaltalk.net/articles/6003/
Government-Cheerleading Bias in Money and Banking Textbooks

Nicholas A. Curott¹, Tyler Watts², and Benjamin R. Thrasher³

LINK TO ABSTRACT

Market failure: The failure of the market to recover from a blow by intervention. —Daniel McCloskey

Do undergraduate money and banking textbooks present a thorough and balanced overview of how commercial banks and central banks impact economic stability? We find that the textbooks are generally of high quality, but that they overemphasize the potential instability of unregulated commercial banks and underemphasize the potential for central banks and financial regulation to negatively impact the economy. The systemic slant amounts to a government-cheerleading bias.

We review the six leading undergraduate money and banking textbooks currently in print and offered for adoption by major textbook publishing companies:⁴

---

¹. Ball State University, Muncie, IN 47304. We thank Thomas Hogan, Steven Horwitz, Jeffrey Rogers Hummel, James McClure, Lawrence H. White, four anonymous referees, and especially Kurt Schuler for helpful comments and suggestions. The financial assistance of the Institute for the Study of Political Economy at Ball State University is also gratefully acknowledged.
². Ferris State University, Big Rapids, MI 49307.
³. Undergraduate student, Ball State University, Muncie, IN 47304.
⁴. To test whether universities use textbooks other than the six offered by major textbook publishers we conducted an informal survey. We were able to find online syllabi for money and banking classes for seven of the universities ranked in the top ten undergraduate economics programs by U.S. News and World Report. Of these, two used Hubbard and O’Brien, one Ball, one Cecchetti and Schoenholtz, one Mishkin, and two didn’t use a standard textbook. Of the U.S. News and World Report top ten public universities we found seven syllabi, and all used Mishkin. We also looked at the ten largest universities in Indiana by enrollment and found six with online syllabi. Of these, four used Hubbard and O’Brien, one Cecchetti and Schoenholtz, and one Mishkin. The difference between the prestigious public universities, which are mainly located on the coasts, and the less prestigious Midwest universities in Indiana is interesting and likely reflects underlying ideological differences. Notably, we didn’t find a single instance of a textbook being used that was not one of the six included for review in this paper. This gives us confidence that these six textbooks
We address seven topics related to the roles that commercial and central banks play in macroeconomic stability: (1) the inherent stability of banks, bank runs, and panics; (2) The historical origins of central banks created before the Fed; (3) the fragility of U.S. banks during the National Banking Era and the origins of the Federal Reserve System; (4) U.S. bank panics during the Great Depression; (5) deposit insurance; (6) monetary policy and the Great Recession of 2008–2009; and (7) the performance of the U.S. economy before and after the Federal Reserve Act of 1913. Each of these topics is significant for monetary theory or for regulatory policy.

Our selection of topics is based, first, on the consideration that these topics cover the bulk of historical information offered by the textbooks as applications of the basic economic theory of money and banking, and second, on the concern that what students learn about these topics can strongly influence their underlying worldviews. For each topic we survey the academic literature and then compare it to the information presented in the textbooks. In each case we find the textbook presentations leave out important historical details or present them in a way that systematically favors one view over another. Near the end of the paper, Table 1 summarizes the textbook views. Every textbook provides a narrative biased in favor of government intervention. We ask whether this might be attributable to consensus bias or status quo bias.

Fortunately, the bias on the seven topics can easily be mitigated by adding a few sentences or altering a paragraph here or there. We recommend that on matters where economists disagree, textbooks should, if possible, present the findings of surveys of economists’ views, a practice already adopted in N. Gregory Mankiw’s (2015) Principles of Microeconomics textbook. We also suggest that author(s) candidly communicate their own political leanings.

have a very large market share.
The inherent stability of banks, bank runs, and panics

There is a spectrum of views about the inherent stability of banks in the academic economic literature. On one end of the spectrum there is the view, which we call the ‘inherent fragility hypothesis,’ which holds that banks are inherently run-prone and that a run on one bank is contagious to other banks, causing a panic. On the other end of the spectrum there is the view, which we call the ‘regulatory weakening hypothesis,’ which holds that, but for interventions, private banking is generally auto-corrective and normally stable in the absence of poor management, and that the source of banking panics is ill-conceived government policies, interventions, or regulations (Selgin 1989).

Advocates of the inherent fragility hypothesis often base their case on the highly influential and abstract mathematical model of Douglas Diamond and Philip Dybvig (1983). Diamond and Dybvig’s basic argument is that a run can be self-justifying from the “me first” problem that depositors face. It is in the interest of an individual depositor to run on a bank if he suspects other depositors might run, because if the suspicion proves correct there won’t be enough funds to go around to pay every depositor. Therefore, any event can trigger a run, even if it is otherwise irrelevant to solvency. Whatever makes depositors anticipate a run will in fact cause them to run, validating the anticipation. Banks are inherently unstable because bank runs can be triggered by any random event, such as sunspots appearing on the sun. These runs cause pre-run solvent institutions to incur losses from hasty liquidation of assets and may lead to bankruptcy.6

Advocates of the inherent fragility hypothesis also claim that a run on one bank can create suspicion in the minds of customers at other banks, initiating further runs and causing a contagion (Allen and Gale 2000). Asymmetric information prompts depositors unsure of the soundness of their own bank’s assets to run if they observe a run on another bank. A run on one bank therefore creates a negative externality that spills over onto other banks. Runs that weaken or destroy

---

5. In this paragraph we draw on White’s (1999, Ch. 6) excellent summary of the Diamond and Dybvig model.
6. The “sunspot” theory was formalized by Diamond and Dybvig in 1983, but the core idea is old. Thomas Jefferson (1813), for instance, made similar arguments: “It is said that our paper [currency] is as good as silver, because we may have silver for it at the bank where it issues. This is not true. One, two, or three persons might have it; but a general application would soon exhaust their vaults … Nothing is necessary to effect it but a general alarm; and that may take place whenever the public shall begin to reflect on, and perceive the impossibility that the banks should repay this sum.”
solvent banks constitute a market failure that requires government intervention in the form of a central bank and financial regulation. According to this view, an unregulated banking system would subject the economy to fire-sale losses, unexpected contractions of the money supply caused by recurring bank panics, and frequent recessions.

Advocates of the inherent fragility hypothesis point to the large number of historical bank failures and panics as evidence for the theory. The historical record shows that banking panics were particularly frequent in England in the early nineteenth century and in America during the pre-Federal Reserve period in the nineteenth and early twentieth centuries.

Advocates of the regulatory weakening hypothesis, in contrast, argue that the Diamond and Dybvig model is theoretically unsound, both because there are problems with the formal model and because the model posits relationships that do not accurately reflect the institutional characteristics of real-world banks (Wallace 1988; Dowd 1992a; 1996, Ch. 9; 2000; White 1999, Ch. 6; McCulloch and Yu 1998).7

Advocates of the regulatory weakening hypothesis also argue that the historical evidence shows bank runs are not random or linked to irrelevant events or rumors. Rather, real-world runs happen when depositors receive bad news indicating that their bank might already be pre-run insolvent (Schuler 1992, 30). Depositors run because knowledge indicates that the bank’s net assets are likely already too low to repay all depositors. Bank failures reflect fundamental deterioration in bank health rather than spontaneous panics which cause viable

---

7. Most notably, the conclusion of the Diamond and Dybvig model—that it is in the interest of an individual depositor to run on a bank if he suspects other depositors might run, because if the suspicion proves correct there won’t be enough funds to go around to pay every depositor—assumes that (a) the bank’s short-term assets are less than its liabilities payable on demand (such as demand deposits or on-demand repurchase agreements), and (b) the bank cannot impose a notice of withdrawal clause to delay the redemption of deposits. Critics of Diamond and Dybvig note that if either of these assumptions does not hold then the inherent instability posited by the model does not exist. Therefore, banks can eliminate the incentive for the public to run on solvent banks either by maintaining adequate capital or by introducing an “option clause” that gives banks the option to delay redemption of demand liabilities for a pre-specified length of time. For further details and a review of the literature on contractual solutions to the supposed inherent instability of banks see Selgin and White (1994b, 1727–1730).

8. Historical evidence suggests bank failures are caused by news of a negative shock to banks’ assets (see, e.g., Gorton 1988; Mishkin 1992; Kaufman 1994). These studies provide evidence that bank panics are not random events or self-confirming equilibria in a situation of multiple equilibria, as in the Diamond-Dybvig model. These studies also cast doubt on the existence of contagion effects. However, in many cases the authors argue bank panics can occur due to asymmetric information. If depositors observe a shock that will likely render some banks insolvent, but they cannot observe whether any individual bank is solvent or not, they may run on all banks, both solvent and insolvent. This leaves open the question of whether government intervention is necessary to solve the asymmetric information problem or if it can be solved by private institutions.
banks to fail (Calomiris and Mason 1997). Asymmetric information does not cause runs on solvent banks because large depositors have an incentive to monitor their banks’ investments and because market institutions such as clearing houses and rating agencies can provide this information at low cost, even for small depositors (Selgin and White 1987; Selgin 1993). Runs that destroy solvent banks are so rare as to be negligible in practice and contagion affects are absent in a system that is not hampered by legal restrictions. The regulatory weakening hypothesis also point to the absence of panics consisting of runs on many pre-run solvent banks in countries with relatively free banking systems as evidence that banks are not inherently unstable and panic prone. They argue that the frequently recurring panics in other countries were caused by bad banking laws, regulations, or interventions, such as restrictions on branch banking and bond collateral requirements for banknote issue. Bank panics were common in countries such as the United States, where banks were subjected to these inefficient government restrictions, but were uncommon in countries such as Scotland and Canada, where banks were not.

Advocates of the regulatory weakening hypothesis also point to the absence of panics consisting of runs on many pre-run solvent banks in countries with relatively free banking systems as evidence that banks are not inherently unstable and panic prone. They argue that the frequently recurring panics in other countries were caused by bad banking laws, regulations, or interventions, such as restrictions on branch banking and bond collateral requirements for banknote issue. Bank panics were common in countries such as the United States, where banks were subjected to these inefficient government restrictions, but were uncommon in countries such as Scotland and Canada, where banks were not. Between the limits of inherent fragility and regulatory weakening lies a spectrum of more moderate views that hold banks are potentially fragile along one or

---

9. Historical evidence of contagion effects is mixed. Surveys show that countries outside the United States have rarely suffered genuine banking panics (Bordo 1986; Schwartz 1986; 1988a; b). In many instances these countries avoided panics even while they lacked central banks or other public lenders of last resort (Selgin 1994).

10. See Dowd (1992b) for case studies on nine historical episodes of free banking. Dowd assesses the historical record of these systems: “most if not all can be considered as reasonably successful, sometimes quite remarkably so” (1992, 2). A possible notable exception would be the so-called Free Banking Era in United States history from 1837–1862. However, whether this episode constitutes a genuine example of free banking is disputable. In the words of Freixas and Rochet (1997, 261): “Although the period from 1837 to 1864 in the U.S. is often referred to as the Free Banking Era, the term is something of a misnomer, for it refers not to a general system of ‘free’ banking in the literal sense described previously, but rather to various state banking systems based on so-called ‘free banking’ laws, which, though they made it unnecessary for new entrants to secure charters (each of which was subject to a vote by the state legislature), nonetheless restricted their undertakings in important ways. Most importantly, U.S. ‘free’ banks were denied the right to establish branch networks, and had to ‘secure’ their notes by purchasing and surrendering to state banking authorities certain securities those authorities deemed eligible for the purpose. The securities in many cases included bonds of the authorizing state governments themselves; and it has been determined that the depreciation of these very securities was the chief cause of ‘free bank’ failures, and indeed of bank failures generally, during the period in question. The lack of branch banking, in turn, caused state-issued banknotes to be discounted at varying rates once they had traveled any considerable distance from their sources. In short, the shortcomings of banks and bank-supplied paper currency during the so-called ‘free banking era’ in the U.S., far from establishing the need for special regulation of banks, testifies to the dangers of unwarranted or unwise regulation.”

11. See White (1995) for details about free banking in Scotland. On the stability of the Canadian system, see Bordo, Rockoff, and Redish (1994). Briones and Rockoff (2005) discuss several historical cases of lightly regulated banking systems that worked well, including Canada.
more dimensions, and that certain bad regulations make them more so. Moreover, many scholars subscribe to the more nuanced view that contagion effects do exist and have been found to weaken banks even if not all bank runs are evidence of irrational behavior or lead to the bankruptcy of solvent banks.

Textbook coverage

The academic literature contains a healthy debate about the inherent stability of banks. Solid theoretical arguments are made both for and against, and detailed historical analysis and institutional investigation finds evidence of instances of successful self-regulating systems, contagion effects, and regulatory weakening. In contrast, the textbook coverage of this topic only presents arguments and evidence clustered toward the inherent-fragility end of the spectrum. Therefore the textbooks fail to provide students with all of the historical information crucial for understanding episodes of banking instability. Since the textbooks only present one side of the story, students are unable to sharpen their analytical understanding of the inherent-fragility view or critically assess its tenets.

Hubbard and O’Brien (2018, 389–393) present an inherent-fragility view under the heading The Origins of Financial Crises. The first subheading, titled The Underlying Fragility of Commercial Banking, claims that banks are inherently fragile due to the liquidity risk caused by maturity mismatch. The second subheading, titled Bank Runs, Contagion, and Bank Panics presents the “me-first” problem facing depositors and states:

In other words, in the absence of deposit insurance, the stability of a bank depends on the confidence of its depositors. In such a situation, if bad news—or even false rumors—shakes that confidence, a bank will experience a run. (Hubbard and O’Brien 2018, 390, emphasis in original)

Hubbard and O’Brien also claim contagions are inevitable in an unregulated banking system due to asymmetric information:

The underlying problem in contagion and bank panics is that banks build their loan portfolios on the basis of private information about borrowers, information banks gather to determine which loans to make. Because this information is private, depositors can’t review it to determine which banks are strong and which are weak. (ibid.)

They conclude the section with the statement: “So, bad news about one bank can raise fears about the financial health of others, resulting in a bank panic” (ibid.).

In a section titled Government Safety Net Mishkin (2019, 217–222) likewise
presents an inherent fragility argument, citing the problems of asymmetric information and the fact that banks operate according to a “sequential service constraint” (i.e., first-come, first-served for depositor withdrawals), which he argues makes the banking system inherently susceptible to bank panics in the absence of a government safety net. Mishkin gives an example of an adverse shock to the economy that causes 5 percent of banks to become insolvent:

> Because of asymmetric information, depositors are unable to tell whether their bank is a good bank or one of the 5% that are insolvent. Depositors at bad and good banks recognize that they may not get back 100 cents on the dollar for their deposits and will want to withdraw them. (Mishkin 2019, 218, emphasis in original)

Ball (2012, 286–292) similarly presents an inherent-fragility viewpoint in a section titled *Bank Runs*. In answer to the question of what causes runs, Ball says that some runs are caused by news of pre-run insolvency, but that other runs can happen on pre-run solvent banks: “This happens if depositors lose confidence in the bank, which can happen suddenly and without good reason” (ibid., 286). Ball concludes the subsection titled *How Bank Runs Happen* by invoking the theory of self-fulfilling expectations: “if people expect a run, then a run occurs. This can happen even if nothing is wrong at the bank before the run” (287).

Brandl (2017, 152–153) presents a simple inherent-fragility view of bank runs that does not include a discussion of contagion effects in a subsection titled *Banks Are Subject to Bank Runs*:

> One of the biggest problems with bank runs is that they can be self-fulfilling prophesies. If people believe that their money is safe in the banking system, and they leave their money in the banks, then their money is safe—the system works as it is designed. If, however, people begin to question the safety and soundness of the banking system, and they respond to this uncertainty by pulling their money out of the banking system en masse, they can trigger a bank run. They can cause the banking system to become unsafe! (Brandl 2017, 152)

Cecchetti and Schoenholtz present an inherent-fragility view in a section titled *The Sources and Consequences of Runs, Panics, and Crises*:

> Banks not only guarantee their depositors immediate cash on demand; they promise to satisfy depositors’ withdrawal requests on a first-come, first-served basis. This commitment has some important implications. Suppose depositors begin to lose confidence in a bank’s ability to meet their withdrawal requests. They have heard a rumor that one of the bank’s largest loans has defaulted, so that the bank’s assets may no longer cover its liabilities. True or not, reports
that a bank has become insolvent can spread fear that it will run out of cash and close its doors. Mindful of the bank’s first-come, first-served policy, frenzied depositors may rush to the bank to convert their balances to cash before other customers arrive. (Cecchetti and Schoenholtz 2021, 360)

They then go on to describe a hypothesis about contagion effects:

If people believe that a bank is in trouble, that belief alone can make it so. When a bank fails, depositors may lose some or all of their deposits, and information about borrowers’ creditworthiness may disappear…. But that is not their main worry. The primary concern is that a single bank’s failure might cause a small-scale bank run that could turn into a system-wide bank panic….Information asymmetries are the reason that a run on a single bank can turn into a bank panic that threatens the entire financial system. (ibid., 361)

Croushore (2015, 179–180) presents a simple inherent-fragility view in the form of a narrative about how fractional reserve banking and the sequential service constraint can cause a bank run, but without using technical terms. Croushore then says:

The worst feature of bank runs is that they tend to spread from one bank to another, a condition known as a contagion. If depositors engage in a run on one bank, depositors at other banks may worry that their bank will be next. If their worry translates into action and each depositor tries to get his funds out before everyone else, the result will be another run on another bank. Before long, every bank may suffer a run, and many will close their doors. (Croushore 2015, 179)

All six textbooks present an inherent-fragility view of banks and tell nearly identical stories about the causes of bank runs. Regardless of which textbook is assigned, students who merely read the textbook will come away with the impression that economists are agreed that banks are inherently unstable, and that bank runs and panics are historically common occurrences in economies where banks are not regulated by government. But economists are not agreed on either of these points. The problem is worsened by the fact that all six textbooks point to the instability of banks in the U.S. in the late 19th and early 20th century as evidence of the inherent instability of banks in the absence of government regulation, but do not mention the legal restrictions that contributed to that instability, nor the examples of stable free-banking systems without any financial panics. We explore this issue separately below.
The origins of central banks created before the Fed

Just as there is a range of views on the inherent stability of banks, there are various explanations for why the first central banks came into existence. One view, which we call the ‘market failure hypothesis,’ is that central banks are created to stabilize the banking system and protect the macroeconomy from disruptions inherent in free monetary and financial markets. According to the market failure view, the proliferation of central banks across the globe is straightforward evidence of the institution’s efficiency. Charles Kindleberger (1994, xi), for instance, says that anyone who questions the market-failure view “has to explain why there seems to be a strong revealed preference in history for a sole issuer.” Charles Goodhart (1988, 1) has notably argued that “the role and functions of central banks have evolved naturally over time” for the theoretical reason that a government central bank is needed to provide efficient supervision of banks and to provide an effective lender of last resort to the financial system during a liquidity crisis.

A second view, which we call the ‘government-interest hypothesis,’ is that central banks first came into existence not to correct for market failure but for fiscal motivations, either to obtain revenue through seigniorage or to fund deficit spending through subsidized borrowing (Selgin and White 1999). Adherents of this view argue the spread of central banks demonstrates their political expedience and not their efficiency-enhancing properties.

According to the government-interest hypothesis, central banks did not develop naturally as a product of market forces. Rather, the first central banks developed unintentionally out of a process by which the government would grant a commercial bank unique legal privileges that other banks did not enjoy. As a result, the privileged bank would grow larger and more central to the financial system, and eventually take on the roles of bankers’ bank and lender of last resort.

---

12. For a thorough discussion of the market failure and government-interest hypotheses see White (1999, ch. 4), who we rely upon here.
13. See also Goodhart (1987; 1994). In a similar vein, Giannini (2011, xxvi–xxvii) argues central banks are the evolutionary institutional solution to the public good of monetary stability: “Any attempt to move beyond commodity money, even in its most advanced from of coinage, must entail an intermingling of money and credit circuit…. The intermingling of money and credit circuit thus set in motion a long and somewhat tortuous process of institutional adaption centered around the figure of the central bank.”
14. See White (1999, ch. 4) for a critique of the view that central banks develop naturally to deal with market failures. In White’s view private clearinghouses could, and historically did, provide the economic roles of commercial bank regulation and lender of last resort. As for central banks, he concludes “The
The historical details regarding the origins of the earliest central banks largely conform to the government-interest hypothesis. The Bank of England, for instance, was granted monopoly privileges for explicitly fiscal reasons. In Walter Bagehot’s (1877, 92) words, the Bank of England “was founded by a Whig Government because it was in desperate want of money.” In 1694 the bank of England was granted an exclusive charter in exchange for loaning £1,200,000 to the Treasury in order to fund the War of the Grand Alliance. Shortly thereafter Parliament granted additional privileges enjoyed by no other bank, including limited liability for shareholders and denying the right to issue notes to any other bank with more than six partners. Over time a series of further legislation, culminating in Sir Robert Peel’s Bank Charter Act of 1844, enabled the Bank of England to secure a monopoly of note issue in England and Wales and to take on the role of bankers’ bank. With the passage of Peel’s Act, the Bank of England arguably acquired all the characteristic functions of a central bank.

The Bank of France was likewise established for clear-cut fiscal reasons. After the turmoil of the French Revolution, there was a brief period of relatively free banking between 1796 and 1803 that operated well (Nataf 1992). Then the Bank of France was created in 1803 to finance Napoleon Bonaparte’s military campaigns (Smith 1990/1936, ch. 4). Both Napoleon and the French government were significant shareholders. Napoleon’s government granted the bank a monopoly of banknote issue, and in return the bank gave prodigious loans to the government.

Adherents of the government-interest hypothesis argue fiscal considerations likewise predominated in the majority of instances of central banks created before the outbreak of World War I, at which date the gold standard operated in most countries without a central bank.

World War I brought about the demise of the classical gold standard as combatant nations suspended specie payments so that central banks could create development is then ‘natural’ in the same sense that comedian Steven Wright suggests that it counts as ‘dying a natural death’ when one is hit by a train: ‘You get hit by a train, naturally you die.’ The standard meaning of ‘natural’ in economics—as in the phrase ‘natural monopoly’—is however, ‘brought about by market forces rather than by government intervention’” (White 1999, 72 n.3). On the central banking role of private clearinghouse associations, see Timberlake 1984; Mullineaux 1987; Gorton and Mullineaux 1987.

16. For details, see Smith (1990/1936, ch. 2). In Smith’s words, “The early history of the Bank was a series of exchanges of favours between a needy government and an accommodating corporation” (ibid., 12). Fiscal motivations likewise pervaded the rechartering of the bank over the subsequent 150 years. The charter was renewed in exchange for a variety of services including low interest or interest free loans to the Crown, direct payments, and loan term conversions for British trade companies (Clapham 1945, ch. 2, ch. 5).
money to finance wartime expenditures. Widespread international adoption of central banking institutions occurred in many countries after the creation of the Federal Reserve System in the United States and the rise of the new international monetary order that formed after the end of World War I.

Besides market failure and government interest there are many other hypotheses, including the hypothesis that at least some central banks may have come into existence: 1) in an effort to correct instability in the banking system that was created by previous inefficient government policies; 2) because of improving technological or economic knowledge; 3) because of the desire of politicians in less advanced economies to copy the institutions in more developed economies; 4) network effects; or 5) because central banking is the result of a natural evolutionary process in financial markets.

Textbook coverage

With the exception of Cecchetti and Schoenholtz (2021), none of the textbooks discuss the factors that motivated the creation of central banks before the establishment of the Federal Reserve System in the United States in 1913, which we discuss in the next section below.

Cecchetti and Schoenholtz discuss the creation of the Bank of England and the Bank of France in a section titled *The Basics: How Central Banks Originated and Their Role Today*. They begin by stating: “The central bank started out as the government’s bank and over the years added various other functions” (2021, 394). In a subsection titled *The Government’s Bank*, Cecchetti and Schoenholtz go on to write: “Governments have financial needs of their own. Some rulers, like King William of Orange, created the central bank to finance wars. Others, like Napoleon Bonaparte, did it in an effort to stabilize their country’s economic and financial system” (ibid., 394). They then provide further details in a footnote:

The Bank of England was charted in 1694 for the express purpose of raising taxes and borrowing to finance a war between Austria, England, and the Netherlands on one side and Louis IV’s France on the other. The Banque de France was created in 1800 in the aftermath of the deep recession and hyperinflation of the French Revolutionary period. For a more detailed discussion, see Glyn Davies’ *A History of Money: From Ancient Times to the Present Day* (Cardiff: University of Wales Press, 2002). (Cecchetti and Schoenholtz 2021, 394 n.2)

We commend Cecchetti and Schoenholtz’s inclusion of some historical information about how the first central banks originated. However, we believe the statement that Napoleon created the Bank of France to stabilize the economic and
financial system is historically inaccurate. As explained above, the economic and financial system was already stabilized by 1796, and Napoleon created the Bank of France for fiscal reasons. Almost immediately afterward Napoleon pressured the bank into an over-issuance of paper currency that destabilized prices by causing inflation.

We agree with Cecchetti and Schoenholtz that it is important for students to understand that central banks are linked to the financial needs of government and that early central banks were created explicitly to finance war. Such information gives historical context for discussions of debt monetization and the 1951 Treasury–Federal Reserve Accord, the seigniorage temptation for governments to create inflation, and central bank independence.

Students unaware of the reasons surrounding the origins of central banks besides the Fed are left with the impression that all central banks came into existence to end bank panics. They also tend to assume all central banks are created out of the public interest and operate to serve the public interest. When presented with the facts about the origins of the earliest central banks, students learn about both market failures and government failures that have happened in different places in different times, and how similar issues might be at play in our own country or in other countries around the world today.

The fragility of U.S. banks during the National Banking Era and the origins of the Federal Reserve System

Bank panics were frequent during the National Banking Era, from the National Banking Act in 1863 until the creation of the Federal Reserve System in 1913. There is a spectrum of views about the causes of these panics. On one
end of the spectrum, the market-failure view suggests that the series of panics in 1873, 1884, 1890, 1893, and 1907\textsuperscript{20} are evidence of the instability that results from the absence of the government fail-safes of a central bank and deposit insurance. The recessions caused by these banking panics are cited as the primary factor that convinced the American people that a central bank was necessary. The reduction in the frequency of bank panics in the period immediately after the creation of the Federal Reserve System in 1913 is offered as evidence of the efficacy of central banking.\textsuperscript{21}

On the other end of the spectrum, adherents of the government-failure theory (Champ, Smith, and Williamson 1989; Calomiris and Haber 2014, 153–184) argue the weakness of the U.S. banking system prior to the creation of the Federal Reserve System was a result of legal restrictions. They point out that financial crises were mainly a U.S. phenomenon during the late 19th century.\textsuperscript{22} The reason, they argue, is that two government policies greatly weakened the U.S. banking system. The first was branching restrictions that limited the size of banks and their ability to diversify assets. During this era, most banks were unit banks with a single location. Under the dual banking system created by the National Currency Act, federally chartered banks were largely unable to branch.\textsuperscript{23} State-chartered banks were not permitted to branch across state lines and were subject to state banking laws that in most instances either prevented or restricted branching. Furthermore, state-chartered banks were subject to a punitive federal tax on note issuance, which inhibited their ability to issue currency. Because of these restrictions, most U.S.

\textsuperscript{20}Jalil (2015) provides a detailed examination of contemporary financial reporting and suggests that only the panics of 1873, 1893, and 1907 were widespread throughout the United States, whereas the events of 1884 and 1890, along with numerous additional financial panic episodes, were localized events that should not be counted as major, or economy-wide, banking panics. Wicker (2006) also indicates only the 1873, 1893, and 1907 episodes count as major, widespread banking panics.

\textsuperscript{21}On this point see Miron (1986), who provides evidence that the regime shift from the National Currency System to the Federal Reserve System dampened seasonal interest rate fluctuations and diminished bank panics.

\textsuperscript{22}According to Calomiris (2010, 5) there were only 10 banking crises worldwide between the years 1875 and 1913, and five occurred in the United States. This fact alone suggests the high frequency of bank crises in the U.S. before 1913 cannot be explained by the contracting structure of banks per se. The contracting structure of banks—borrowing short-term to provide opaque long-term loans subject to a first-come, first-served constraint—has been essentially the same since the earliest beginnings of commercial banking. Yet some countries have experienced frequent bank crises, whereas some countries have only had one bank crisis, and others have had none at all. Moreover, the countries with frequent panics were the ones with the most severe regulatory restrictions, and the countries without any crises were regulated the least.

\textsuperscript{23}The text of the National Banking Act did not expressly prohibit branching by national-chartered banks. Some scholars have argued that Hugh McCulloch, the first Comptroller of the Currency and thus chief regulator of national banks, established de facto prohibition on branching through an overly strict interpretation of the National Banking Act (see Selgin 2016, 4–5; McCulley 1992, 13–14).
banks were tiny, undiversified, and prone to failure.\(^\text{24}\)

The second government policy was bond collateral restrictions on note issue. Any federally chartered national bank during this era was permitted to issue paper banknote currency. To issue notes national banks were required to buy and deposit at the Treasury federal government bonds equal in value to ten-ninths of the value of the banknotes they issued.\(^\text{25}\) The reasons for imposing bond collateral requirements were to provide a uniform currency with sound backing and to help finance the government’s deficit spending during the Civil War.\(^\text{26}\) But the unintended consequence was that the supply of paper currency in the U.S. tended to vary with bond prices and became limited by the value of outstanding government bonds available for purchase.\(^\text{27}\) The bond collateral requirement created an inelastic national currency supply because the national banks that issued the currency could not easily adjust the volume in circulation to cope with seasonal changes in the demand (Selgin and White 1994a).\(^\text{28}\)

Adherents of the legal restrictions view argue that it was the inability of U.S. banks to satisfy the seasonal demand for currency that led to recurring panics (Laughlin 1898; Horwitz 1990; White 1987; Lowenstein 2015, 49–55; Selgin 2016). During the harvest season farmers had an increased need for currency to pay

---

24. For instance, in 1910 there were over 19,000 banks, but only 292 banks operated branches and the total number of branches was 548 (Federal Reserve Committee on Branch, Group, and Chain Banking 1931, 123).

25. State-chartered banks were subject to a prohibitive 10 percent tax on banknote issue, which was designed to encourage banks to seek a national charter. This tax effectively curtailed the issue of notes by state banks and contributed greatly to the prevalence of unit banking. For a review of the motives that led to the 10 percent tax and its effects, see Selgin 2000.

26. The mix of fiscal motivation and public interest justification for the National Currency Act affords both a market failure and a government failure interpretation. See Rockoff (1974; 1975a; b; 1985) for details about the institutional arrangements that existed during the 1837–1862 U.S. “Free Banking” Era and for an evaluation and explanation of the economic performance of American banks during that period. Rolnick and Weber (1983, 1090) find little evidence of contagion effects even during the U.S. Free Banking era, stating: “Our preliminary conclusion from this evidence is that it is misleading to characterize the overall free banking experience as a failure of laissez-faire banking.”

27. For instance, the upward limit on the quantity of paper currency that could be printed in the U.S. when the National Banking Act was passed in 1863 was approximately $300 million. The quantity of government-issued United States Notes, known as ‘greenbacks,’ which were the only other form of paper currency in circulation at that time, was fixed and would be gradually reduced in the postbellum period. Therefore, when the federal government began retiring the national debt after the end of the Civil War the limit on the size of national currency decreased as the face value of outstanding bonds decreased. This led to a further problem of a paper currency shortage because the currency supply was forced to shrink at the same time that the demand for currency was increasing due to a growing U.S. economy.

28. Two of the main difficulties that banks faced were buying suitable bonds at prices that made it profitable to issue banknotes and avoiding delays before new notes could be put into circulation. These problems were exacerbated during panics. James (1976) and White (1987) provide details concerning currency shortages during the National Banking Era.
farmhands. Since national banks could not easily or profitably increase the quantity of paper banknotes in circulation, farmers withdrew gold and silver coins from their deposit accounts to pay farmhands instead. The reserve drain on banks led to yearly credit contractions, seasonally high interest rates, and full-fledged financial panics in 1873, 1884, 1890, 1893, and 1907.29 During these panics there were widespread bank runs and suspensions. Advocates of the regulatory weakening hypothesis argue the susceptibility of U.S. banks to runs and panics during the National Banking Era were policy inflicted. They point out that during these same years there were zero bank failures and far less banking system distress in Canada, which also featured a relatively large agricultural sector and the same seasonal shifts in currency demand. Canadian banks were allowed to branch and were therefore much larger, more diversified, and fewer in number than American banks. Canadian banks also did not have bond collateral requirements on note issue, and would simply issue a larger amount of paper banknotes to meet seasonal needs, and withdraw the extra notes from circulation when they were no longer needed.30 Finally, adherents of the legal restrictions view point out that the defects of the dual banking system in the United States during National Banking Era were not lost on contemporary banking experts. During the latter part of the nineteenth century an Asset Currency Movement gained traction, and multiple bills were introduced into Congress that would have abolished the existing bond-secured currency and replaced it with currency that was backed by general bank assets, while at the same time allowing for nationwide branch banking (Lowenstein 2015, 55).31 Adherents

29. These panics were typically triggered by the failure of a large New York bank, a tightening of the New York money market, and a fall in stock prices that exacerbated the New York banks’ seasonal liquidity problems by making it impossible for them to call in loans and prompting suspensions. The problems of the New York banks were caused by the pyramiding of liabilities in New York banks during the National Banking Era.

30. It is worth noting that even during the unstable National Banking Era in the U.S. there is not much evidence of contagion effects. In the words of Selgin and White (1994b, 1726): “of the five or six major panics (1873, 1884, 1890, 1893, 1907; perhaps also 1896) during the National Banking era, only three involved suspensions of payments. Among them, failure statistics suggest that only the 1893 panic involved a nationwide contagion (George Kaufman 1988, 566; 1994). Runs against National Banks appear to have been triggered by news indicating probable bank insolvency, contrary to the theory that depositors stage runs simply out of fear that others might run (Gorton 1985; 1988). Furthermore, bank customers in the National Banking era panics attempted to redeem deposits for currency, but generally did not attempt to redeem banknotes for gold or other legal tender (R. Alton Gilbert 1988, 137–138 n.3); in Northern states holdings of Canadian banknotes also increased. These facts suggest that fear of currency shortage (banknote issue was legally restricted) rather than fear of bank failure was at work.”

31. For more details regarding the fate of asset currency reform movements, see Wicker 2005; McCulley 1993, 42–75. The most notable attempt at reform was the Indianapolis Monetary Commission’s 1898 proposal. In addition, multiple Asset Currency bills were introduced into the House of Representatives by Charles N. Fowler, the Chair of the House Committee on Banking and Currency. After making their way through the House these bills were ultimately rejected in the Senate.
of the regulatory weakening hypothesis argue these proposals were economically sound but blocked for political reasons by a rent-seeking coalition comprised of large New York banks, small unit banks, farmers, and the political machinations of Senator Nelson Aldrich. They claim the National Monetary Commission, which ultimately led to the Federal Reserve Act, was a façade behind which Aldrich set aside the sounder asset currency proposals in favor of an alternative proposal that benefited agrarian debtors and preserved the dominance of New York banks but did not address the restrictions on branching that was the root cause of bank instability (Selgin 2015).

Textbook coverage

We see three shortcomings in the textbooks. First, none clearly explain that the restrictions on U.S. banks during the National Banking Era were unusual restrictions. Second, every textbook presents the recurring banking panics in the U.S. as straightforward evidence that banks are inherently unstable and prone to frequent panics in the absence of a government safety net. Presenting the U.S. experience in this light is questionable considering U.S. banks were subject to unusual restrictions and that bank panics were very infrequent in most other advanced economies by the late 19th century and many of those countries did not have a central bank. It is an error of omission that none of the textbooks mention these facts. And third, the creation of the Federal Reserve System is presented as the only option for stopping bank panics in the U.S., when in fact there was another viable alternative, namely, the asset-currency movement that was blocked for political reasons.

Brandl (2017, 79–81) specifically addresses the causes of the Panic of 1907 and presents the most information on the topic:

In their 2007 book titled The Panic of 1907: Lessons Learned from the Market’s Perfect Storm, Robert Bruner and Sean Carr state the Panic of 1907, like all financial crises, was not caused by one single event. Rather, they state the Panic of 1907, like most financial crises, was caused by the culmination of a number of bad things happening at once. The Panic of 1907 was triggered by wild speculation in the stock market; excessively loose lending by banks and trusts . . . ; a need to divert cash to San Francisco for rebuilding after the 1906 earthquake; and a lack of effective oversight of financial markets. (Brandl 2017, 79)

Under the Lessons Learned box at the end of the chapter under the caption The Need for a Central Bank, Brandl summarizes the chapter as follows:
As the dust cleared from the Panic of 1907, the lessons to be learned from the experience became clear. The US banking system had become so large and so important to the rest of the economy that it needed a ‘lender of last resort’ during a time of crisis. In addition, to avoid financial crises, the United States needed a single currency used nationwide instead of thousands of different bank notes. Simply put, the United States needed a central bank. (Brandl 2017, 81)

Brandl does point out in a section titled The Bank of Canada (ibid., 179–181) that “The private, large, countrywide banks had branches in rural areas as well as urban areas, with few bank failures or bank runs.” However, he does not relate this to the U.S. experience.

In discussing the dual banking system, Mishkin (2019, 238) says: “To eliminate the abuses of the state-charted banks (called state banks), the National Bank Act of 1863 (and subsequent amendments to it) created a new banking system of federally chartered banks (called national banks).” He goes on to state: “This legislation was originally intended to dry up sources of funds to state banks by imposing a prohibitive tax on their banknotes while leaving the banknotes of the federally chartered banks untaxed.” Later on, in a section titled The Origins of the Federal Reserve System, Mishkin says:

The termination of the Second Bank’s [Second Bank of the United States] national charter in 1836 created a severe problem for American financial markets, because there was no lender of last resort that could provide reserves to the banking system to avert a bank panic. Hence, in the nineteenth and early twentieth centuries, nationwide bank panics became a regular event, occurring every twenty years or so, culminating in the Panic of 1907. The 1907 Panic resulted in such widespread bank failures and such substantial losses to depositors that the public was finally convinced that a central bank was needed to prevent future panics. (Mishkin 2019, 295)

Hubbard and O’Brien (2018, 389) state the U.S. banking system was unstable in the pre-Fed era due to the lack of a government safety net: “For most of the nineteenth and early twentieth centuries, … neither federal deposit insurance nor the Federal Reserve existed. As a result, banks were subject to periodic bank runs.” In a subsection titled Government Intervention to Stop Bank Panics they do not mention the pre-Fed panics specifically by date or provide historical details, but merely say:

Congress reacted to bank panics by establishing the Federal Reserve System in 1913. Policymakers and economists argued that the banking industry needed a “bankers’ bank,” or lender of last resort. (Hubbard and O’Brien 2018, 390, emphasis in original)
As noted above, Ball (2012) endorses an inherent-fragility view of banks. He presents the instability of U.S. banks during the National Banking Era as evidence:

Nationwide bank panics were once common in the United States. Between 1873 and 1933 the country experienced an average of three panics per decade. Bank panics occur because a loss of confidence is contagious. A run on one bank triggers runs at others, which trigger runs at others, and so on. (Ball 2012, 290)

Ball (2012, 16) makes note of unit banking and says that it results in less efficiency and lower economic growth, but he never explicitly draws the connection between unit banking and the unusual instability in the U.S. banking system. Ball also makes note of the political opposition to branch banking: “Many unit banks were happy with the status quo, which gave them local monopolies. The American Bankers Association lobbied against branching” (ibid., 227). Finally, Ball presents a fairly balanced discussion of Abraham Lincoln’s motivations behind the National Bank Act:

As president, Lincoln proposed the National Bank Act, which Congress passed in 1863….Lincoln was motivated partly by episodes of fraud at state banks. In addition, like Alexander Hamilton, he hoped to unify the nation’s currency….Finally, national banks helped finance Union spending on the Civil War, because they were required to purchase Treasury Bonds. (Ball 2012, 227)

Cecchetti and Schoenholtz (2021, 398–400) do not give the dates of pre-Fed panics or provide any details about them. Rather, in a section titled Stability: The Primary Objective of All Central Banks they merely say:

The rationale for the existence of a central bank is equally clear. While economic and financial systems may be fairly stable most of the time, when left on their own they are prone to episodes of extreme volatility. Prior to the advent of the Fed, the U.S. financial system was extremely unstable. It was plagued by frequent panics. (Cecchetti and Schoenholtz 2021, 398)

Croushore (2015, 180) discusses the restrictions on branch banking but does not relate them to bank runs and instability:

From 1864, when the national banking system was established until 1927, when the McFadden Act was passed, commercial banks with a national charter from The Comptroller of the Currency were forbidden to have any branches. The economic impact of the restrictions was to keep most banks inefficiently small. The restriction also prevented well-run banks from expanding to compete with poorly managed banks. (Croushore 2015, 180)
Later, Croushore says:

Bank runs, which occurred frequently in the late 1800s and early 1900s, have been eliminated almost completely. Financial crises that led to serve recessions were commonplace before World War I; such crises are now much less common and have far less impact on the economy. (Croushore 2015, 184)

The textbooks present the straight-line narrative that the recurring bank panics before 1913 are evidence of the inherent instability of commercial banks. The Federal Reserve Act is presented as the necessary and only logical solution for ending these panics. This narrative is incomplete and potentially misleading. Most of the textbooks mention that the U.S. had unit banking, but none of the textbooks explain why lack of branching factored into the instability of the U.S. banking system prior to the Fed. Nor do they mention the problems inherent in the bond collateral requirement or the asset-currency alternative to a central bank.

U.S. bank panics during the Great Depression

Scholars have identified numerous factors that may have contributed to the length and severity of the Great Depression. These include the monetary policy of the Federal Reserve in the 1920s, the stock market crash of 1929, an autonomous collapse of investment spending, tariffs and declining international trade, the shock-amplifying mechanism of the gold standard, various New Deal policies, and monetary contraction and banking failures. The largest concentration of bank suspensions in U.S. history occurred between 1930 and 1933. More than 9000 banks failed during those years, representing approximately 30 percent of banks that had been in business at the end of 1929. Economists disagree about whether and to what extent these bank failures played a causal role in worsening the depression. Economists also disagree about whether and to what extent contagions of panic played a role in causing banks to fail. We identify three main views that are prominent in the academic literature.33

One view that has been prominent since John Maynard Keynes’s (1936) publication of The General Theory of Employment, Interest and Money is that the bank failures were panics caused by a decline in income and interest rates. So according

---

32. See Mitchener and Richardson (2019), who find evidence of network contagion.
33. Our classification of views as ‘Keynesian,’ ‘Monetarist,’ or ‘fundamentals’ is useful for discussion purposes but an oversimplification and permits of combinations. For instance, Temin’s (1976; 1989) influential work combines some aspects of both Keynesian and Monetarist views while rejecting other aspects of both views. There are also many heterodox views as well.
to what we'll call the ‘Keynesian panic view,’ the banking panics did not cause a decline in income, interest rates, and the money supply, but instead were caused by them. The decline in income, interest rates, and the money supply were caused by an autonomous drop in investment spending and a collapse in confidence after the stock market crash in 1929. On this view, the Depression is evidence of macroeconomic failure inherent in the free market and the banking crisis did not play an exogenous causal role.

A second view, which we call the ‘Monetarist panic view,’ has been prominent since Milton Friedman and Anna Schwartz (1963) published *A Monetary History of the United States: 1867–1960*. On this view, panic-induced deposit withdrawals caused the currency-deposit ratio to rise and provoked waves of bank suspensions. The money supply therefore decreased independently and caused an exogenous decline in aggregate demand that caused unemployment to rise and output to fall. The bank suspensions are offered as evidence of contagion effects that played a causal role in propagating the depression. According to this view the Federal Reserve had sufficient power to cut short the process of monetary contraction and banking collapse. The Depression is presented as an example of government failure by the central bank in which the banking crisis played a key role.

A third view, which we will call the ‘fundamentals’ view, holds that fundamental shocks to bank solvency, such as increased default risks to banks’ loan portfolios or a fall in the value of bonds held by banks, caused banks to become insolvent (Calomiris and Mason 1997; Boughton and Wicker 1979; 1984; Wicker 1980; 1996; Ramirez 2003). According to this view, bank failures before 1933 were mainly the result of local shocks that proved fatal to many banks weakened by legal restrictions such as those on branching. These were not genuine panics in the sense that illiquidity caused by unwarranted deposit withdrawals caused many solvent banks to become insolvent. Rather, the vast majority of the banks that failed during the Great Depression were tiny unit banks that could not branch or diversify their assets, leaving them prone to failure. According to the fundamentals

---

34. While the fundamentals view agrees with the Monetarist panic view that bank failures contributed to the length and severity of the Great Depression, it sharply contrasts with the Monetarist panic view that these bank failures were caused by waves of panics that caused solvent banks to become insolvent. In the Monetarist view panicking by the public was an exogenous source of instability unrelated to banks’ asset conditions and could have been prevented by correct action by the Federal Reserve. The fundamentals view holds that bank failures were endogenous. On this point see Calomiris (2007, 6), who says: “Endogenous contractions of deposits and loans, just like unwarranted contractions, will limit the supply of money and credit, and thus they will exacerbate the macroeconomic decline that caused them. Thus, according to the fundamentals view, banking distress can magnify economic downturns even if banks are not the originators of shocks; banks will tend to magnify macroeconomic shocks through their prudential decisions to curtail the supplies of loans and deposits in response to adverse shocks, even if banks are passive responders to shocks and even if depositors avoid engaging in unwarranted runs or panics.”
view, neither the proliferation of unit banks nor the large number of bank failures were natural features of a free-banking system, but were the result of nation-wide restrictions on branching. Adherents of the fundamentals view point out that California was one of the few states that allowed branching, and California banks were more efficient and had lower rates of bank failures in the 1930s than other states (Carlson and Mitchener 2009). Likewise, in Canada, where most banks had nation-wide branching, there was not a single bank failure. 35

Contagion effects appear to have played a limited role in bank failures prior to 1933. 36 It has even been argued that the one genuine banking panic in the U.S., in February/March 1933, was not caused by random or irrelevant events, nor by a general fear amongst the public, nor by a mistrust of banks. Rather, it was not a run on banks at all. Rather, it was a run on gold caused by the perception that Franklin Roosevelt would devalue gold after his impending inauguration. 37 This perception soon proved to be correct, validating the anticipation.

The fundamentals view sees the bank failures during the Great Depression as primarily the result of the government failure of imposing harmful legal restrictions that left banks susceptible to insolvency after real economic shocks—and not a result of unwarranted panics. It argues that the collapse of deposits and loans reduced the money supply and caused a contraction of credit, which exacerbated the economic decline of the Great Depression.

Textbook coverage

Hubbard and O’Brien (2018, 397–403) provide an entire section on *The Financial Crisis of the Great Depression*. After addressing several initial factors, a subsection entitled *The Bank Panics of the Early 1930s* presents an outline of the

---

35. Grossman (1994) offers details about the stability of the Canadian banking system and ten other countries during the Great Depression. Grossman finds that “macroeconomic policy—especially exchange-rate policy—and banking structure, but not lenders of last resort, were systematically responsible for banking stability” (1994, 1).

36. See Selgin and White (1994b, 1726–1727): “Contagion effects also appear to have played a more limited role than is usually supposed during the ‘Great Contraction’ of 1930 to 1933. Prior to 1932, bank runs were confined mainly to banks that were either pre-run insolvent themselves or affiliates of other insolvent firms (Elmus Wicker 1980). Serious regional contagions erupted in late 1932, but these were aggravated if not triggered by state governments’ policy of declaring bank ‘holidays’ in response to mounting bank failures (George J. Benston et al. 1986, 52).”

37. Wigmore (1987) provides evidence that the banking crisis of 1933 was caused primarily by a gold drain due to a speculative attack on the dollar. The anticipation that Franklin Roosevelt would devalue the dollar provoked a sharp increase in foreign and domestic demand for gold that exhausted the gold reserves of the Federal Reserve Bank of New York. The drain on gold reserves of the banking system that led to the Bank Holiday of 1933 was a currency crisis precipitated by government currency manipulations and was not caused by domestic hoarding or a contagion of fear.
Monetarist panic view: “Many economists believe that the series of bank panics that began in the fall of 1930 greatly contributed to the length and severity of the Depression. The bank panics came in several waves” (ibid., 399). Without identifying the root cause—unit banking—they go on to state: “The large number of small, poorly diversified banks—particularly those that held agricultural loans as commodity prices fell—helped fuel the panics” (399). In a section on The Failure of Federal Reserve Policy During the Great Depression they give four reasons why the Fed worsened the depression: “No one was in charge”; “The Fed was reluctant to rescue insolvent banks”; “The Fed failed to understand the difference between nominal and real interest rates”; and “The Fed wanted to ‘purge’ speculative excess” (400–401).

Brandl, too, provides a Monetarist panic view, and says: “Many also blame an increase in the amount of cash held by the nonbank public as contributing to the Great Depression of the 1930s” (2017, 156). Brandl does not provide detailed information about the bank panics. In a subsection titled Financial Markets during the Great Depression he provides details about the stock market crash and castigates the Fed: “Where was the Federal Reserve in all this? Wasn’t it established in 1913 in response to the Panic of 1907 to avoid just such a financial and economic meltdown?” (ibid., 84). He pins this failure to act on “a weak leader” (82) and “the Burgess-Rifler [i.e., Real Bills] doctrine” (84). He says: “What was really needed was expansionary monetary policy to get the economy going again” (84).

Ball (2012, 290–292) in a section titled Bank Panics in the 1930s also presents a Monetarist panic view:

Major trouble began in 1930. Failures rose at rural banks in the Midwest, and this made depositors nervous about other banks in the region. ... A psychological milestone was the failure of the New York-based Bank of the United States in December 1930. ... Other events eroded confidence further. ... In the 1932 election campaign, Democrats publicized banking problems to criticize the Republican government. The stream of worrisome news produced a nationwide panic. (Ball 2012, 291)

Ball says that the bank panics ended after Roosevelt’s bank holiday because of a restoration of confidence [for unspecified reasons]: “President Roosevelt understood the psychology of panics. His famous statement that ‘we have nothing to fear but fear itself’ referred partly to banking. It captures the fact that panics result from self-fulfilling expectations” (ibid., 292), and directs the reader in a footnote to Friedman and Schwartz (1963) for more details.

Cecchetti and Schoenholtz alone among the textbooks present a mainly endogenous view of bank failures during the Great Depression: “The next series of bank panics occurred during the Great Depression of the 1930s, when output fell
by roughly one-third. Bank panics usually start with real economic events or their prospect, not just rumors” (2021, 363), and direct the reader to Gary Gorton (2012) in a footnote. Cecchetti and Schoenholtz later say:

The series of three bank panics that occurred during the Great Depression of the 1930s is one example of the failure of a lender of last resort. While the Federal Reserve had the capacity to operate as a lender of last resort in the 1930s, it chose not to do so. In effect, policymakers acted as if the “fire” would burn itself out. Instead, the conflagration spread and intensified. The result was the worst financial disaster in the 100-plus-year history of the Federal Reserve. (Cecchetti and Schoenholtz 2021, 365)

Mishkin (2019) presents a Monetarist panic view with a touch of the fundamentals view. In an application he titles The Mother of All Financial Crises: The Great Depression, Mishkin says:

What might have been a normal recession turned into something far worse, however, when severe droughts in the Midwest led to a sharp decline in agricultural production, with the result that farmers could not pay back their bank loans. The resulting defaults on farm mortgages led to large loan losses on bank balance sheets in agricultural regions. The general weakness of the economy, and of the banks in agricultural regions in particular, prompted substantial withdrawals from banks, building to a full-fledged panic in November and December of 1930, with the stock market falling sharply. For more than two years, the Fed sat idly by through one bank panic after another, the most severe spate of panics in U.S. history…. With a greatly reduced number of financial intermediaries still in business, adverse selection and moral hazard problems intensified even further. Financial markets struggled to channel funds to borrower-spenders with productive investment opportunities. The amount of outstanding commercial loans fell by half from 1929 to 1933, and investment spending collapsed, declining by 90% from its 1929 level…. The ongoing deflation that accompanied declining economic activity eventually led to a 25% decline in the price level. This deflation short-circuited the normal recovery process that occurs in most recessions. The huge decline in prices triggered a debt deflation in which real net worth fell because of the increased burden of indebtedness borne by firms and households. The decline in net worth and the resulting increase in adverse selection and moral hazard problems in the credit markets led to a prolonged economic contraction in which unemployment rose to 25% of the labor force. The financial crisis of the Great Depression was the worst ever experienced in the United States, which explains why the economic contraction was also the most severe ever experienced by the nation. (Mishkin 2019, 273–275)
Croushore (2015) does not provide any detailed information on the bank failures during the Great Depression.

The five textbooks that address the topic present a panic view of 1930s bank failures. All of those except Cecchetti and Schoenholtz present the bank failures as an exogenous causal factor that contributed to the severity of the Great Depression, and blame the Federal Reserve for significantly worsening the Great Depression (albeit in the wake of an exogenous market panic)—and notably so, as this is the only topic surveyed where the textbooks generally present a government-failure view. And as we show next, all the books say that the banking disturbances finally ended in 1933 partly due to the introduction of deposit insurance and partly due to the improved performance of the Fed, which learned from its mistakes of the 1930s.

**Deposit insurance**

In the economics literature deposit insurance has many defenders and many critics. The theoretical case for deposit insurance largely rests on the argument of Diamond and Dybvig (1983). In the Diamond and Dybvig model it is possible for the government to intervene with a policy they call deposit insurance that eliminates the bad bank-run equilibrium and allows the good, run-free equilibrium to dominate. By eliminating bank runs deposit insurance winds up costing the government nothing because the insurance never has to pay out. The more general argument for deposit insurance is that it removes the public’s incentive to run on banks by convincing them that deposits are insured by the government. Advocates of deposit insurance therefore conclude that deposit insurance is desirable because at minimal cost it decreases liquidity risk in the banking system.

Critics of deposit insurance argue that deposit insurance increases insolvency risk within the banking system and leads to bad economic outcomes—the moral hazard problem. Deposit insurance protects depositors from the downside risk of the bank’s investments not performing. Therefore, depositors have no incentive to shop around for a bank that meets their risk preferences. Instead, depositors have a perverse incentive to bank at the riskiest banks because these banks can share higher interest returns from riskier investments with depositors in the form of paying greater interest on deposits or other benefits desired by depositors.

Critics of deposit insurance also argue that deposit insurance eliminates the incentive for depositors to monitor their bank for changes to the riskiness of its investments (White 1999, Ch. 6). Without deposit insurance, depositors have an incentive to withdraw funds from banks that take on too much risk, and to run on insolvent banks. Insolvent banks close promptly, limiting the harm the managers might do by gambling for recovery and looting the bank’s assets. But with deposit
insurance, the public relies on regulators to close down insolvent banks. Regulators are often worse monitors than depositors with skin in the game. During the Savings and Loan Crisis of the 1980s, thousands of insolvent thrifts were left open for years. The zombie thrifts cost the economy and the taxpayer billions of dollars, all under the watchful eye of the regulators. Historical experience shows that regulators can often be asleep at the switch and practice too much forbearance. Deposit insurance does not, overall, protect the public from the risk generated by the avarice of bankers, as many of its supporters claim and most of the public believes. Rather, it subverts sound banking and results in costs for taxpayers.

Most defenders of deposit insurance point to the experience of the Great Depression to support the claim that banks are inherently unstable and prone to runs and panics. The Federal Deposit Insurance Corporation (FDIC) was established to oversee deposit insurance as part of the 1933 Banking Act, which was passed in the wake of more than 9000 bank failures between 1929 and 1933. The critics of deposit insurance argue it is a widespread misconception that deposit insurance was implemented to protect the public. They claim deposit insurance was not created to protect small depositors or to make the banking system safer. Rather, it was created to maintain the economic interests of unit banks by preserving the unstable unit banking system that was then in place (Calomiris and Jaremski 2016).

Some critics of deposit insurance argue that the bank failures between 1929 and 1932 were not caused by liquidity panics but real economic shocks that left many legally restricted unit banks insolvent. Deposit insurance, which is meant to prevent liquidity panics, cannot address solvency shocks and would not have been a viable remedy for preventing those bank failures. The only effective solution would have been to repeal the legal restrictions that made banks vulnerable. During the debates leading up to the creation of the FDIC, many thoughtful reformers wanted to create a more stable banking system by allowing branch banking, instead of instituting deposit insurance. Scholars, the public, and many politicians at that time were still scarred by the experience of prior unsuccessful government deposit insurance.

38. Dotsey and Kurianov (1990) give historical details on how the Savings and Loan Crisis was perpetuated and worsened by regulatory forbearance. Dotsey and Kurianov provide evidence that regulators lacked the financial resources to pay to close insolvent thrifts and delayed closure, hoping that insolvent thrifts might return to profitability if action was forestalled (1990, 12–14). Additionally, Dotsey and Kurianov argue that regulators had perverse political incentives to preserve mismanaged and insolvent thrifts (ibid., 20–23).

39. A companion deposit insurance entity for savings and loan institutions—the now defunct Federal Savings and Loan Insurance Corporation (FSLIC)—was created by the National Housing Act of 1934.

40. In the words of Calomiris and Haber (2014, 190): “Although the civics textbooks used by just about every American high school portray deposit insurance as a necessary step to save the banking system, all the evidence indicates otherwise: it was a product of lobbying by unit bankers who wanted to stifle the growth of branch banking.”
insurance programs in various U.S. states. But deposit insurance had a strong advocate in Representative Henry Steagall, the chairman of the House Committee on Banking and Currency. Steagall was motivated in part to protect the private interests of farmers and unit banks within his Alabama district and blocked all branch banking reforms and pushed through deposit insurance with the help of a powerful lobby that consisted mainly of rural farmers and unit banks.

Critics of deposit insurance also argue that contrary to widespread misperceptions, deposit insurance did not put an end to the banking panic in 1933—which was the last, and in their view only genuine Depression-era bank panic. Indeed, it could not have done so, because it was implemented after the panic was already over. The panic came to an end after the end of the bank holiday on March 13, 1933. But the Banking Act of 1933 was not passed until June 13, 1933, and temporary deposit insurance was not put into effect until January 1934.

Critics of deposit insurance also claim that deposit insurance has high costs without providing any compensating benefits. From 1933 until the Savings and Loan Crisis of the 1980s deposit insurance cost the public hundreds of billions of dollars in welfare losses in the form of monopoly quasi-rents garnered by banks by paying depositors lower interest and charging customers higher interest on loans than would have been possible in the market without the existence of deposit insurance. During this period, the prohibition against paying interest on deposits—known as Regulation Q—largely prevented banks from sharing the proceeds from riskier investments with depositors. Regulation Q, combined with limits on the size of deposits covered by deposit insurance, kept a check on the moral hazard.

41. President Franklin Roosevelt, the Secretary of the Treasury, and the Comptroller of the Currency all opposed deposit insurance, because they were familiar with the calamitous failures of state-level taxpayer guarantee plans. Roosevelt threatened to veto any bill that authorized government deposit insurance. Roosevelt backed off later to gain populist support for his other programs, but only after complaining “this bill has more lives than a cat” (“Roosevelt Hails Goal, He Calls Recovery Act Most Sweeping Law in Nation’s History, Johnson Administrator, Million Jobs by October 1, Employer’s Urge to Hire More Men with Government Stopping Unfair Competition,” New York Times, June 17, 1933, p. 1). See Flood (1992) on how the debate over deposit insurance in the early 1930s was informed by the moral-hazard problem and failures created by state taxpayer guarantee plans.


43. Calomiris and Haber (2014, 190) remark upon how “the banking crisis of 1932–33 ended months before the establishment of FDIC insurance.”

44. The 1933 Banking Act authorized the Temporary Deposit Insurance Fund, which began coverage on January 1, 1934, and a permanent plan that was to take effect on July 1, 1934. The permanent plan was delayed and full deposit insurance was actually put into effect July 1, 1935 (FDIC 1998, 30). Moreover, only deposits up to $2,500 were covered, which left approximately two-thirds of all bank deposits uninsured (Board of Governors of the Federal Reserve 1933, 28). It is difficult to see how deposit insurance could account for the banking calm as it was implemented after the crisis was ended and left most deposits uninsured.
problem for a time by preventing overly risky banks from offering high interest on deposits and thereby attract depositors at the expense of prudent banks and by leaving large depositors and most deposits uninsured.\(^{45}\)

High inflation in the 1960s and 1970s and mounting political pressure impelled regulators to begin to loosen the regulations that prevented paying interest on deposits. Meanwhile, the limits on deposit insurance were continually raised faster than the general rate of inflation. The combination of these factors unshackled the moral hazard problem and set the stage for the Savings and Loan Crisis of the 1980s.\(^{46}\) Between 1986 and 1995, 1,043 out of the 3,234 U.S. savings and loan associations failed. The U.S. General Accounting Office (1996, 13) estimated the cost of the crisis at $160.1 billion, approximately 4 percent of average annual GDP in the 1980s.

Canada was the next country to adopt deposit insurance, in 1967. Subsequently, many other countries followed suit, and today 146 countries have some form of government deposit insurance.\(^{47}\) In none of these cases did a country adopt deposit insurance because of any preceding banking crisis. A growing empirical literature finds that deposit insurance leads to more frequent bank failures and also bigger losses (Demirgüç-Kunt and Detragiache 1998; 2002; Demirgüç-Kunt and Kane 2002; Barth, Caprio, and Levine 2004; Demirgüç-Kunt and Huizinga 2004).\(^{48}\) The findings are supported by a large number of case studies on individual countries (Carr, Mathewson, and Quigley 1995; Mondschean and Opiela 1999; Beek 2002; Chernykh and Cole 2011). Empirical evidence also suggests that deposit insurance reduces overall economic growth by inhibiting financial development (Cecchetti and Krause 2005; Cull, Senbet, and Sorge 2005).\(^ {49}\)

Critics of deposit insurance conclude that the costs of deposit insurance likely exceed the benefits. The primary purported benefit of deposit insurance, that it reduces liquidity risk within the banking system, can be achieved by other means. First and foremost, an effective lender of last resort is sufficient to prevent a liquidity crisis. A lender of last resort and deposit insurance are therefore substi-

---

45. How important statutory limits on deposit insurance were in containing moral hazard during this period is debatable given that regulators in practice acted to keep all depositors whole and it was generally known that deposits beyond the size officially covered by deposit insurance were implicitly guaranteed.

46. Many studies find the high levels of risk taken by the S&Ls were primarily the result of moral hazard created by deposit insurance, e.g., Kane 1989; Barth 1991; Cebula 1993.

47. According to the International Association of Deposit Insurance [(link)](link).

48. According to Calomiris, banking crises worldwide were ten times more frequent and five times more severe in the period from 1980–2013 than they were in the period from 1874–1913. He argues that the widespread adoption of deposit insurance is a major factor contributing to the instability of the more recent period.

49. For a review of the recent literature on deposit insurance see Hogan and Johnson (2016).
tutes in the sense that they are two different institutions that exist to provide liquidity to the banking system. Moreover, private deposit insurance without mandated coverage is an alternative to the current system of government mandated deposit insurance. A number of private insurance systems existed in the United States before the FDIC. Historical evidence suggests that private insurance can provide all of the same benefits to depositors and banks as government mandated insurance, but at a lower cost. More importantly, private insurers would have a profit incentive to structure the terms and conditions and charge an actuarially sound premium for coverage, mitigating moral hazard.

Textbook coverage

Whether deposit insurance creates net benefits is a hotly debated topic of academic research. And there is contention over whether deposit insurance helped end the banking panic of 1933. The burgeoning empirical literature generally finds that overall the spread of deposit insurance has led to more frequent bank failures and larger losses. All of this contrasts with much of what is presented in the textbooks.

Hubbard and O’Brien say:

As we will see in this chapter, the Fed failed to stop the bank panics of the early 1930’s, which led Congress to create the Federal Deposit Insurance Corporation (FDIC) in 1934. By reassuring depositors that they would receive their money back even if their bank failed, deposit insurance effectively ended the era of commercial bank panics in the United States. (Hubbard and O’Brien 2018, 391)

50. Hogan and Johnson (2016) propose an array of alternatives to the FDIC, including privatization. They write: “Historical evidence of deposit insurance prior to the FDIC indicates that private mechanisms such as clearinghouses, coinsurance programs, and systems of self-regulation are likely to emerge to stem bank risk. The empirical evidence indicates that these proposals are likely to increase efficiency and stability in the U.S. banking system” (2016, 441–442).

51. See Calomiris (1990), who surveys both private and government insurance systems in the United States prior to the FDIC and finds: “In both the antebellum period and in the 1920s, insurance systems that relied on self-regulation, made credible by mutual liability, were successful, while compulsory state systems were not” (p. 283).

52. Part of the problem with government deposit insurance, and the main feature that contributes to the moral hazard problem, is that government deposit insurance providers do not charge premiums that are correctly adjusted for risk. For many years the FDIC charged a flat rate premium of one twelfth of one percent to all banks regardless of the riskiness of their investments, which was assessed against total deposits and not insured deposits. In the wake of the Savings and Loan crisis the FDIC was prompted to try to provide pricing with a sounder actuarial basis. Hogan and Luther (2014; 2016) provide evidence that premiums were substantially below the actuarially fair rate between 1999 and 2007, which contributed to the moral hazard problem and increased taxpayer losses due to bank failures between 2007 and 2010.
Later, in a subsection titled *The Bank Panics of the Early 1930s*, Hubbard and O’Brien say: “Of the 24,500 commercial banks operating in the United States in June 1929, only 15,400 were still operating in June 1934” (2018, 399). They do also say: “The large number of small, poorly diversified banks—particularly those that held agricultural loans as commodity prices fell—helped fuel the panics” (ibid., 399). But they do not explain why there were so many small undiversified banks and that the situation was different in other countries. Hubbard and O’Brien mention in passing that deposit insurance may cause moral hazard, but don’t elaborate (358). The Savings and Loan Crisis is mentioned but not discussed at length.

Mishkin in a subsection titled *Bank Panics and the Need for Deposit Insurance* presents the view that creation of deposit insurance in 1934 stabilized the banking system. He says:

> With fully insured deposits, depositors don’t need to run to the bank to make withdrawals—even if they are worried about the bank’s health—because their deposits will be worth 100 cents on the dollar no matter what. From 1930 to 1933, the years immediately preceding the creation of the FDIC in 1934, bank failures averaged more than 2,000 per year. After the establishment of the FDIC in 1934, bank failures averaged fewer than 15 per year until 1981. (Mishkin 2019, 218)

In the section *The Spread of Government Deposit Insurance Throughout the World: Is This a Good Thing?*, Mishkin says: “The answer seems to be no under many circumstances. Research at the World Bank has found that, on average, the adoption of explicit government deposit insurance is associated with less banking sector stability and a higher incidence of banking crises” (2019, 219). But he goes on to qualify: “However, the negative effects of deposit insurance appear only in countries with weak institutional environments.” In the subsection *Moral Hazard and the Government Safety Net*, Mishkin says: “Although a government safety net can help protect depositors and other creditors and prevent, or ameliorate, financial crisis, it is a mixed blessing. The most serious drawback of the government safety net stems from moral hazard…. Financial institutions with a government safety net have an incentive to take on greater risks than they otherwise would, because taxpayers will foot the bill if the bank subsequently goes belly up” (ibid., 220). Mishkin’s

53. Hubbard and O’Brien’s Figure 12.5 shows that with the establishment of the FDIC in 1934, bank suspensions fell to low levels. The caption to Figure 12.5 states: “Bank suspensions, during which banks are closed to the public either temporarily or permanently, soared during the bank panics of the early 1930s before falling to low levels following the establishment of the FDIC in 1934” (Hubbard and O’Brien 2018, 399). Notably, the graph in Figure 12.5 begins in 1920 and ends in 1940, reinforcing the idea that deposit insurance put an end to banking crises. Later on page 330 a graph shows bank failures in the U.S. between 1960 and 2016, which immediately draws the reader’s attention to the 1980s.
discussion in an online appendix titled *The 1980s Savings and Loan and Banking Crisis* presents a fairly balanced overview that discusses prior financial innovations and deposit insurance, concluding: “As a result of these forces, commercial banks and savings and loans did take on excessive risks and began to suffer substantial losses.”

Cecchetti and Schoenholtz, in their subsection *Government Deposit Insurance*, say: “Congress’s response to the Federal Reserve’s inability to stem the bank panics of the 1930s was to create nationwide deposit insurance” (2021, 367). They go on to say: “Since its inception, deposit insurance clearly helped to prevent runs on commercial banks” (ibid., 368). In the subsection *Problems Created by the Government Safety Net* they elaborate on the moral hazard problem and say: “In protecting depositors, then, the government creates moral hazard” (ibid.). The solution to the moral hazard problem is said to be further regulation: “But this safety net causes bank managers to take on too much risk, leading to the regulation and supervision that we will discuss later in the chapter” (364). The Savings and Loan Crisis is not covered.

Brandl presents deposit insurance as a beneficial institution but is careful to mention that critics disagree. He says:

> Since the 1930s the banking system has sought to instill confidence in depositors through government-sponsored deposit insurance. ... This insurance scheme seems to have worked well; the number of bank runs since the 1930s in the United States has fallen to essentially zero. As we will see later, however, deposit insurance is not a panacea. Critics of deposit insurance argue that the current system may cause as many problems as it solves. (Brandl 2017, 152)

In a subsection titled *The Savings & Loan Crisis*, Brandl says that the Savings and Loan Crisis was caused by the structure of the industry. He describes the financial deregulation and the role of deposit insurance:

> As a result of these misaligned incentives, many Savings & Loans wrote very risky loans...If these loans were successful and did not default, the Savings & Loan would be very profitable and could share this increased profit with depositors through high interest rates on deposits. If these risky loans did not pan out and fell into default, causing the Savings & Loan to fail, the depositor could simply turn to the government, who insured the deposits through the Federal Deposit Insurance Corporation, for their money. (Brandl 2017, 97–98)

In the *Lessons Learned* box titled *Need to Address Causes of Problems, Not Symptoms*, Brandl concludes: “Finally, in 1989, Congress did address the cause of the problem—the structure of the industry—and closed down the failing Savings and Loans” (2017, 98). Brandl notes: “Others have argued that even FIRREA [the
Financial Institutions Reform, Recovery, and Enforcement Act in 1989 did not address some of the deeper fundamental structural flaws in our financial system.”

Croushore says, in a subsection titled Providing a Federal Safety Net to Prevent Bank Runs: “Deposit insurance is a powerful method for preventing bank runs” (2015, 184). He also says: “Moral hazard arises from both deposit insurance and the lender-of-last-resort function. When a bank knows that its depositors do not care what the bank does because their deposits are insured by the government, the bank might make riskier investments” (ibid., 185). He then says: “To prevent these asymmetric-information problems the government must supervise and regulate banks” (185). In a subsection titled The Savings and Loan Crisis under the main section titled Failures of the Banking System, Croushore says the initial insolvency of the S&Ls was “…a classic case in banking in which a set of institutions based their decisions on historical behavior, in this case the behavior of interest rates” (162). He notes in 1980 the government enacted new laws including raising deposit coverage on accounts from $40,000 to $100,000 and says: “At this point the regulators and legislators created a moral-hazard problem. We now had a situation in which S&Ls were bankrupt but know that the regulators were not about to close them down” (163).

Ball, in a section titled Deposit Insurance, says: “Runs have occurred at individual banks but are rare, because the government has figured out how to solve the problem: deposit insurance” (2012, 292). Next, in a subsection titled Misuses of Deposits, Ball describes the moral hazard problems of excessive risk and looting of insolvent banks by bank management (ibid., 293–294). In a subsection The Problems with Deposit Insurance Ball says:

We saw that nervous depositors can cause bank runs. But they also have a positive effect: they discourage bankers from misusing deposits…. With deposit insurance…. A surge of failures can force the government to absorb part of the costs, as in the S&L crisis. Moral hazard and the absence of monitoring can end up hurting taxpayers. (Ball 2012, 296)

In a case study box titled Deposit Insurance and Banking Crises, Ball discusses the findings of Asli Demirgüç-Kunt and Enrica Detragiache (2002) and says:

Overall, the World Bank–IMF study found that the negative effects of deposit insurance outweigh the positive effects….However, there is an important qualification….The study found that deposit insurance makes crises more likely in countries with weak supervision but less likely in countries with strong supervision. (Ball 2012, 297)

All of the textbooks say deposit insurance is a necessary government safety
net. All give the impression or say outright that the 1930s bank failures were ended by deposit insurance. All also explain that deposit insurance causes moral hazard, and several books say it played a role in the Savings and Loan crisis. However, none question that the benefits of deposit insurance exceed the costs (although Brandl points out critics have this view and Mishkin and Ball note deposit insurance may destabilize less developed countries). Rather, they present moral hazard as an unfortunate consequence of necessary deposit insurance that must be monitored carefully by regulators. The textbooks present a similar view of too-big-to-fail policy, which we explore next.

Monetary policy and the Great Recession of 2008–2009

Beginning in the mid-1990s, rising housing demand combined with inelastic housing supply led to an unprecedented and sustained bout of home-price appreciation in the United States. Indices of home prices, especially in urban and coastal areas, peaked at historic highs in the summer of 2006 (Shiller 2015, 260). Home price inflation stalled and then reversed sharply in 2007. Declining home prices led to a near cessation of home construction activity in much of the country and diminished or eradicated many homeowners’ home equity, leading to high rates of home mortgage default and significant negative wealth effects. A large decline in aggregate demand that began in the home finance and construction industries rippled throughout the US economy. The housing bust culminated in the financial crisis of 2008 and triggered the Great Recession of 2008–2009, the largest and most sustained economic downturn in the United States since the Great Depression of the 1930s.

What caused this economically devastating sequence of: (1) housing boom; (2) housing bust; (3) financial crisis; and (4) recession? And what, if any, role did central banks and monetary policy play in this chain of events? While details and points of emphasis differ among competing narratives, explanations of the crisis can be broadly categorized into two major lines of argument. The first view, which we label the market-failure position, centers on the culpability of inadequately regulated, profit-seeking bankers who put short-term profits above long-term stability and engaged in excessive risk taking. On this view, mortgage lenders, commercial banks, and investment banks overindulged in subprime mortgage lending and sowed the seeds of a credit crisis and recession that was mitigated by the Federal Reserve’s responses. The second view, the government-failure position, locates the origins of the housing boom primarily in government policies
that encouraged home ownership and the erosion of lending standards. On this view, additional mistakes by monetary policymakers exacerbated the financial crisis and contributed to the length and severity of the recession that followed.

Advocates of the market-failure position are especially critical of what they term the ‘deregulation’ of financial markets during the 1990s and 2000s, which, they argue, blurred the lines between investment and commercial banks and led to excessive risk taking by both types of institutions. Advocates of this view claim that financial markets are inherently unstable and face the prospect of spontaneously emerging credit bubbles and financial crises. Regulators should have nipped excess mortgage risk in the bud during the housing boom, but they had been defanged by institutional changes over the past two decades, and then were asleep at the switch when the financial crisis was brewing (Stiglitz 2009, 332–333; Financial Crisis Inquiry Commission 2011, 53–56; Jeffers 2013; Duffie 2019). On this view, preventing solvency and liquidity crises requires strict government regulations that enforce adequate capital requirements and limit banks’ lending risk (Johnson and Kwak 2010, 205).

Advocates of the market-failure view further argue that market innovations in mortgage securitization contributed to the financial crisis by masking the apparent risks of subprime lending and reducing incentives for lender prudence. They argue that the eagerness of government sponsored enterprises54 (GSEs) and investment banks to snap up loans made by mortgage originators for the express purpose of acquiring raw material for Mortgage Backed Securities (MBS) removed originators’ skin in the game, which eroded lending standards and heightened mortgage lending risk. Additionally, they argue investors’ appetite for highly-rated MBS, particularly those composed of higher-yielding subprime mortgages, created a massive conduit of funding for high-risk subprime mortgage loans, and resulted in the “financial alchemy” of turning risky loans into highly-rated mortgage securities (Lewis 2015, 72ff.).

Advocates of the government-failure position, in contrast, argue that perverse government policies, not deregulation, were the chief cause of the financial crisis.55 While noting the role that securitization played in the financial crisis, they

---

54. The term ‘government sponsored enterprise’ refers to government-chartered mortgage banking institutions, which includes Fannie Mae and Freddy Mac. GSEs were created to channel additional funding into home mortgage markets and thereby encourage additional lending. Originally the GSEs issued investment-grade bonds which were funded by large pools of underlying ‘conforming’ (prime) mortgages. However, during the height of the housing boom, GSEs became significant participants in issuance of subprime mortgage backed securities (Financial Crisis Inquiry Commission 2011, 38ff.). GSEs mortgage investments carried an implicit government guarantee which was made explicit in the Treasury’s ‘bailout’ and takeover of the GSEs in 2008.

55. Advocates of the government-failure position also point out that deregulation can cut both ways. For instance, the partial repeal of the 1933 Glass-Steagall Act, which was one of the most significant
argue that GSEs dominated the secondary market for mortgages and that the U.S. Department of Housing and Urban Development guidelines eroded lending standards by requiring GSEs to extend more loans to low income borrowers, to accept smaller down payments, and to make larger loans relative to borrowers’ income. And while advocates of the government failure position do not necessarily defend the actions of bankers vis-à-vis subprime mortgage investments, they do argue that their incentives were shaped largely by institutional rules and perverse government policies which were the root cause of the problems (Calomiris and Haber 2014, 256–277). Additionally, advocates of the government-failure position argue that the Federal Reserve and other financial regulators contributed to the financial crisis and the Great Recession of 2008–2009 in the following three ways:

1. Monetary policy: Monetary policy was too accommodative during the housing boom and then too tight during the recession. First, the Federal Reserve inflated the housing bubble by keeping interest rates “too low for too long” (Gjerstad and Smith 2009; Taylor 2009). Excessively low interest rates helped ramp up housing demand, and hence, home prices, as lower interest rates enable buyers to affordably finance more expensive homes. Next, during the financial crisis the Fed initially sterilized its emergency lending and then implemented the policy of paying interest on bank reserves. At that point the Fed’s monetary policy was too tight to keep nominal GDP growing at its prior rate and caused deflation, high real interest rates, and an increase in cyclical unemployment (Hummel 2011; Sumner 2011; Selgin 2018, 67–97).

2. Lender of last resort and moral hazard: The existence and inconsistent application of the Fed’s ‘too big to fail’ (TBTF) policy exacerbated the financial crisis. First, during the housing boom, investment risk already encouraged by the Fed’s low-rate policy was greatly exacerbated by TBTF’s ‘financial crisis insurance’ protection against catastrophic losses. The implicit bailout guarantee encouraged deregulation measures cited as encouraging banks’ overindulgence in subprime lending, proved beneficial during the financial crisis. The Graham-Leach-Bliley Act of 1999 removed Glass-Steagall’s separation of commercial and investment banks. The market-failure view holds that removing the firewall between commercial and investment banks led to mergers which exacerbated risk levels and the too-big-to-fail problem. While this instance of moral hazard is certainly worthy of discussion, it should be also noted that eliminating the commercial-investment bank barrier allowed failing investment banks to tap into Federal Reserve lending and rescue packages. This turned out to be a crucial factor in the stabilization of Morgan Stanley and Goldman Sachs during the financial crisis. White (2010) and McDonald (2016) provide detailed discussions of Glass-Steagall, its repeal, and its role in the financial crisis of 2008. 56

Ironically, the problem of too little bank consolidation in the 1800s and 1900s that arose from restrictions on branching was turned on its head, giving way to too much bank consolidation in the 2000s. Calomiris and Haber (2014, ch. 7) argue political intrusion into bank regulation in the 1980s and 1990s led to increased market power of banks and too-big-to-fail financial institutions: “Chapter 7 drives that point home by examining how the U.S. banking system, freed of restrictions on branching and competition—a change that should have made the system more stable—became positioned during the 1990s for the
financial institutions to make highly leveraged, risky investments in MBS and other housing-related derivatives (Roberts 2010). Later, during the financial crisis in 2008, the leveraged buyout of Bear Stearns reinforced the bailout expectation, raising moral hazard in the financial system. Then, against expectations, the Fed allowed Lehman Brothers to fail, which caused a financial panic and a credit freeze. Finally, testimony in Congress on September 23rd by acting Fed Chairman Ben Bernanke and Treasury Secretary Hank Paulson created uncertainty regarding the Fed’s policy response and the TARP bailout plan, and precipitated the stock market crash. These inconsistent, on-again off-again bailout policies worsened the financial crisis (Taylor 2010, 169–173; 2012, 1022–1030).

3. Regulatory inadequacies: In the case of subprime mortgage lending in particular, it was not “deregulation,” but rather adjustments to existing financial regulations that had the unintended consequence of increasing subprime mortgage risk. According to proponents of the “regulatory capital arbitrage” theory, changes in capital regulations had an unintended consequence of incentivizing banks’ investments in shaky subprime MBS. The Basel Accord capital requirements, first promulgated in 1988, were initially envisioned as a regulatory enhancement of banks’ capital adequacy. Yet when combined with the increase in mortgage securitization, the risk-weighted Basel capital requirements created an incentive structure that prompted increased subprime lending and ultimately led to the excessive subprime mortgage risk that blew up beginning in 2007. Complacent ratings agencies allowed junk-rated subprime mortgages to be packaged into putatively investment-grade securities. Banks eventually realized that, under Basel I, they could cut the capital requirement for home mortgage portfolios from 4 percent to 1.6 percent by switching from booking loans they originated to selling off their mortgages for securitization and then investing the proceeds into MBS and derivatives (Kling 2009, 22–28). Thus the Basel Accord changes to capitalization rules, rather than “deregulating” the banks, merely “re-regulated” them in ways that would prove to be destabilizing.

Textbook coverage

Given the complexity and interconnectedness of the arguments outlined above, few economists limit the cause of the financial crisis to just one of these factors. Indeed, many if not most adherents of the government-failure viewpoint draw on several or all of the elements listed. Some in the market-failure camp draw on certain of the government-failure arguments as well. The immense global impact of the financial crisis and recession has generated a rich field for academic
study and given rise to multiple competing theories and vast literatures of supporting claims and evidence. To provide a balanced and complete treatment of this complex event for money and banking students, textbook authors should at least acknowledge and briefly outline the major competing arguments. However, all of the textbooks under review here omit discussion of at least one and often several possibly important causal factors that played into the financial crisis and recession.

Hubbard and O’Brien present a fairly balanced account, mainly descriptive but with some discussion of a few potential market and government failures. In a subsection titled Did a Global Saving Glut Cause the U.S. Housing Boom? Hubbard and O’Brien present the global savings glut explanation for low interest rates in the mid-2000s, but are careful to point out the following: “Some economists have argued that the Fed persisted in a low-interest-rate policy for too long a period, thereby fueling the housing boom.” They also conclude the section by noting: “[John] Taylor argues that Federal Reserve policy, rather than a global saving glut, fueled the housing bubble in the United States” (2018, 136). Hubbard and O’Brien (ibid., 73, 638–641) discuss the securitization of mortgages and its effects on the financial industry in descriptive terms, but note potential significant problems in the mortgage market, including political pressures that weaken mortgage standards for the GSEs. In a section titled The Financial Crisis of 2007–09 Hubbard and O’Brien describe the rise and fall of housing prices and the bank runs at Bear Stearns and Lehman Brothers in largely descriptive terms without ascribing causal factors. The next section, Financial Crisis and Financial Regulation, discusses TBTF policy and notes the moral hazard problem created by Federal Reserve bailouts of insolvent institutions (ibid., 408). Hubbard and O’Brien also question one aspect of the Fed’s handling of the financial crisis in the section Could the Fed Have Saved Lehman Brothers?, which presents the view that fear of increasing moral hazard led the Fed to allow Lehman Brothers to fail, while noting Bernanke’s claim that the Fed was legally prevented by the Federal Reserve Act from undertaking a bailout. Overall, however, Hubbard and O’Brien mainly present the view that Fed policy after the financial crisis helped shorten the recession, for example stating: “The economy started to recover in mid-2009 only after the risk premium began to decline to more normal levels. The Fed helped reduce the risk premium by undertaking unconventional policies such as buying mortgage-backed securities issued by Fannie Mae and Freddy Mac” (ibid., 638). But in the concluding box of Chapter 12, titled Answering the Key Question, Hubbard and O’Brien note: “Some economists believe that policy errors by the Federal Reserve and policy uncertainty during and after the recession explain the severity of the recession and the weakness of the recovery” (ibid., 421).

Mishkin presents an account of the financial crisis and recession that mostly
hews toward a spontaneous market-failure storyline, although he does include some nods to potential government-failure complications. In a section titled *The Global Financial Crisis of 2007–2009*, Mishkin states that “financial innovation in mortgage markets, agency problems in mortgage markets, and the role of asymmetric information in the credit-rating process” were the “three central factors” underlying the crisis (2019, 275). Mishkin says the housing bubble was driven by growth in the subprime mortgage market, with further stimulus coming from three sources: (1) low interest rates driven by large capital inflows from countries like China and India; (2) government mandates for the GSEs to invest heavily in MBS; and (3) Federal Reserve monetary policy that reduced mortgage interest rates (ibid., 278).

Mishkin blames problems in the mortgage market on agency problems of market participants including investors, brokers, commercial banks, investment banks, and rating agencies. He notes that the rating agencies in particular were problematic, saying that that they “were subject to conflicts of interest because the large fees they earned from advising clients on how to structure products that they themselves were rating meant that they did not have sufficient incentives to make sure their ratings were accurate” (2019, 277).

In another chapter Mishkin notes: “One problem with the too-big-to-fail policy is that it increases the moral hazard incentives for big banks” (2019, 221). In his expanded chapter on the financial crisis, Mishkin states that the TBTF problem “was an important factor that contributed to the global financial crisis” (ibid., 286). In his Chapter 15, *The Tools of Monetary Policy*, Mishkin presents an unqualified view that the Fed’s monetary policy and lender-of-last-resort actions during and subsequent to the financial crisis were helpful (2019, 343–365). The Inside the Fed box titled *Fed Lending Facilities During the Global Financial Crisis* states: “During the global financial crisis, the Federal Reserve became very creative in assembling a host of new lending facilities to help restore liquidity to different parts of the financial system” (ibid., 359).

Brandl does not provide any detailed overview of the Great Recession of 2008–2009, but does touch on some related topics. In a section called *The Mortgage Market, Government Policies, and the Global Financial Crisis* Brandl presents elements of both market and government failures that led to the housing bubble:

Certainly one of the major causes of the current crisis was the financial markets misuse of Gaussian copulas….Another thing driving the expansion of the CMO [collateralized mortgage obligation] market was the rapid relaxation of underwriting standards in the mortgage markets….Adding fuel to this housing asset bubble were policymakers in Washington. Both the Clinton and George W. Bush Administrations pursued increased home ownership as one of their important economic policy objectives. (Brandl 2017, 377)
Brandl presents a balanced discussion of TBTF, stating:

Many critics complain that the TBTF policies that date back to the 1980s are one of the fundamental causes of the global financial crisis that started in 2008...The debate over TBTF rages on: Do they reduce systemic risk or do they exacerbate a moral hazard that led to the worst economic slowdown since the Great Depression? (Brandl 2017, 299)

Cecchetti and Schoenholtz present mainly a spontaneous market-failure account, simply noting that lenders “relaxed their lending standards” and became “complacent” about the upward trend in home prices. They claim that the housing bust and increased defaults caught banks which had “bet the house” on subprime MBS flat-footed, leading to insolvency and a liquidity crisis (Cecchetti and Schoenholtz 2021, 161). Cecchetti and Schoenholtz do provide some details of the role of securitization in the crisis. They state that each party to the MBS chain was too reliant upon information from others and failed to perform due diligence in assessing subprime mortgage risk; the crisis thus represented a market-failure outcome involving asymmetric information and the free rider problem (ibid., 280). Cecchetti and Schoenholtz do address the regulatory inadequacy component of the government-failure argument, noting that banks had learned to “evade or game” regulations, specifically by engaging in regulatory capital arbitrage, swapping booked mortgages for highly-rated MBS, in order to reduce the burden of regulatory capital requirements under Basel (ibid., 375, 379).

Croushore offers primarily a market-failure view of the crisis, suggesting that excessive subprime mortgage risk emerged spontaneously due to myopic, under-regulated bankers who became “willing to extend mortgage loans to just about anyone who wanted one, making subprime loans (to borrowers who were high risks), and even making so-called NINJA loans (to people with No Income, No Job, or Assets)” (2015, 163). He further notes the role of extremely high leverage ratios in heightening banks’ susceptibility to negative shocks (ibid., 163), but does not comment on the possible linkage between high leverage and the Fed’s low interest rate policies. Croushore does acknowledge the moral hazard argument as a contributing factor in the crisis, noting that, with the prospect of bailouts, investment banks “may make riskier investment decisions as a result” (ibid.).

Ball likewise presents mainly a market-failure view, noting that mortgage lenders were simply myopic in extending subprime credit and performed insufficient credit underwriting. Ball also notes the role played in the buildup of subprime mortgage risk by MBS, but does not link the growth in securitization to the regulatory structure (2012, 234–235).

All of the textbooks present primarily a market-failure view of the Great Recession of 2008–2009 and an endorsement of the Fed’s unprecedented actions...
as lender-of-last-resort and implementation of unconventional monetary policy tools—although Hubbard and O’Brien and Mishkin note some competing theories. Some of the textbooks present one or another aspect of government failure, but none present a full and balanced overview of all aspects of the government-failure arguments.

The performance of the U.S. economy before and after the Federal Reserve

The Federal Reserve’s mandate for monetary policy is to promote the goals of maximum employment, stable prices, and moderate long term interest rates. The Fed also has the stated goal, as lender-of-last-resort, of containing financial disruptions and limiting their spread outside the financial sector. Whether and to what extent the Fed has or has not achieved its stated goals is a matter of much debate in the economics literature.

Merely looking at the standard Kuznets-Kendrick historical GNP series data sets shows that prices, output, and unemployment have been slightly less stable in the U.S. after the creation of the Federal Reserve System in 1913 than they were in the pre-Fed era. However, many economists prefer to exclude the early period under the Fed between 1913 and 1945 from this comparison on the grounds that this was a learning period during which the Fed had gained missing knowledge about how to conduct monetary policy successfully. Comparison between the pre-Fed era and the post-WWII Fed era formed the basis for the common claim in the 1960s and 1970s that the Fed had helped stabilize the macro-economy by reducing the frequency and duration of recessions.

Christina Romer’s (1986; 1989; 2009) influential research cast doubt on these earlier claims by arguing the seemingly greater stability in the post-WWII Fed era is merely an artifact of differences in the way the standard data sets measure output and unemployment for different time periods. Using revised data, Romer found that the volatility of output, the volatility of unemployment, and the rate of inflation were all higher in the post-WWII era than in the pre-Fed era. Romer’s findings provoked a large body of follow-up research in which some studies find the economy performed better during the pre-Fed era while other studies find the economy has performed better in the post-WWII Fed era.57

Textbook coverage

By and large, the textbooks merely assert the goal of the Fed is to stabilize unemployment and inflation. While the texts generally demur when it comes to assessing the Fed’s overall historical performance, they tend to be optimistic about the Fed’s ability to learn from past mistakes and improve macroeconomic stability going forward.

Cecchetti and Schoenholtz do not offer a general assessment of the Fed’s historical performance, yet are optimistic that the Fed and other central banks have learned from past mistakes. They cite the pre-Fed instability of the U.S. financial system as the main rationale for the creation of the Fed, but then note that “[t]he historical record is filled with examples of [central bank] failure, like the Great Depression of the 1930s… Economic historians blame the Federal Reserve for the severity of that episode” (2021, 398). They briefly note the US inflation crisis of the late 1970s but do not discuss the Fed’s role in contributing to this episode. They do partly inculpate the Fed for the Great Recession of 2008–2009: “The Fed also bears considerable responsibility for the crisis of 2007–2009. It was largely passive as intermediaries took on increasing risk amid an unprecedented housing bubble, and it allowed the financial hurricane to intensify for more than a year after the storms began” (ibid., 402). However, they go on to praise the Fed’s response to the crisis in 2008, saying: “the Fed used all its emergency authority in historically unprecedented ways to steady the financial system when the crisis peaked in 2008. The Fed’s tenacity and flexibility helped promote a huge recovery of financial conditions in 2009 and avoid a second Great Depression” (ibid.). They close on a fairly sanguine forecast of central banks’ ability to maintain economic stability in the future, stating: “Over the years, central bankers have learned from their mistakes” and are now well-positioned to engage in last resort lending during crises and deliver price stability for the economy (515).

Mishkin—a former member of the Fed’s Board of Governors—likewise offers no general assessment of the Fed’s performance. Mishkin does lay blame for the severity of the 1930s financial crises at the feet of the Fed, noting that “For more than two years, the Fed sat idly by through one bank panic after another, the most severe spate of panics in U.S. history” (2019, 273). Mishkin also documents the Fed’s role in, and response to, the inflation of the 1970s (ibid., 575), another episode widely viewed as a Federal Reserve policy error.

Ball also does not offer an overall assessment of the Fed’s long-run performance in achieving macroeconomic stabilization. Ball documents the Fed’s causal role in the Great Inflation of the 1970s and early 1980s (2012, 373–374, 460), and briefly discusses the Fed’s failure to engage in its lender of last resort function in the 1930s financial crises (ibid., 550).

The record of persistent inflation since World War II, particularly the high rates of inflation during the late 1970s and early 1980s, undercuts the claim that the Fed has emphasized price stability. Other economists argue that the Fed’s record on price stability is relatively good and that the high inflation rates of the 1970s were primarily due to soaring oil prices that took the Fed by surprise. (Hubbard and O’Brien 2018, 446)

Croushore also does not offer an overall assessment of the Fed’s long-run performance in achieving macroeconomic stabilization. He does not discuss the Fed’s role in the Great Depression or Great Recession of 2008–2009. Croushore does not present a detailed discussion of the Great Inflation of the 1970s, but in discussing inflation under the different Fed chairs says, “Perhaps the worst performance was turned in by Arthur Burns in the 1970s, under whose chairmanship inflation remained at a high level despite several recessions that helped to reduce it” (Croushore 2015, 321).


A thorough textbook on the subject of money and banking must address how well the Fed has achieved its mission. None of the textbooks provide this. Rather than making assertions about the Fed’s goals, the textbooks should provide a brief and unbiased summary of the statistical findings of the post-Romer literature on the topic.

**Summary table and discussion of possible biases**

The following table summarizes the views that each textbook presents on the seven topics.
### TABLE 1. The views of the textbooks on the seven topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>Hubbard and O'Brien</th>
<th>Mishkin</th>
<th>Ball</th>
<th>Brandl</th>
<th>Cecchetti and Schoenholtz</th>
<th>Croushore</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bank Stability, Runs, Panics</strong></td>
<td>All present only the sunspot hypothesis of inherent bank fragility; none presents regulatory weakening hypothesis or historical evidence comparing the stability of restricted vs. unrestricted banking systems</td>
<td>All present only the sunspot hypothesis of inherent bank fragility; none presents regulatory weakening hypothesis or historical evidence comparing the stability of restricted vs. unrestricted banking systems</td>
<td>All present only the sunspot hypothesis of inherent bank fragility; none presents regulatory weakening hypothesis or historical evidence comparing the stability of restricted vs. unrestricted banking systems</td>
<td>All present only the sunspot hypothesis of inherent bank fragility; none presents regulatory weakening hypothesis or historical evidence comparing the stability of restricted vs. unrestricted banking systems</td>
<td>All present only the sunspot hypothesis of inherent bank fragility; none presents regulatory weakening hypothesis or historical evidence comparing the stability of restricted vs. unrestricted banking systems</td>
<td>All present only the sunspot hypothesis of inherent bank fragility; none presents regulatory weakening hypothesis or historical evidence comparing the stability of restricted vs. unrestricted banking systems</td>
</tr>
<tr>
<td><strong>Origins of Central Banks</strong></td>
<td>Market failure</td>
<td>Market failure</td>
<td>Market failure</td>
<td>Market failure</td>
<td>Government interest and market failure</td>
<td>Market failure</td>
</tr>
<tr>
<td><strong>National Banking Era; Origins of Fed</strong></td>
<td>Strictly market failure—no coverage of legal restrictions that led to banks’ fragility; No mention of alternative plans or proposals to Federal Reserve for preventing financial crises</td>
<td>Strictly market failure—no coverage of legal restrictions that led to banks’ fragility; No mention of alternative plans or proposals to Federal Reserve for preventing financial crises</td>
<td>Strictly market failure—no coverage of legal restrictions that led to banks’ fragility; No mention of alternative plans or proposals to Federal Reserve for preventing financial crises</td>
<td>Strictly market failure—no coverage of legal restrictions that led to banks’ fragility; No mention of alternative plans or proposals to Federal Reserve for preventing financial crises</td>
<td>Strictly market failure—no coverage of legal restrictions that led to banks’ fragility; No mention of alternative plans or proposals to Federal Reserve for preventing financial crises</td>
<td>Strictly market failure—no coverage of legal restrictions that led to banks’ fragility; No mention of alternative plans or proposals to Federal Reserve for preventing financial crises</td>
</tr>
<tr>
<td><strong>Great Depression Bank Panics</strong></td>
<td>Monetarist panic view; government failure</td>
<td>Monetarist panic view; government failure</td>
<td>Monetarist panic view; government failure</td>
<td>Monetarist panic view; government failure</td>
<td>Endogenous view of bank suspensions; market and government failure</td>
<td>No detailed coverage</td>
</tr>
<tr>
<td><strong>Deposit Insurance</strong></td>
<td>Favorable view; mentions moral hazard in passing</td>
<td>Overall favorable view; presents both positive and negative aspects; cites critical literature</td>
<td>Overall favorable view; presents both positive and negative aspects; addresses moral hazard in S&amp;L crisis; cites critical literature</td>
<td>Overall favorable view; presents both positive and negative aspects; addresses moral hazard in S&amp;L crisis; cites critical literature</td>
<td>Overall favorable view; presents both positive and negative aspects; no coverage of S&amp;L crisis</td>
<td>Overall favorable view; presents both positive and negative aspects; addresses moral hazard in S&amp;L crisis</td>
</tr>
<tr>
<td><strong>Financial Crisis of 2008</strong></td>
<td>Market failure and government failure; discusses low interest rate policy and role of moral hazard</td>
<td>Market failure view; notes some sources of government failure</td>
<td>Simple market failure view</td>
<td>No detailed discussion of the Recession; market failure and government failure in housing bubble</td>
<td>Market failure; details on securitization; mentions some aspects of regulatory inadequacy</td>
<td>Market failure; mentions role of moral hazard</td>
</tr>
</tbody>
</table>
Whichever textbook is used, students basically find the following narrative: Commercial banks are naturally and inherently unstable, so central banks are created to stabilize the banking system. The United States had frequent bank panics before the creation of the Fed. A central bank was needed to end these panics, so the Fed was created for this purpose. The Fed blundered in the 1930s and made the Great Depression worse, but it learned from its mistakes and that is also why a second safety net of deposit insurance was implemented. Deposit insurance put an end to the Panic of 1933 and has virtually eliminated bank runs. Deposit insurance creates moral hazard, but the benefits are worth the costs; government can effectively manage moral hazard risk through financial regulations, although sometimes lack of sufficient regulation has manifested in problems. The financial crisis of 2008 was the product of insufficient regulation of financial markets, but the Fed had the courage to act and pursue a too-big-to-fail policy and implement unconventional policy tools. Overall, the post-WWII Fed has made the economy more stable than it was in the pre-Fed era.

Every one of these claims is disputed by prominent economists in research articles published in top academic journals. However, the textbooks do not point out that a significant number of economists have alternative views. Additionally, the textbook narrative relies on a biased selection of facts. The cumulative effect of these oversights is an unbalanced textbook narrative that could potentially mislead student readers.

To put this in perspective and illustrate how biased the textbook narrative is consider the following opposite narrative that could be constructed based solely on arguments by prominent economists in top journals: Commercial banks are generally stable and there was historically an absence of bank panics in countries with free banking systems. Legal restrictions on banks made them unstable. Central banks came into existence to support the fiscal needs of government or deal with instability engendered by legal restrictions. The Fed was created after the Asset Currency Movement was blocked for political reasons. Legal weakening of banks and the Fed helped make a recession beginning in 1929 become the Great Depression by causing the banking system to collapse. The banking crisis was halted by Roosevelt’s currency reforms but subsequently deposit insurance was unnecessarily implemented. Deposit insurance was costly but seemed to end bank failures for a while until moral hazard exploded in the Savings & Loan crisis. A combination of bad government policies and bad Fed policies created a financial crisis in 2008 and subsequently worsened the Great Recession of 2008–2009. Prices and unemployment have been less stable than they were before the Fed was created. Our desire is not for textbooks to champion one or the other of these narratives, but merely to lay out all possibilities discussed in the economic literature in a fair and balanced way and make students aware of them.
Here we suggest two underlying biases that may or may not be at play but could potentially account for the observed narrative:

1. False consensus bias. Textbook authors may simply be presenting views they perceive to be matters of general consensus in the economics profession. We believe a simple survey of economists would show there is no such consensus. A survey by Robert Whaples (1995) included questions related to two of the topics reviewed in this paper and found disagreement. When presented with the following proposition, 32 percent agreed, 43 percent disagreed, and 25 percent agreed with provisos: “Throughout the contractionary period of the Great Depression, the Federal Reserve had ample powers to cut short the process of monetary deflation and banking collapse. Proper action would have eased the severity of the contraction and very likely would have brought it to an end at a much earlier date.” On the proposition “The cyclical volatility of GNP and unemployment was greater before the Great Depression than it has been since the end of World War Two” Whaples found 54 percent agreed, 22 percent disagreed, and 24 percent agreed with provisos. 58

2. Status quo bias. At times the textbooks can almost read like an ex-post justification of whatever financial institutions or policies happen to exist at the time of publication. This suggests to us status quo bias may be at play. Taking a cursory look through out-of-print textbooks, and even at earlier editions of currently in-print textbooks, it appears to us that only matters currently under debate by the Federal Reserve’s Federal Open Market Committee are presented in the textbooks as actual matters of theoretical debate. As the status quo of the Fed shifts, so do the textbooks. For instance, the practice of central bank purchases of assets other than short-term treasury bonds is presented as inappropriate in older textbooks but appropriate in newer textbooks. This change coincides in time with the Fed’s changes to its own operating system. Status quo bias may also explain why other topics that are not matters of Fed consensus, such as rules versus discretion in monetary policy, are treated in a less biased way. 59

Another factor that may serve to reinforce the status quo bias is the Fed’s role in academic research in monetary economics (White 2005). Table 2 lists affiliations

58. It is worth noting the Whaples survey was conducted before the Great Recession, and that even during the height of the Great Moderation 22 percent of economists disagreed with the view that the Fed had made the economy more stable in the post-WWII era than in the pre-Fed era.

59. Scientism may help account for why economics textbooks in general tend to assume government intervention can successfully correct market failures because it leads to the hubris that technological knowledge alone is sufficient for government planners to successfully remold the social order of the marketplace. Scientism may also explain why textbooks tend to present a Whig view of history because it leads to the assumption that scientific knowledge is always progressing and the mistaken presumption that institutions that exist later in time, such as central banks or deposit insurance, must be superior to those that previously existed simply because they came later. On scientism generally, see Hayek 2010/1952.
that the textbook authors have had with central banks. We surmise that if the textbooks were written by authors previously employed by or affiliated with commercial banks instead of central banks the characterization of the stabilizing properties of commercial banks vis-à-vis central banks would be largely reversed.

**TABLE 2. Central bank affiliations of the authors of the six textbooks**

<table>
<thead>
<tr>
<th>Authors</th>
<th>Affiliation and Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brandl</td>
<td>None known.</td>
</tr>
<tr>
<td>Cecchetti and Schoenholtz</td>
<td>Cecchetti: Former Executive Vice President and Director of Research at the Federal Reserve Bank of New York. Schoenholtz: Visiting scholar at the Bank of Japan.</td>
</tr>
<tr>
<td>Croushore</td>
<td>Former 14 Year Vice-President of the Federal Reserve Bank of Philadelphia.</td>
</tr>
</tbody>
</table>

**Recommendations**

Our main recommendation is that the textbooks treat the seven topics surveyed in this paper with the same balance that they treat other topics such as rules versus discretion in monetary policy or the debate over central bank independence. Specifically, all of the textbooks should mention: the stability of the Canadian banking system in the absence of a central bank; the fiscal motivations for the creation of the Bank of England and other early central banks; the role that legal restrictions played in creating an unstable banking system in the U.S. during the National Banking Era; the disagreement over whether panic is to blame for the banking failures during the Great Depression and over the causes of panics; the disagreement over whether deposit insurance in fact ended the Bank Panic of 1933 and whether deposit insurance creates net benefits; the role government regulations and policies possibly played in the subprime mortgage boom and subsequent recession, including monetary policy, the Basel capital requirements, creation of the ratings agency cartel, and various housing policies that specifically required or encouraged subprime mortgage origination; and Christina Romer’s findings about the performance of the economy during the pre-Fed and post-Fed eras. In most cases these changes can be made through the addition of a few sentences, or minor revisions to existing paragraphs. These small changes alone would make for improved textbooks. We encourage the textbooks to embrace
more detailed comparative institutional analysis informed by the experiences of historical and world central banks and banking systems. In sharp contrast to popular textbooks of the past, the current textbooks present too much pure macroeconomic theory and not enough pure monetary theory and especially not enough monetary history.

We also recommend that the textbooks be screened as much as possible for false consensus and status quo bias. On matters where economists disagree, the textbooks should inform the readers of that, perhaps by presenting findings of surveys of economists’ views. Greg Mankiw (2015) has already adopted such a practice in his Principles of Microeconomics textbook, and it would make a nice addition to the money and banking textbooks. Moreover, we encourage authors to consider confessing their own ideological leanings.

The textbooks should embrace the complicated interpenetration of theory and history. Instead of ignoring the great debate of ideas in the economics profession, the textbooks should let students in on it. By presenting alternative views, when appropriate, textbooks would increase the analytical reasoning skills of students and improve their understanding of the issues. Also, in our own experience as teachers, we find that students enjoy the subject more when, instead of being treated like passive vessels of official knowledge, they are invited to question officialdom and to join an ongoing discussion.

References


**Federal Reserve Committee on Branch, Group, and Chain Banking.** 1931. *Statistical Data*. Submitted to the Subcommittee of the Committee on Banking and Currency of
the United States Senate, November 6. Link


About the Authors

Nicholas Curott is an assistant teaching professor and the assistant director of student programming for the Institute for the Study of Political Economy at Ball State University, where he teaches macroeconomics and monetary policy, and coaches the Fed Challenge Team. His research focuses on the history, theory, and political economy of banking institutions. He holds a B.A. in Economics from the University of Colorado at Boulder and a Ph.D. in Economics from George Mason University. His email address is nacurott@bsu.edu.

Tyler Watts joined the faculty of Ferris State University in 2018. His research interests include monetary history, entrepreneurship, and institutional analysis. He has published articles in *Eastern Economic Journal* and *The American Economist*. He is originally from Colorado and prior to his academic career he worked in the residential construction industry. His email address is tylerwatts@ferris.edu.

Ben Thrasher is a student at Ball State University, studying to graduate with a B.S. in mathematical economics. He is a Buchanan Scholar with the Institute for the Study of Political Economy at the Miller College of Business. Aside from a focus on mathematics, his economic study is focused on the Federal Reserve System. His email address is bthrasher@bsu.edu.

Go to archive of Intellectual Tyranny of the Status Quo section
Go to March 2020 issue

Discuss this article at Journaltalk:
https://journaltalk.net/articles/6004/
The Stewart Retractions: A Quantitative and Qualitative Analysis

Justin T. Pickett

This study analyzes the recent retraction of five articles from three sociology journals—Social Problems, Criminology, and Law & Society Review. Analyzing the retractions is important for several reasons:

• The retraction notices are vague, providing little information about what went wrong.
• The authors have continued to promote their retracted findings in print, insisting that “the main substantive results are correct” (Law & Society Review 2020).
• Other articles by the authors have some of the same irregularities (e.g., Mears et al. 2013; Mears et al. 2017; Stewart and Simons 2010; Stewart et al. 2006; Stewart et al. 2009), but thus far only one of these has been corrected and none have been retracted.
• Examining how coauthors and journal editors respond to learning about irregularities in articles sheds light on the sociology of science.

The retracted articles’ titles and abstracts are provided later; the authors of each article are as follows:


The only coauthor on all five retracted articles was Dr. Eric Stewart. He was the data holder and analyst for each article. I coauthored one of the retracted articles, Johnson et al. (2011), but here I analyze all five. I organize my analysis of the quantitative and qualitative data into three sections: (1) what happened in the articles, (2) what happened among the coauthors, and (3) what happened at the journals. Everything—data, code, emails, text messages, Excel files, drafts, and university documents—needed to verify my claims is provided online (link).

The correspondence discussed herein is public record under law, per the Freedom of Information Act. The present study received Institutional Review Board approval from the author’s university. The timeline is as follows. In February 2019, Drs. Nick Brown and James Heathers first emailed Dr. Mears to raise concerns about one of the articles. In May 2019, someone calling himself “John Smith” sent a list of irregularities in the five articles to the coauthors, journal editors, and administrators at Florida State University (FSU) (link). That email instigated a misconduct inquiry at FSU (see Pickett 2020). The retractions were announced in November 2019.

What happened in the five articles

Consistently incorrect means and standard deviations

The means and standard deviations for binary variables are connected mathematically (Heathers and Brown 2019; Schumm et al. 2019). For any specific sample size $N$, a binary variable with a given mean $P$—the proportion of the sample coded “1”—will have a standard deviation equal to:

$$SD = \sqrt{\frac{N}{N-1} \times P(1-P)}$$

In Johnson et al. (2011), the standard deviations are wrong for nine binary
variables, given the listed means. As Table 1 shows, six of these discrepancies are very large, too large to have resulted from rounding. Although these discrepancies could be due to error, it would not be a one-time error. All of the retracted articles have impossible standard deviations: there are discrepancies for six binary variables in Stewart et al. (2015), and five of them are large. In Stewart, Johnson et al. (2019), there are nine discrepancies, and six are large. In Stewart et al. (2018), there are four discrepancies and all four are large. Those are just the retracted articles. Impossible standard deviations also appear in many of Dr. Stewart’s other published articles. There are six in Mears et al. (2013, 706), three in Stewart and Simons (2010, 604), three in Stewart et al. (2009, 865), one in Stewart (2003, 591), and one in Stewart et al. (2006, 17). There were five in Mears et al. (2017, 230), and they were corrected in an erratum.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Published article</th>
<th>Mean</th>
<th>SD</th>
<th>Correct SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td></td>
<td>.86</td>
<td>.41</td>
<td>.35</td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>.10</td>
<td>.33</td>
<td>.30</td>
</tr>
<tr>
<td>Hispanic</td>
<td></td>
<td>.04</td>
<td>.22</td>
<td>.20</td>
</tr>
<tr>
<td>Married</td>
<td></td>
<td>.59</td>
<td>.31</td>
<td>.49</td>
</tr>
<tr>
<td>Education level</td>
<td></td>
<td>.42</td>
<td>.31</td>
<td>.49</td>
</tr>
<tr>
<td>(college graduate)</td>
<td></td>
<td>.43</td>
<td>.31</td>
<td>.50</td>
</tr>
<tr>
<td>Political conservative</td>
<td></td>
<td>.78</td>
<td>.33</td>
<td>.41</td>
</tr>
<tr>
<td>Own home</td>
<td></td>
<td>.17</td>
<td>.41</td>
<td>.38</td>
</tr>
<tr>
<td>Southwest</td>
<td></td>
<td>.17</td>
<td>.41</td>
<td>.38</td>
</tr>
<tr>
<td>South</td>
<td></td>
<td>.44</td>
<td>.39</td>
<td>.50</td>
</tr>
</tbody>
</table>

Notes: Large differences are those exceeding five points.

Dr. Stewart and his coauthors have offered different explanations for the discrepancies, none of which are credible. Dr. Mears told an editor they included “incorrect standard deviations for binary measures” because “the ones that we presented were mistakenly based on the formula for continuous measures.” But that is obviously untrue, for two reasons. First, there is not a separate formula for calculating standard deviations for continuous measures. Second, even if there was, modern statistical programs would not apply the wrong formula. In an email, Dr. Johnson wrote: “the standard deviations were wrong in some cases because they were based on categorical rather than binary measures (e.g., gender coded male, female, unknown).” That explanation also cannot be true. If it was, the standard

2. Email from Dr. Mears to Dr. Sterett, May 30, 2019.
3. Email from Dr. Johnson to me, October 31, 2019.
deviations would have been identical for the different racial groups (Whites, Blacks, and Hispanics) and regions (East, West, Northwest, and South). They were not.

There are other conceivable explanations for such discrepancies (e.g., misreporting sample size, unreported imputation). One possibility is that “the descriptive statistics have simply been fabricated” (Heathers and Brown 2019, 7). Fabrication can lead to such discrepancies when researchers are either unaware or forget that binary variables’ means and standard deviations are connected (Schumm et al. 2019).

Non-uniform terminal-digit distributions

One method for identifying fabricated numbers is to test whether the distribution of terminal (or rightmost) digits in reported statistics differs significantly from uniform (Diekmann 2007; Mosimann et al. 2002). The U.S. Office of Research Integrity has used this method to identify several cases of scientific fraud (Mosimann et al. 1995; 2002). It is based on Benford’s law, which describes the logarithmic distribution of the first significant (nonzero) digit, with the implication “that the distribution of higher-order digits increasingly approximates the uniform distribution” (Diekmann 2007, 323). For example, the probability of observing a specific number (0–9) in the second digit \(d_2\) is given by the formula:

\[
\text{Prob}(D_2 = d_2) = \sum_{d_1 = 0}^{9} \log \left(1 + \frac{1}{d_1 d_2}\right), \quad d_2 \in \{0, 1, \ldots, 9\}.
\]

Working through this formula reveals the second digit’s expected distribution: slightly more zeros (12 percent) than nines (9 percent), with the percentages of the other numbers (1–8) falling in between. By the third digit, each number (0–9) should appear roughly 10 percent of the time. Specifically, the expected distribution is: 0 = 10.18 percent, 1 = 10.14 percent, 2 = 10.10 percent, 3 = 10.06 percent, 4 = 10.02 percent, 5 = 9.98 percent, 6 = 9.94 percent, 7 = 9.90 percent, 8 = 9.86 percent, and 9 = 9.83 percent (Nigrini 2012).

Regression coefficients and standard errors should be Benford-distributed (Diekmann 2007; Günnel and Tödter 2009). For three-decimal regression coefficients and standard errors, the distribution of terminal digits should be approximately uniform, especially with rounding from the unreported fourth digit.\(^4\) This means that approximately 10 percent of reported terminal digits should be zeros.\(^5\)

---

4. The exception being coefficients and standard errors with less than two significant digits (e.g., \(b = .000\) or \(.001\)).
5. The terminal-digit distributions for coefficients and standard errors with only two significant digits (e.g., \(b = .020\)) should have slightly more zeros (10–12 percent, with rounding).
Fabricators, however, have difficulty generating the expected distributions for all but the first digit (Diekmann 2007). Mostly, this reflects their inability to create uniform distributions. For example, they tend to avoid ending numbers with zero, which results in terminal-digit distributions lacking the expected number of zeros (Mosimann et al. 1995).

In our article, Johnson et al. (2011), less than 2 percent of the regression coefficients and standard errors end with zero, and the terminal-digit distribution differs significantly from uniform ($\chi^2 = 26.18, p = .002$). If the true underlying distribution is uniform, we would expect to see such an extreme sample distribution by chance roughly 1 in 500 times. The numbers in the second article using the 2008 data are just as improbable. Less than 2 percent of the regression coefficients and standard errors in Stewart et al. (2015) end with zero, and the terminal-digit distribution differs significantly from uniform ($\chi^2 = 31.22, p < .001$). In each of Dr. Stewart’s more recent articles (Mears et al. 2019; Stewart et al. 2018; Stewart, Johnson et al. 2019), 2 percent or less of the coefficients and standard errors end with zero, and the terminal-digit distribution differs significantly from uniform ($\chi^2 = 61.00, p < .001$; $\chi^2 = 113.20, p < .001$; $\chi^2 = 43.70, p < .001$). In fact, in Stewart, Johnson et al. (2019), none of the coefficients or standard errors end with zero.

What does this mean? To answer this question we need to know whether the distribution of terminal digits “from a sample of non-manipulated articles is in accordance with Benford’s law” (Diekmann and Jann 2010, 398). Accordingly, I searched for articles that examined criminal justice topics and used similar methodologies—specifically, that had comparably large (or larger) samples, used multilevel modeling, estimated a series of stepwise models building from a baseline specification, and reported coefficients and standard errors to three decimals. To increase the likelihood that the comparison articles were “non-manipulated,” I only included articles that used data available to outside researchers. I also excluded articles that involved Dr. Stewart or his coauthors on the five questionable articles. I used the first ten articles I found that met these criteria.

Dr. Stewart’s five articles reported a total of 1,582 coefficients and standard errors. To generate comparison groups with a similar number of articles and statistics, I block-randomized (on the basis of the number of reported statistics)

---

6. Odds ratios and $t$-statistics are excluded because they are calculated from the coefficients and standard errors.
7. After Dr. Stewart was notified of the unusual terminal-digit distributions in these articles, he corrected two, one between ‘online first’ and print publication (Mears et al. 2019) and the other after print (Stewart et al. 2018). The numbers I report are for the original articles. The ‘corrected’ articles also have non-uniform distributions, though.
the ten articles I found into two comparison groups of five articles each. The first comparison group reported 1,332 coefficients and standard errors; the second comparison group reported 1,232. Figure 1 shows the terminal-digit distributions for Dr. Stewart’s five questionable articles and both comparison groups.

Less than 1 percent of the terminal digits in Dr. Stewart’s articles are zeros, whereas 10 percent are in the first comparison group, and 9.4 percent are in the second. Unsurprisingly, panel A in Figure 1 shows that the terminal-digit distribution in Dr. Stewart’s articles differs significantly from uniform ($\chi^2 = 252.07$, $p < .001$). ‘Differs significantly’ is an understatement; the p-value is incredibly small ($p = 3.65 \times 10^{-49}$). By contrast, in neither comparison group does the terminal digit distribution differ significantly from uniform (see panels C and E).

One possible explanation is that the standard errors in Dr. Stewart’s articles are far more stable (across models) than in the comparison articles, which means that the same terminal digits are being counted multiple times. To examine this possibility, I restricted the analysis to terminal digits in full models, defined as the most complete specification in a given model set estimated with a specific sample and outcome variable. The full-model-only, terminal-digit distributions are in panels B, D, and F. The conclusion is the same. Neither comparison group has a terminal-digit distribution that differs significantly from uniform. By contrast, the distribution for Dr. Stewart’s articles is extremely non-uniform ($\chi^2 = 87.73$, $p < .001$). If the true underlying distribution is uniform, we would expect a sample distribution as extreme as reported in Dr. Stewart’s articles by chance about 1 in two hundred trillion times ($p = 4.64 \times 10^{-15}$).

Here is Dr. Stewart’s explanation for the low frequency of terminal-digit zeros in his articles: “Although there generally weren’t a lot of zeros in the 3rd decimal place, I round 3rd place zeros either up or down. For example, if a coefficient was .000007, I would round the value to .007 or .007x10^{-3}.” (None of the articles reported rounding or scientific notation.) Unfortunately, this explanation cannot withstand empirical scrutiny. It applies only to statistics with two leading zeros (e.g., $b = .007$), yet even if we exclude all such statistics, the terminal digit distribution in Dr. Stewart’s articles is extremely non-uniform ($\chi^2 = 242.22$, $p < .001$). Additionally, even if Dr. Stewart rounded 3rd place zeros, the distribution of non-zero (1–9) terminal digits should still be approximately uniform, but it is not ($\chi^2 = 98.17$, $p < .001$). By contrast, in the two comparison groups, the distribution of non-zero terminal digits is approximately uniform (group 1: $\chi^2 = 12.10$, $p = .147$; group 2: $\chi^2 = 5.85$, $p = .663$). Regardless of how we restrict the analysis, then, the statistics in Dr. Stewart’s articles stand out for their improbability, given real data.

---

9. Memo from Dr. Stewart to Dr. Thomas Blomberg, May 28, 2019.
Figure 1. Third decimals in five Stewart articles and in two comparison groups, each including five articles by other authors.

Notes: The Figure shows the percentage of third decimals in each numerical category for the regression coefficients and standard errors in five articles for which Dr. Stewart did the analysis versus two comparison groups of five articles each. The comparison articles did not involve Dr. Stewart, but all had similar methodologies (used multilevel modeling with large samples, and estimated stepwise models building from a baseline equation).

Unverifiable surveys

Three of the retracted articles reported using data from a large ($N = 2,736$), nationally representative, dual frame (landlines and cellphones) telephone survey conducted in 2013, with a 60.8 percent response rate (Mears et al. 2019; Stewart
et al. 2018; Stewart, Johnson et al. 2019). The response rate is surprising, because it greatly exceeds the typical rate of less than 10 percent obtained by professional polling organizations (Keeter et al. 2017). It is also surprising that none of the articles listed a funding source, because a survey of this size should cost in excess of $100,000 (Guterbock et al. 2018).

Perhaps most surprising, however, is that none of the articles named the survey organization that conducted the 2013 survey. When asked via email in 2018 about the survey organization, Dr. Stewart wrote:

I teamed up with some of my grad school buddies on their survey. I helped them with some analysis on a few of their projects. In turn, they agreed to add some questions for me on their broader telephone survey. They used their students to make the calls because many of the vendors were charging so much. They did a fairly good job for their first and only telephone survey. Have you ever considered training and using undergraduate students and renting time in a CATI lab at Albany? I remember Gertz did this some when he had his research outfit at FSU. 10

According to Dr. Stewart, then, the 2013 survey was done by his friends, using their students as interviewers, and it was their first survey. The last sentence is also important. It mentions Dr. Marc Gertz, a former professor at FSU, and explains that he did similar surveys before he closed his polling firm, The Research Network (TRN).

Later, Dr. Stewart changed his story. In July 2019, he told FSU’s Inquiry Committee that it was Dr. Gertz and TRN staff who conducted the 2013 survey, but wrote: “I do not have the correspondence from Dr. Gertz and/or the Research Network staff providing the 2013 data files because they were given to me on a jump drive.” 11 There are three problems with Dr. Stewart’s new explanation. First, TRN closed in 2010. Second, when asked if he conducted the 2013 survey, Dr. Gertz wrote: “Not me, wish it were.” 12 Third, the former TRN director, Jake Bratton, wrote that he never provided Dr. Stewart data for any survey conducted after 2009.

Two of the retracted articles reported using data from a large ($N = 1,184$ to $1,379$), nationally representative, dual frame telephone survey conducted in 2008.

10. Email from Dr. Stewart to me, March 27, 2018.
11. Inquiry documents from Dr. Stewart to Inquiry Committee, July 2019. Dr. Stewart did not provide the committee any documentation or data from the 2013 survey. Instead, he provided a copy of a 2017 email from Mr. Bratton, which showed a data attachment. The data attachment was for surveys conducted by other researchers on various topics between 2000 and 2009. In the email, Mr. Bratton told Dr. Stewart, “not all of these have anything to do with racial typification [sic], so you’ll have to sort through them.”
12. Email from Dr. Gertz to me, March 27, 2018.
by TRN, with a 54.8 percent response rate (Johnson et al. 2011; Stewart et al. 2015). The articles each reported one survey, but Dr. Stewart told his coauthors and FSU’s Inquiry Committee that TRN ran multiple surveys for him in 2008 and he combined them. To support this claim, he provided copies of two emails he received from Mr. Bratton in 2008 that each showed data attachments. He did not provide the raw data to the Committee. However, Mr. Bratton disputed Dr. Stewart’s account repeatedly in writing. In one email, he wrote:

That survey in the article and those questions are N = 500. The second file sent per my email you cite was a match file of census data to merge on respondent ID based on self-report zipcode, none of the original data was included. I have no record or recollection of asking that dependent variable in a following survey and TRN was closed in 1Q 2010.13

Identical statistics after changes in...everything else

TRN finished the survey for Johnson et al. (2011) in January 2008. When my coauthors and I presented our findings over a year later, in November 2009, we reported 868 respondents. In late 2010, when we were putting the final touches on our manuscript before submitting it, we still reported 868 respondents. However, the sample size reported in our published article is 1,184. There are two problems with this. First, the source of the 316 new respondents is a mystery. “I didn’t notice differences from earlier to later versions of the paper in terms of sample sizes,” Dr. Johnson wrote after I pointed out the mysterious new respondents in 2019.14

Second, the addition of 316 new respondents had almost no effect on any of the reported statistics—means, standard deviations, regression coefficients, or standard errors. Approximately 90 percent of the statistics reported in the presentation, manuscript draft, and published article are identical to the third decimal place. To illustrate, Table 2 presents the regression results from the first three models in the manuscript draft and published article. Numbers that change are in boxes, those that do not are unboxed. Although the article has 316 more respondents than the draft, and includes an additional county-level variable (Percent Republican), almost all of the coefficients and standard errors are identical.

Years later, the same problem—sample size growth without other changes—happened again. Stewart et al. (2015) analyzed data from the same 2008 survey we used in Johnson et al. (2011), and reported the same 54.8 percent response rate, the same 96 percent completion rate, the same 10 percent verification rate

13. Email from Mr. Bratton to me, November 11, 2019.
14. Email from Dr. Johnson to me, June 6, 2019.
TABLE 2. Johnson et al. (2011) manuscript draft vs. published article: Mostly identical statistics (numbers that change are in boxes)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Manuscript draft</th>
<th>Published article</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td></td>
<td>$b$</td>
<td>SE</td>
</tr>
<tr>
<td>Criminal threat</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Econ. threat</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Political threat</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pet. Hispanic</td>
<td>-</td>
<td>-.104</td>
</tr>
<tr>
<td>Hispanic grth</td>
<td>-</td>
<td>.288*</td>
</tr>
<tr>
<td>Homicide rate</td>
<td>.016</td>
<td>.033</td>
</tr>
<tr>
<td>Concent’d dis.</td>
<td>.052</td>
<td>.086</td>
</tr>
<tr>
<td>Percent Repub.</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Percent Black</td>
<td>.043</td>
<td>.173</td>
</tr>
<tr>
<td>Pop. structure</td>
<td>-.053</td>
<td>.173</td>
</tr>
<tr>
<td>Black</td>
<td>-.639*</td>
<td>.248</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-.997*</td>
<td>.359</td>
</tr>
<tr>
<td>Age</td>
<td>.016</td>
<td>.005</td>
</tr>
<tr>
<td>Male</td>
<td>.148*</td>
<td>.056</td>
</tr>
<tr>
<td>Married</td>
<td>.208</td>
<td>.201</td>
</tr>
<tr>
<td>Education level</td>
<td>-.099</td>
<td>.198</td>
</tr>
<tr>
<td>Family income</td>
<td>-.026</td>
<td>.061</td>
</tr>
<tr>
<td>Employed</td>
<td>-.191*</td>
<td>.082</td>
</tr>
<tr>
<td>Political con.</td>
<td>.374*</td>
<td>.146</td>
</tr>
<tr>
<td>Own home</td>
<td>-.137</td>
<td>.246</td>
</tr>
<tr>
<td>Southwest</td>
<td>-.111</td>
<td>.274</td>
</tr>
<tr>
<td>Northeast</td>
<td>-.183</td>
<td>.281</td>
</tr>
<tr>
<td>West</td>
<td>.082</td>
<td>.264</td>
</tr>
<tr>
<td>Gen'l punitive</td>
<td>.191*</td>
<td>.062</td>
</tr>
<tr>
<td>Intercept</td>
<td>-.832*</td>
<td>.102</td>
</tr>
<tr>
<td>Variance</td>
<td>10%</td>
<td>15%</td>
</tr>
</tbody>
</table>

| N   | 868  | 1,184 |

*Note: Numbers that change between manuscript draft and published article are in boxes. *p < .05 (two-tailed).
2008 and 2011), and then to 1,379 respondents (between 2011 and 2015). Where these 511 respondents came from is unclear. Remarkably, the response, completion, verification, and agreement rates all remain unchanged after such substantial growth in the sample size.  

More baffling still are the descriptive statistics. The samples in these articles differed in many ways. Whereas the Johnson et al. (2011) analytic sample was racially diverse, Stewart et al.’s (2015) analytic sample was racially homogenous, including only non-Latino Whites. The Johnson et al. (2011) respondents lived in 91 counties, and those in Stewart et al.’s (2015) lived in 88 counties. In sum, these articles had different total sample sizes, different analytic sample sizes at both levels (individual and county), and analytic samples with different racial compositions. As Table 3 shows, despite all of these differences, most of the descriptive statistics in the two samples are identical (unboxed).

**TABLE 3. Johnson et al. (2011) vs. Stewart et al. (2015):**
Mostly identical descriptive statistics (numbers that change are in boxes)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Johnson et al. (2011)</th>
<th>Stewart et al. (2015)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of ethnicity in punishment</td>
<td>.31</td>
<td>-</td>
</tr>
<tr>
<td>Punitive Latino sentiment</td>
<td>-</td>
<td>10.75</td>
</tr>
<tr>
<td><strong>Independent variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived Hispanic/Latino threat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latino criminal threat</td>
<td>4.93</td>
<td>4.93</td>
</tr>
<tr>
<td>Hispanic/Latino economic threat</td>
<td>1.72</td>
<td>1.72</td>
</tr>
<tr>
<td>Hispanic/Latino political threat</td>
<td>4.38</td>
<td>4.38</td>
</tr>
<tr>
<td>Aggregate Hispanic/Latino threat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Hispanic/Latino</td>
<td>.12</td>
<td>.16</td>
</tr>
<tr>
<td>Hispanic/Latino growth</td>
<td>.26</td>
<td>.03</td>
</tr>
<tr>
<td><strong>County characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homicide rate (per 100,000)</td>
<td>3.96</td>
<td>3.96</td>
</tr>
<tr>
<td>Concentrated disadvantage</td>
<td>1.09</td>
<td>1.48</td>
</tr>
<tr>
<td>Percent Republican</td>
<td>53.04</td>
<td>.53</td>
</tr>
<tr>
<td>Percent Black</td>
<td>.10</td>
<td>.19</td>
</tr>
<tr>
<td>Population structure</td>
<td>5.39</td>
<td>5.39</td>
</tr>
</tbody>
</table>

15. In their National Science Foundation proposal, Stewart and Martinez (2010) listed the same 2008 survey with the same rates, but claimed it had a total sample size of 929, instead of 1,184 or 1,379.
In December 2010, after a colleague read a draft of our manuscript and suggested we control for county political climate, Dr. Stewart asked me to collect county voting percentages and sent me an Excel file to use. It was an individual-level file without any variables that only included case numbers and geographic identifiers. He sent this file just a few weeks before we submitted our manuscript to *Criminology*, and long after he analyzed the data and produced results mostly identical to those in the published article (see above), so it should have included the right number of respondents and counties. It did not. The file included 1,000 respondents in 292 counties, not 1,184 respondents in 91 counties (more on this shortly).

The discrepancy in sample size has an important implication: the descriptive statistics for the variable I collected should differ from those in the published article. This is true regardless of the explanation for the sample size discrepancy. In the file I sent back, the mean percent voting Republican was 53.04 with a standard deviation of 13.02. This variable (Percent Republican) had the same mean (53.04)
and standard deviation (13.02) in our published article, even though the sample size at both levels changed (from \(N_1 / N_2 = 1,000 / 292\) to \(N_1 / N_2 = 1,184 / 91\)). Although provided in proportion rather than percentage format, the variable also had the same mean (.53) and standard deviation (.13) in Stewart et al. (2015), even though the sample size at both levels changed again in that article. This kind of stability in statistics despite changes in sample size does not happen with real data.

### Inexplicable sample sizes and statistics

After we received two emails in 2019 identifying data irregularities in our article and in four others coauthored by Dr. Stewart, I asked my coauthors to send me the full data for Johnson et al. (2011). Two things happened. First, Dr. Johnson told me he did not have a copy and had never seen the data. Second, I encountered difficulties getting the data from Dr. Stewart (more on this later). Consequently, I examined the limited Excel file I already had, which Dr. Stewart had sent in December 2010, shortly before we submitted our article. I discovered the file had only 1,000 respondents, not the 1,184 reported in the article, and that only 500 of those respondents were unique; the other 500 were duplicates. I informed my coauthors about the duplicates and other issues. Dr. Gertz then contacted the former TRN director, who confirmed that the survey he ran for us included only 500 respondents.

On June 10, 2019, Dr. Stewart finally shared with Dr. Johnson and me a copy of the data for our article. At that time he admitted “there are 300+ county units and 500 individuals.” Dr. Stewart wrote me to explain how the duplication happened: “I thought I was merging files from two different surveys for which I had questions. I merged the wrong the file.” This explanation is problematic, because Johnson et al. (2011) and Stewart et al. (2015) both described one survey, not two. Ten days later, Dr. Stewart gave Dr. Johnson the same explanation for the duplicates, “I received multiple files, but mistakenly merged the incorrect one,” but he added something new: “I found the correct data files that should have been merged.” Several things here are problematic.

First, the former director of the TRN has said repeatedly that there was only one sample, and it included only 500 respondents. He did send Dr. Stewart two files, but they were for the same survey and included the same respondents. In one email, Mr. Bratton wrote: “The second file sent per my email you cite was a match file of census data to merge on respondent ID based on self-report zipcode, none

---

16. Email from Dr. Stewart to me, June 10, 2019.
17. Email from Dr. Stewart to me, June 10, 2019.
18. Email from Dr. Stewart to Dr. Johnson, June 20, 2019.
of the original data was included.” In another email, he wrote:

Dr. Stewart was sent exactly 500 records that January 2008 … [and] in May of 2008 I sent another SPSS file that only had the names/census info of the counties for the first ~420 and I think he had to [look] up the county names by zip code in the file for the other 80 by hand to do his analysis.  

Second, accidentally doubling the sample would not have yielded the numbers reported in our published article. Dr. Stewart claimed he accidentally included the 500 duplicates in his analysis for our article because he “merged the wrong the file.” He clarified that “the correct data files that should have been merged” were not. This is important, so it bears repeating. Dr. Stewart has said that the duplicates were included in the analysis for our published article and that the explanation for the duplicates is accidental doubling. However, accidentally doubling the full sample of 500 respondents, would have resulted in 1,000 respondents, not the 1,184 reported in our article, nor the 1,379 reported in Stewart et al. (2015).

**Figure 2.** Johnson et al. (2011): Published article vs. shared data

Additionally, the descriptive statistics in the published article should match those in the data, even if the 500 respondents were accidentally doubled. Doubling the sample would increase the sample size, but it would not change its composition. However, the descriptive statistics in the published article differ substantially from the shared data. The outcome variable in our analysis is public support for the use

---

19. Email from Mr. Bratton to me, November 11, 2019.
20. Email from Mr. Bratton to Dr. Gertz, June 7, 2019.
21. Email from Dr. Stewart to me, June 10, 2019.
22. Email from Dr. Stewart to Dr. Johnson, June 20, 2019.
of defendants’ ethnicity in sentencing decisions. The distribution of the outcome variable by respondents’ race is shown on page 419 of Johnson et al. (2011). Figure 2 shows how it compares to the shared data ($N = 500$). Doubling the sample and then adding another 184 respondents (to get the reported sample size of 1,184) cannot explain the discrepancies. For example, even doubling the sample to 1,000 and then adding 184 Black respondents who all oppose ethnicity-based sentencing would reduce the percent of Blacks supporting it from 38 percent to 13 percent, not to the article’s 3 percent.

TABLE 4. Johnson et al. (2011): Descriptive statistics in published article vs. shared data (significant differences are in boxes)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Published article</th>
<th></th>
<th>Shared data</th>
<th></th>
<th>p-value for difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use of ethnicity in punishment</td>
<td>.31</td>
<td>.46</td>
<td>.37</td>
<td>.48</td>
<td>.007</td>
</tr>
<tr>
<td>Hispanic criminal threat</td>
<td>4.93</td>
<td>1.66</td>
<td>2.85</td>
<td>2.67</td>
<td>.000</td>
</tr>
<tr>
<td>Hispanic economic threat</td>
<td>1.72</td>
<td>1.13</td>
<td>1.64</td>
<td>1.13</td>
<td>.137</td>
</tr>
<tr>
<td>Hispanic political threat</td>
<td>4.38</td>
<td>1.41</td>
<td>1.95</td>
<td>1.59</td>
<td>.000</td>
</tr>
<tr>
<td>Percent Hispanic</td>
<td>.12</td>
<td>.11</td>
<td>9.52</td>
<td>11.61</td>
<td>.000</td>
</tr>
<tr>
<td>Hispanic growth</td>
<td>.26</td>
<td>1.53</td>
<td>3.18</td>
<td>3.11</td>
<td>.000</td>
</tr>
<tr>
<td>Homicide rate (per 100,000)</td>
<td>3.96</td>
<td>4.37</td>
<td>3.92</td>
<td>4.25</td>
<td>.819</td>
</tr>
<tr>
<td>Concentrated disadvantage</td>
<td>1.09</td>
<td>1.53</td>
<td>1.40</td>
<td>.94</td>
<td>.000</td>
</tr>
<tr>
<td>Percent Republican</td>
<td>53.04</td>
<td>13.02</td>
<td>52.56</td>
<td>13.07</td>
<td>.415</td>
</tr>
<tr>
<td>Percent Black</td>
<td>.10</td>
<td>.14</td>
<td>11.63</td>
<td>11.98</td>
<td>.000</td>
</tr>
<tr>
<td>Population structure</td>
<td>5.39</td>
<td>.70</td>
<td>5.11</td>
<td>1.00</td>
<td>.000</td>
</tr>
<tr>
<td>White</td>
<td>.86</td>
<td>.41</td>
<td>.85</td>
<td>.36</td>
<td>.606</td>
</tr>
<tr>
<td>Black</td>
<td>.10</td>
<td>.33</td>
<td>.10</td>
<td>.30</td>
<td>1.000</td>
</tr>
<tr>
<td>Hispanic</td>
<td>.04</td>
<td>.22</td>
<td>.05</td>
<td>.21</td>
<td>.494</td>
</tr>
<tr>
<td>Age</td>
<td>47.12</td>
<td>19.72</td>
<td>46.41</td>
<td>16.98</td>
<td>.352</td>
</tr>
<tr>
<td>Male</td>
<td>.47</td>
<td>.50</td>
<td>.46</td>
<td>.50</td>
<td>.591</td>
</tr>
<tr>
<td>Married</td>
<td>.59</td>
<td>.31</td>
<td>.61</td>
<td>.49</td>
<td>.275</td>
</tr>
<tr>
<td>Education level (college graduate)</td>
<td>.42</td>
<td>.31</td>
<td>.42</td>
<td>.49</td>
<td>.902</td>
</tr>
<tr>
<td>Family income</td>
<td>$62,700</td>
<td>$14,210</td>
<td>$61,196</td>
<td>$22,593</td>
<td>.137</td>
</tr>
<tr>
<td>Employed</td>
<td>.46</td>
<td>.50</td>
<td>.55</td>
<td>.50</td>
<td>.000</td>
</tr>
<tr>
<td>Political conservative</td>
<td>.43</td>
<td>.31</td>
<td>.70</td>
<td>.46</td>
<td>.000</td>
</tr>
<tr>
<td>Own home</td>
<td>.78</td>
<td>.33</td>
<td>.78</td>
<td>.41</td>
<td>.829</td>
</tr>
<tr>
<td>Southwest</td>
<td>.17</td>
<td>.41</td>
<td>.16</td>
<td>.37</td>
<td>.721</td>
</tr>
<tr>
<td>Northeast</td>
<td>.15</td>
<td>.35</td>
<td>.15</td>
<td>.36</td>
<td>.802</td>
</tr>
<tr>
<td>Midwest</td>
<td>.24</td>
<td>.43</td>
<td>.24</td>
<td>.43</td>
<td>.917</td>
</tr>
<tr>
<td>West</td>
<td>.17</td>
<td>.38</td>
<td>.17</td>
<td>.38</td>
<td>.812</td>
</tr>
<tr>
<td>South</td>
<td>.44</td>
<td>.39</td>
<td>.43</td>
<td>.50</td>
<td>.787</td>
</tr>
<tr>
<td>General punitive attitudes</td>
<td>6.84</td>
<td>2.16</td>
<td>4.69</td>
<td>3.57</td>
<td>.000</td>
</tr>
</tbody>
</table>
There are other noteworthy distributional differences. Table 4 compares all of the descriptive statistics in the published article to those in the shared data. Some of the differences are simply impossible. In the article, for example, the 1,184 respondents had a mean of 4.38 on the Hispanic political threat index. In the shared data, the 500 respondents have a mean of only 1.95. Working through the math reveals that the 684 additional respondents included in the article (but not in the data) must have had an average score of 6.16 on the index. The problem is that this average score is higher than the highest possible value on the index, which is 6.

Similarly, the published article claims that 43 percent of respondents are political conservatives. In the shared data, 70 percent are political conservatives. Even doubling the sample to 1,000 and then adding 184 liberals would only drop the percentage of conservatives in the sample to 59 percent, not to the 43 percent reported in the article. Additionally, the mean for Hispanic criminal threat is almost two points higher (mean = 4.93 vs. 2.85), and the mean for general punitive attitudes is over two points higher (mean = 6.84 vs. 4.69), in the published article than in the shared data. Even doubling the sample and then adding 184 respondents with the highest possible value (a value of 9) on these two variables would only increase their means to 3.81 and 5.36, respectively—both still a point lower than in the article.

TABLE 5. Johnson et al. (2011):
Interaction effects in published article vs. shared data
(notable differences are in boxes)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Published article</th>
<th>Shared data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( b )</td>
<td>SE</td>
</tr>
<tr>
<td>Perceived Hispanic threat</td>
<td>.183*</td>
<td>.079</td>
</tr>
<tr>
<td>Criminal threat</td>
<td>.272*</td>
<td>.111</td>
</tr>
<tr>
<td>Economic threat</td>
<td>.008</td>
<td>.116</td>
</tr>
<tr>
<td>Political threat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggregate Hispanic threat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent Hispanic</td>
<td>-.089</td>
<td>.766</td>
</tr>
<tr>
<td>Hispanic growth</td>
<td>.334**</td>
<td>.127</td>
</tr>
<tr>
<td>Interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criminal ( \times ) His. Growth</td>
<td>.126</td>
<td>.051</td>
</tr>
<tr>
<td>Economic ( \times ) His. Growth</td>
<td>.175</td>
<td>.073</td>
</tr>
<tr>
<td>Political ( \times ) His. Growth</td>
<td>-.101</td>
<td>.087</td>
</tr>
<tr>
<td>Intercept</td>
<td>-.848***</td>
<td>.119</td>
</tr>
</tbody>
</table>

Notes: * \( p < .05; \)** \( p < .01; \)*** \( p < .001 \) (two-tailed).

Accidentally doubling the sample of 500 respondents would leave the regression coefficients unscathed—they would be identical if all respondents were
duplicated. However, the regression results in the published article differ substantially from those in the shared data. Most notably, the main findings in the article—the interaction effect of perceived Hispanic threat (criminal and economic) and Hispanic growth—do not emerge with the shared data. Those findings are reported on page 423 of Johnson et al. (2011). In the shared data, none of the coefficients for the interaction terms are statistically significant and they are all much smaller. I show the comparison in Table 5.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Published article</th>
<th>Shared data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent Hispanic</td>
<td>-.104 .662 .901 .021 .020 .021</td>
<td>.021 .020 .021</td>
</tr>
<tr>
<td>Hispanic growth</td>
<td>.288** .102 1.334 .070 .053 .933</td>
<td>-.013 .033 .987 -.031 .030 .970</td>
</tr>
<tr>
<td>Homicide rate (per 100,000)</td>
<td>-.013 .033 .987 -.031 .030 .970</td>
<td>.051 .086 1.052 .033 .144 1.034</td>
</tr>
<tr>
<td>Concentrated disadvantage</td>
<td>.051 .086 1.052 .033 .144 1.034</td>
<td>.001 .005 1.000 .009 .009 1.009</td>
</tr>
<tr>
<td>Percent Republican</td>
<td>.001 .005 1.000 .009 .009 1.009</td>
<td>-.038 .173 .963 .003 .012 1.003</td>
</tr>
<tr>
<td>Percent Black</td>
<td>-.038 .173 .963 .003 .012 1.003</td>
<td>-.051 .174 .950 .068 .126 1.070</td>
</tr>
<tr>
<td>Population structure</td>
<td>-.051 .174 .950 .068 .126 1.070</td>
<td>.001 .005 1.000 .009 .009 1.009</td>
</tr>
<tr>
<td>Black</td>
<td>-.628* .248 .534 .949** .361 2.582</td>
<td>-.013 .033 .987 -.031 .030 .970</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-.982** .359 .375 -.752 .547 .472</td>
<td>.051 .086 1.052 .033 .144 1.034</td>
</tr>
<tr>
<td>Age</td>
<td>.016* .005 .016 .004 .006 1.004</td>
<td>.016* .005 .016 .004 .006 1.004</td>
</tr>
<tr>
<td>Male</td>
<td>.164** .056 1.178 -.111 .199 .895</td>
<td>.056 .016 1.056 .016 .016 1.056</td>
</tr>
<tr>
<td>Married</td>
<td>.208 .201 1.231 .425 .210 1.529</td>
<td>.001 .005 1.000 .009 .009 1.009</td>
</tr>
<tr>
<td>Education level</td>
<td>-.099 .198 .906 .215 .226 1.240</td>
<td>-.099 .198 .906 .215 .226 1.240</td>
</tr>
<tr>
<td>Family income</td>
<td>-.026 .061 .974 .062 .096 1.064</td>
<td>-.026 .061 .974 .062 .096 1.064</td>
</tr>
<tr>
<td>Political Conservative</td>
<td>.367* .146 1.443 .360 .218 1.434</td>
<td>.367* .146 1.443 .360 .218 1.434</td>
</tr>
<tr>
<td>Own home</td>
<td>-.137 .246 .872 .158 .281 1.171</td>
<td>-.137 .246 .872 .158 .281 1.171</td>
</tr>
<tr>
<td>Southwest</td>
<td>-.111 .274 .895 .721 .369 2.057</td>
<td>-.111 .274 .895 .721 .369 2.057</td>
</tr>
<tr>
<td>Northeast</td>
<td>-.183 .281 .832 .522 .334 1.686</td>
<td>-.183 .281 .832 .522 .334 1.686</td>
</tr>
<tr>
<td>Midwest</td>
<td>.054 .230 1.055 .102 .291 1.107</td>
<td>-.183 .281 .832 .522 .334 1.686</td>
</tr>
<tr>
<td>West</td>
<td>.082 .264 1.085 -.350 .362 .705</td>
<td>.082 .264 1.085 -.350 .362 .705</td>
</tr>
<tr>
<td>Intercept</td>
<td>-.858*** .105 -.640*** .097</td>
<td>-.858*** .105 -.640*** .097</td>
</tr>
</tbody>
</table>

The main effect of Hispanic growth also fails to replicate in the shared data; indeed, the coefficient is in the opposite direction. This is the case even when the interaction terms are removed from the model. In the published article, the
main effect of Hispanic growth is shown in Model 2 on page 420, and is positive and statistically significant ($b = .288, p < .01$). In the shared data, the coefficient is negative and non-significant. Table 6 compares the estimates in Johnson et al. (2011) to those from the shared data. The differences are striking, extending to many other variables besides Hispanic growth. For example, the coefficient for Black in the published article is negative and significant ($b = -0.628, p < .05$), but it is positive and significant in the actual data ($b = 0.949, p < .01$). Again, the regression results would be identical if the sample was accidentally doubled. They are not, so it is impossible that the only change to the data was accidental doubling.

**Unreported, implausible county clusters**

In the data Dr. Stewart shared, there are 500 respondents and they are nested in 326 counties. In Johnson et al. (2011), however, we claimed to have 1,184 respondents nested in 91 counties. Similarly, Stewart et al. (2015) used the same data and claimed the respondents in their analytic sample lived in 88 counties. Dr. Stewart has acknowledged that the county numbers reported in both articles are wrong. The explanation he gave to FSU’s Inquiry Committee is that he created “county units” by clustering the 326 counties. There are at least three problems with the explanation.

First, there is little justification for grouping together so many counties in either Johnson et al. (2011) or Stewart et al. (2015), given the time period of the studies. It is not desirable to group counties, because it throws away geographic detail and creates “meaningless socio-political entities” (Hagen et al. 2013, 770). Typically, researchers only group counties when their boundaries change during measurement years. Therefore, county clusters normally appear only in historical studies that examine data over a large number of decades or across centuries, and even in those studies the researchers only create county clusters for those specific counties that have boundaries that changed during the time period examined (e.g., King et al. 2009; Messner et al. 2005). To see if there was any justification for grouping counties in Johnson et al. (2011) or Stewart et al. (2015), I investigated how the specific counties in our data changed during the measurement years for the aggregate variables. For these studies, boundary changes would justify the creation of exactly one cluster, made from four counties in Colorado: Adams County, Boulder County, Jefferson County, and Weld County ([link]). That would leave 322 separate counties and one cluster, or 323 county units.

Second, neither article mentioned county clusters or reported any grouping together of counties. FSU’s Inquiry Committee noted this in its final report:

Dr. Stewart acknowledged that, although the paper referred to counties and
reported descriptive statistics for counties, in performing the analyses, he aggregated the counties into clusters. He provided no explanation for this decision, but committee members note that aggregating the counties allowed Dr. Stewart to use what was then an emerging statistical approach that could not have been supported by the county-level data.

**Figure 3.** Dr. Stewart’s description of samples in FSU inquiry document

|-----------|--------------------------------------------------------------------------------------------------|
| Related data file(s): | • Final Dataset 010808 (500 cases)  
• Fsua050108 (343 cases)  
• Fsua121407 (164 cases) |
| Log(s) for the data file(s): | (See Appendix B) |

|-----------|--------------------------------------------------------------------------------------------------|
| Related data file(s): | • Final Dataset 010808 (500 cases)  
• Fsua050108 (343 cases)  
• Fsua121407 (164 cases)  
• Final Fsuo08 (425 cases)  
• RACE FINAL 2013 (1079 cases) |
| Log(s) for the data file(s): | (See Appendix B)  
**Preliminary analyses are available. No corrections have been submitted to the journal yet. (Appendix C)** |

Actually, the situation is quite a bit worse, as Figure 3 shows. Specifically, in the documents Dr. Stewart submitted to FSU, he claimed he included the same data used in our 2011 article in each of his later articles. Of course, that differs from what the 2018 and 2019 articles reported; they all reported a single 2013 survey of 2,736 respondents: “The telephone surveys were conducted during the spring, summer, and fall of 2013” (Stewart, Johnson et al. 2019, 200). Nevertheless, Dr. Stewart told the Committee the 2013 sample only included 1,079 respondents (see Figure 3), and he combined that smaller sample with the data he used in
Johnson et al. (2011). The problem is that one of the files Dr. Stewart supposedly combined—“Final Dataset 010808 (500 cases)”—included 326 counties by itself. However, the five articles only reported between 79 and 168 total counties. That leaves two possibilities: Dr. Stewart clustered the counties down to a smaller number in all five articles but failed to report it in any of them, or he fabricated the number of counties in each article.

**Figure 4.** Research Network data: The 326 counties clustered liberally into 176 ‘county units’ (clusters and counties)

![Image](image.png)

*Notes:* Counties clustered with adjacent counties in the same state are in blue (N = 67).
Counties isolated from other counties in the state are in red (N = 109).

Finally, there does not appear to be any reasonable clustering method that would yield the number of counties reported in the articles. For example, consider the article I coauthored, Johnson et al. (2011), which reported 91 counties. Dr. Stewart has not provided the code for generating county clusters or any cluster-level data. However, I have tried different clustering methods, and none yield anywhere close to 91 county units (clusters and separate counties). The most liberal method, which yields the smallest number of county units, is to group together any counties in the data that are within the same state and share a border, either directly or indirectly via another county. Figure 4 shows the results of this clustering method. County clusters are in blue, and separate counties—counties that are not adjacent to other sampled counties in the same state—are in red. Even this method...
fails to come close to producing 91 county units. Instead, it yields 109 separate counties and 67 county clusters, for a total of 176 county units.

**What happened among the coauthors**

The coauthors of the five retracted articles include two past editors of *Criminology*, the flagship journal of the American Society of Criminology (ASC), as well as three ASC Fellows and two ASC vice presidents. Two coauthors, Brian Johnson and Eric Baumer, have written articles about the importance of research ethics (e.g., “What Scholars Should Know about ‘Self-Plagiarism’” (Lauritsen et al. 2019); “Salami-Slicing, Peek-a-Boo, and LPUS: Addressing the Problem of Piecemeal Publication” (Gartner et al. 2012)). Thus, how these scholars responded to learning about irregularities in their own articles is likely to be informative about the research norms and practices in criminology and sociology.

To my knowledge, none of the coauthors have spoken publicly about what happened in the retracted articles, except to insist in the retraction notices that the irregularities resulted from “coding mistakes” and “transcription errors” (*Law & Society Review* 2020; *Criminology* 2020a, b), and to defend the accuracy of the retracted findings (*Law & Society Review* 2020). Additionally, several of the coauthors also coauthored other papers with Dr. Stewart that have irregularities (e.g., Mears et al. 2013; Stewart et al. 2009).

**Drs. Eric Stewart and Brian Johnson**

On May 28, 2019, when I first asked Dr. Stewart for the data for Johnson et al. (2011), he said to wait a week. A week later, he delayed again, claiming: “I have to wait until after I talk to Research Board.” I asked him, “why won’t they let you share the data with your coauthor?” He ignored my question. So, I contacted FSU’s Office of Research (OR), who disputed his claim: “Sharing copies of the data is perfectly okay. Please just do not alter the original data set in any way.” I forwarded the OR’s email to all of my coauthors, including Dr. Johnson, but Dr. Stewart still withheld the data. It was only after I identified the 500 duplicates in the limited file that I already had that Dr. Stewart shared a copy of the data. He explained, “I had to recreate the data file with no duplicates.” In response,

---

23. Text message from Dr. Stewart to me, June 6, 2019.
24. Text message from me to Dr. Stewart, June 6, 2019.
25. Email from Ms. Diana Key to me, June 6, 2019.
26. Email from Dr. Stewart to me, June 9, 2019.
I asked for the original data with duplicates included, so I could try to replicate the published results. Dr. Stewart replied saying he had destroyed the original file: “corrections have been made and saved.”27 He did this despite the OR’s directive not to alter the data in any way.

Later, when Dr. Johnson reached out to Dr. Stewart—“I haven’t heard from you in a bit … have you had any luck locating the final dataset with the same sample size used in the 2011 paper?”—Dr. Stewart replied saying that he found another 425 respondents from a second survey. But neither he nor Dr. Johnson told me. When Dr. Stewart eventually sent Dr. Johnson the new output for the combined sample (N = 925), he told Dr. Johnson he was not going to share it with me. Eventually, I heard about the new developments from someone who was not a coauthor, and I emailed Dr. Stewart. Dr. Stewart refused to answer my questions about the second survey. So I contacted Dr. Johnson, explained that I could not get a response from Dr. Stewart, and asked him to send me the new output. “Sorry I’m just getting back to you, I needed a few days to consider your request,” Dr. Johnson replied after several days, “Eric asked that I not share additional output…[and] I think it is best under the circumstances that any data/output comes directly from Eric.”29

Dr. Johnson asked Dr. Stewart for the combined data in June, “if you don’t mind sending me a copy of the updated 925 data I’d like to look through it to see if I can refute any of Justin’s claims.”30 Dr. Stewart replied to him immediately, “I will send it. Let me pair [sic] it down to just the variables we use. I will send in a bit.”31 However, he never followed through. “Eric never did send me the data but he did allow me to remotely access his computer while he was running the new models,” Dr. Johnson wrote months later.32 Despite never receiving the data, despite knowing that Dr. Stewart had deleted data against the ORI’s directive, and despite being the lead author on our article, Dr. Johnson went along with Dr. Stewart’s request to withhold information from a coauthor (me) for four months.

Dr. Marc Gertz

Dr. Marc Gertz coauthored two of the five retracted articles. After I told Dr. Gertz there were hundreds of duplicates in Johnson et al. (2011), he emailed Jake Bratton, the former TRC director, to double-check. Mr. Bratton confirmed to him...

27. Email from Dr. Stewart to me, June 10, 2019.
28. Email from Dr. Johnson to Dr. Stewart, June 19, 2019.
29. Email from Dr. Johnson to me, July 1, 2019.
30. Email from Dr. Johnson to Dr. Stewart, June 27, 2019.
31. Email from Dr. Stewart to Dr. Johnson, June 27, 2019.
32. Email from Dr. Johnson to me, October 31, 2019.
that there was only one 2008 sample and that it only had 500 respondents, writing: “Basically the choices are inexperience in blending data resulting in loading the data twice and not noticing or more sample size was needed to get to \( p < .05 \).” What did Dr. Gertz do after learning from two different sources—Mr. Bratton and me—that an article he coauthored reported over twice as many respondents as were actually interviewed? He emailed Dr. Stewart a letter of support. In it, he wrote:

> At a minimum, Eric Stewart, my colleague, received data sets from me in 2008, 2009, 2013, 2017, and 2018 investigating various topics around the influence of racism in our culture. It is possible he had access to other data sets from me as well. Anyone who claims Eric had access to only one national survey is clearly incorrect.

The letter did two important things. It seemingly disputed my claim that there was only one survey done for our 2011 article, and it implied that the 2013 data came from Dr. Gertz. Dr. Gertz wrote this letter despite what Mr. Bratton told him, and despite having previously written to me that he did not conduct the 2013 survey (“Not me, wish it were”). Dr. Stewart then provided the letter to his coauthors, the journals, and FSU to support his claims about both the 2008 and 2013 surveys. The letter became a key piece of evidence in their investigations. As late as September, one editor wrote: “the only response I have from the authors is an email from someone confirming he did the survey,” to which another editor replied “we received the same letter from the person who supposedly conducted the surveys in question.”

**Dr. Daniel Mears**

Dr. Mears was the lead author on the 2019 *Law & Society Review* article and a coauthor on Stewart et al.’s (2018) *Criminology* article. After being contacted by Drs. Brown and Heathers in February, Dr. Mears coauthored corrections to both articles that were mathematically impossible. For example, the reported age difference between the original and corrected samples required the 140 dropped respondents to have a negative mean age (Polyacantha 2019). When alerted to these impossibilities, Dr. Mears sent an email to Dr. Sterett, the editor of *Law & Society Review*, disputing them and attacking Drs. Brown and Heathers. He stressed “the lack of clear credentials that these individuals have for leveling post-hoc critiques,”

33. Email from Mr. Bratton to Dr. Gertz, June 7, 2019.
34. Email from Dr. Gertz to Dr. Stewart, August 6, 2019.
35. Email from Dr. Sterett to Dr. Linders, September 25, 2019.
36. Email from Dr. Linders to Dr. Sterett, September 25, 2019.
and said “‘data thugs appear intent on maligning researchers, journal editors, programs, and universities under the guise of ‘advancing’ science.’”

Months later, after I posted a preprint on SocArXiv describing the problems in the 2008 data (Pickett 2019), Dr. Mears wrote another email to Dr. Sterett disputing what I wrote. He also provided Dr. Gertz’s letter of support to prove he conducted the 2013 survey, writing: “there was no funding source … The survey data were freely provided to Dr. Stewart by Dr. Marc Gertz.” Over a month later, Dr. Mears emailed Dr. Sterett and once again defended Mears et al. (2019), claiming that the 61 percent response rate he reported “was comparable to well-conducted national studies using random-digit dialing techniques (Czajka and Beyler 2016).”

The article he cited focused on government surveys; it said the National Immunization Survey (NIS) and the Behavioral Risk Factor Surveillance System, two highly-funded telephone surveys, achieved 2013 response rates of 62 percent and 46 percent, respectively. NORC’s five-year contract to administer the NIS is valued up to $163,658,456 (link). Well conducted, indeed.

Dr. Mears did these things—coauthored the corrections, disparaged “data thugs,” impugned me, repeatedly reassured the editor—apparently without analyzing the data himself, or even verifying the accuracy of the data description in the articles. For example, Dr. Stewart told FSU that in both articles he combined samples from different years, but failed to report it. However, Dr. Mears said nothing about this in his emails to the editor, which means he either did not know about it or he withheld the information. Recall, too, that Dr. Mears gave the editor a false explanation for the mean/SD discrepancies, so he apparently did not verify the information that he himself communicated, either.

**Drs. Eric Baumer and Patricia Warren**

Drs. Baumer and Warren each coauthored two of the retracted articles, and also coauthored corrections to their articles that were mathematically impossible, apparently without verifying the data themselves. On June 10, I alerted them, and five other coauthors, about the problems in the 2008 data (e.g., inclusion of duplicates, county number discrepancy, differences in findings, etc.). I reminded them the information “has direct implications for the 2015 Social Problems article (Stewart, Martinez, Baumer, and Gertz), which uses the same data, and indirect implications for all of the other articles.” My email included one, seemingly

---

37. Email from Dr. Mears to Dr. Sterett, May 30, 2019.
38. Email from Dr. Mears to Dr. Sterett, August 7, 2019.
39. Email from Dr. Mears to Dr. Sterett, September 29, 2019.
40. Dr. Stewart said he combined 1,432 respondents interviewed in 2007 and 2008 with 1,079 interviewed in 2013. That sums to a total sample size of 2,511, which is 225 less than the 2,736 reported in the articles.
noncontroversial recommendation: “I am sending you this because, if you haven’t yet, you may want to request a copy of the data from the article you are on, and examine it yourself.”

Dr. Baumer replied to everyone, writing: “my own view is that it is important to give Eric Stewart a chance to address these questions before drawing conclusions or taking further steps.”

His clear suggestion was that none of the coauthors should ask for the data, which is the only thing that I had recommended they do. Shortly thereafter, Dr. Baumer and I talked on the phone and he told me he refused to request the data from Dr. Stewart, who he emphasized was his close friend. He also said that although he had never seen the data, he had nonetheless consulted with an editor about the data irregularities, and had also offered to travel to FSU to meet with university officials on Dr. Stewart’s behalf.

After receiving my June 10 email about the data problems, Dr. Warren also replied to everyone, and seemingly suggested that it would be unprofessional or even disrespectful to ask for the data. She wrote: “I plan to operate with full professionalism and scholarly respect. I too choose to give Eric Stewart time to work through all the issues.”

Almost two months later, she contacted journal editors to defend Dr. Stewart and, apparently, to blame me. “The email we just received from Patricia Warren…ends up shifting the blame to the whistle blower instead,” the editors of Social Problems wrote on August 6.

Later in August, Dr. Warren filed a complaint about me with the police. It was about my June 10th reply to her and Dr. Baumer’s emails, where I had written, “Let me stop this before it turns into an assault.” I meant ‘an assault on me,’ and I apologized for my email’s tone shortly after sending it. To say I was shocked when the police contacted me in late August would be an understatement. The last communication I had with Dr. Warren was two months earlier, when she responded to my apology: “Thank you for your apology. All will be well.”

In response to publicity surrounding my preprint and Dr. Stewart’s articles, the American Society of Criminology held a Forum on Scientific Integrity at its annual meeting in November 2019. At the forum, despite having coauthored two of the articles, having emailed editors, and having contacted the police about my email, Dr. Warren told the ASC executives, “I am not necessarily attached to the incident.” Her question for them was about how they planned to deal with it when “there is a public war…attached to what’s going on.”

---

41. Email from Dr. Baumer to me and six other coauthors, June 10, 2019.
42. Email from Dr. Warren to me and six other coauthors, June 10, 2019.
43. Email from Dr. Linders to Dr. Wright, August 6, 2019.
44. Email from Dr. Warren to me, June 12, 2019.
45. Dr. Warren’s comments start at 39:53 in this video.
What happened at the journals

The Committee on Public Ethics (COPE) says that journal editors should investigate when “a published article is criticised via direct email,” regardless of whether the sender is anonymous, and emphasizes that “it is important not to try to ‘out’ people who wish to be anonymous” (link). An analysis of emails obtained under the Freedom of Information Act reveals that all of the editors were alerted in May about the data irregularities in all five of Dr. Stewart’s articles, either by an anonymous sender—“John Smith”—or by another editor.

Law & Society Review, regarding Mears et al. 2019

After receiving the anonymous email, Dr. Susan Sterett, the journal’s editor, shared it with the authors, writing: “it seems to imply pretty egregious misconduct—points 4 and 8 especially.”46 Then she emailed the editors of the other journals and tried to get them all on the same page. She wrote: “I would like to have a coordinated response, including possibly ignoring the email” (my emphasis).47 She also explained that she tried to discover the source’s identity: “I asked ‘John Smith’ to give me more information about himself and he would not.”48

In July, after I posted my preprint, Dr. Sterett contacted the other editors again to reiterate her position, “I am not interested in asking for a response from the authors to an anonymous email. However, to my mind it’s worth knowing that the issue isn’t going away.”49 But she also explained that if any of the other editors ever decided to do anything, she wanted to be included: “I’d appreciate knowing, and I’d appreciate doing something together.”50 In late August, she contacted the other editors again, and forwarded them a discussion by Dr. Jeremy Freese of the mathematical impossibilities in the articles. In the same email, Dr. Sterett noted that she had received Dr. Gertz’s letter of support, and once more reiterated her stance on the data irregularities: “I want to treat the issue as closed unless someone wants to question the survey in detail.”51 Months after she closed the issue, the article was retracted at the authors’ request.

46. Email from Dr. Sterett to Dr. Mears, May 29, 2019.
47. Email from Dr. Sterett to Drs. Linders and Wright, May 31, 2019.
48. Email from Dr. Sterett to Drs. Linders and Wright, May 31, 2019.
49. Email from Dr. Sterett to Drs. McDowall, Linders, and Wright, July 15, 2019.
50. Email from Dr. Sterett to Drs. McDowall, Linders, and Wright, July 15, 2019.
51. Email from Dr. Sterett to Drs. McDowall, Linders, and Wright, August 26, 2019.
**Social Problems**, regarding Stewart et al. 2015 and Stewart, Johnson et al. 2019

Drs. Annulla Linders and Earl Wright, the journal’s co-editors, received an email in May from Dr. Stewart listing some of the accusations and irregularities, and, in relation to Dr. Brown, Dr. Heathers, and Mr. Smith, asserting that “data thugs…demand data and if they do not receive it, they contact editors and universities and threaten to write blogs and tweets about the errors uncovered.”

Drs. Linders and Wright also received emails in May and July from Dr. Sterett about Dr. Stewart’s articles. The May email included a full list of the irregularities in all five articles. The co-editors did not investigate.

Two weeks after they got Dr. Sterett’s second email, and two months after they received Dr. Stewart’s email, Drs. Linders and Wright received an email from a reporter, Thomas Bartlett, asking if they were looking into the irregularities in Stewart et al. (2019). They replied, “no question concerning this paper has been brought to our attention.” Before replying to the reporter, however, Dr. Wright wrote to Dr. Linders: “a writer from the Chronicle of Higher Education is sniffing around. Is the paper he cites below the one inquired into by the ‘data thugs?’ Of course, I won’t respond until we get a plan together.”

When I found out what Drs. Linders and Wright had told the reporter, I emailed them my preprint. It turns out they already had it. “Earl just stumbled on the (damaging) information below,” Dr. Linders wrote about the preprint when they first found it online. Dr. Wright had circulated the preprint before I emailed them; he noted that “one of the author’s ‘outted’ is Stewart,” to which Dr. Linders responded: “Ok, so more damage.” Still, it apparently took a direct email from a non-anonymous source to get them to investigate. “So now we have a formal complaint to justify an investigation,” Dr. Linders wrote after receiving my email.

**Criminology** (regarding Johnson et al. 2011 and Stewart et al. 2018)

*Criminology*’s co-editors—Drs. Brian Johnson, Janet Lauritsen, David McDowall, and Jody Miller—adopted a comment-and-reply response to the anony-

---

52. Email from Dr. Stewart to Drs. Linders and Wright, May 30, 2019.
53. Email from Dr. Wright to Mr. Bartlett, July 31, 2019.
54. Email from Dr. Wright to Dr. Linders, July 31, 2019.
55. Email from Dr. Linders to Dr. Wright, August 2, 2019.
56. Email from Dr. Wright to Dr. Linders, August 2, 2019.
57. Email from Dr. Linders to Dr. Wright, August 2, 2019.
58. Email from Dr. Linders to Dr. Sterett, August 5, 2019.
mous email. They “invited ‘John Smith’ to submit a comment or comments about the articles in question” (Johnson et al. 2019; Stewart et al. 2018), but demanded he reveal his identity before they would move forward.⁵⁹ One co-editor, Dr. Johnson, kept Dr. Stewart in loop. When the co-editors first learned of the data irregularities, Dr. Johnson told Dr. Stewart: “I’m not sure it will amount to anything but thought you’d want to know. If anything comes down the pipeline I’ll keep you informed.”⁶⁰ When the co-editors eventually offered ‘John Smith’ the comment-and-reply opportunity, Dr. Johnson let Dr. Stewart know, explaining that:

One of the provisions of the invitation is that the critique could not be anonymous and it would also be subject to external reviews. I think if we end up submitting a correction first that would make the comment pointless, and I think David et al. just felt like they had to provide some type of formal response to the email.⁶¹

Later, a professor from another discipline, Dr. Walter Schumm, who also had analyzed the irregularities in Dr. Stewart’s articles, did send Criminology a non-anonymous critique, but the co-editors denied the professor the comment-and-reply opportunity. Dr. McDowall, the lead editor, replied to Dr. Schumm, writing: “I do not think additional commentary will be useful at this point, but I appreciate your offer to contribute.”⁶²

In June, I emailed the co-editors and asked them to retract Johnson et al. (2011). I sent them evidence that neither the findings nor sample reported in the article existed, and I told them that Drs. Stewart and Johnson were refusing to share data or even output with me. The co-editors replied saying they were going to give my coauthors a few months to work through their reanalysis. During that time period, my coauthors were free to withhold data, output, and even basic information from me. That is so unbelievable it bears repeating: for several months, the co-editors let Drs. Stewart and Johnson refuse to share data and output with a coauthor. About the evidence I sent, Dr. McDowall told the Chronicle of Higher Education that he “didn’t read it in great depth,” and that he thought it was “pretty hostile for Justin to start making these claims” (Bartlett 2019).

After Stewart, Mears et al.’s (2019) corrigendum was published, accusations quickly surfaced that it was mathematically impossible. Mr. Bartlett, the Chronicle reporter, asked Dr. McDowall about those accusations, and Dr. McDowall responded by outlining his negative views of the accusers:

---

⁵⁹. See email from Dr. Johnson to Dr. Stewart, June 20, 2019.
⁶⁰. Email from Dr. Johnson to Dr. Stewart, May 29, 2019.
⁶¹. Email from Dr. Johnson to Dr. Stewart, June 20, 2019. Dr. Johnson was referring to his co-editors.
⁶². Email from Dr. McDowall to Dr. Schumm, September 27, 2019.
If regular circumstances prevailed, I imagine that the corrigendum would have passed without notice after it appeared. Given the current situation, I was surprised that it took five days before the trolls on www.socjobrumors discovered it and began savaging it. I also know, of course, that a Stanford sociologist with a perhaps unjustifiably high sense of self-esteem has tweeted disparagingly about the correction. From my point of view, some of the socjobrumors postings offer better and more thoughtful criticisms than did the high self-esteem Stanford sociologist.  

Dr. McDowall also discussed the timeline of the corrigendum and explained why it was not the journal’s responsibility to ensure its accuracy:

Eric Stewart asked sometime around February if he could submit a correction to his 2018 article. This was well before the appearance of “John Smith” or Justin Pickett. It may have been after the “data thugs” contacted him, but he did not mention that as a motivation … The document does not address any specific criticism that the journal has or will publish, and it is not itself an original peer reviewed contribution. It is simply an author’s attempted correction to a set of results, and it is unnecessary and out of place for me to offer a defense of it.

But there is more to the story than that. The Stewart, Mears et al. (2019) corrigendum was published in mid-August. The anonymous email Dr. McDowall and his coeditors received three months before, in May, did more than just list the irregularities in the original Stewart et al. (2018) article. It also explained that Dr. Stewart and his coauthors had sent a correction for the same 2013 sample to another journal, after receiving outside criticism in February; that the correction appeared to cover up the original irregularities, rather than explain them; and that the correction had many new irregularities. The May email listed those irregularities, which included the same mathematical impossibilities that Dr. Freese (the Stanford sociologist) and others later pointed out in the Criminology corrigendum. The co-editors either did not read the anonymous email or ignored its content.

Dr. McDowall also expressed to Mr. Bartlett disapproval of the criticism directed at Dr. Stewart and provided a potential explanation for the irregularities. Even though the corrigendum listed a single coding error and reported using the same 2013 data as the original article, Dr. McDowall wrote:

I will nevertheless suggest the outlines of a defense, since I think that Stewart

63. Email from Dr. McDowall to Mr. Bartlett, August 27, 2019.
64. Email from Dr. McDowall to Mr. Bartlett, August 27, 2019.
has been treated unfairly about it … the descriptive results are not possible if the original and corrected versions used exactly the same data … I have not worked through whether variations in missing data patterns could in fact account for the different summary statistics. It seems to be a reasonable possibility, however, and I offer the matter to the trolls and Stanford University professors … in fairness please note that Stewart does not represent the data as being identical in both samples. Again, this would not be a substantial issue absent the lynch mob atmosphere that was only beginning to emerge as he completed the correction document. 65

Mr. Bartlett asked whether the co-editors would request Dr. Stewart’s data. Dr. McDowall said they would not. He explained, “We will not be rushed into one-sided decisions to satisfy the demands of internet bullies or Stanford University professors, no matter how high their apparent self-esteem.” 66 Both of the Criminology articles were eventually retracted, at the authors’ request. Afterward, the co-editors published a statement claiming that science was coming “under growing attacks from…those who are trying to establish themselves as self-appointed guardians (and often entrepreneurs) of science” (McDowall et al. 2020).

Conclusion and recommendations

The articles

Scientific fraud occurs all too frequently—approximately 1 in 50 scientists admit to fabricating or falsifying data (Fanelli 2009)—and I believe it is the most likely explanation for the data irregularities in the five retracted articles. Dr. Stewart’s current claim about the source of the 2013 survey differs from his previous claim and from what the survey firm’s owner and director have said. His claim about the number of 2008 samples also differs from the director’s account. When asked for the 2008 data, Dr. Stewart claimed he destroyed the original file, even though FSU officials said not to change it. More generally, many aspects of the data and findings are impossible, and others are so implausible or improbable as to be preposterous.

Knowing whether the retracted articles are fraudulent is important because Dr. Stewart has several other articles with irregularities (e.g., Mears et al. 2013; Mears et al. 2017; Stewart 2003; Stewart and Simons 2010; Stewart et al. 2006;
Stewart et al. 2009). The retraction notices say honest error, not fraud, is the explanation. Fortunately, if that is true, Dr. Stewart could easily prove it: recreate the original sample \(N = 1,184\) that produces the findings in Johnson et al. (2011) and then publicly explain how he did it. Dr. Stewart claims he got from \(N = 500\) to \(N = 1,184\) through accidental duplication, but then dropped the duplicates when I asked for the data. Because dropping duplicates is reversible, Dr. Stewart should be able to duplicate his way back to the same sample again, if he is telling the truth. I have made the sample of 500 respondents available publicly. All that is needed now is to know which of the 500 respondents to duplicate, and how many times, to recreate the original sample \(N = 1,184\) that produces the findings in our published article. Dr. Stewart could also post code showing how he clustered the 326 counties in the data I released \(N = 500\) down to 91 county units.

**Coauthors and data**

How did a group of competent researchers end up publishing five unsound, unsalvageable articles? Monopolization of the data seems to be part of the answer. “I never worked with (or even saw) any version of the data,” Dr. Johnson wrote about Johnson et al. (2011).67 “Dr. Stewart had conducted the analyses and created the tables for all five papers,” Dr. Mears explained.68 To my knowledge, none of Dr. Stewart’s coauthors ever analyzed, or even laid eyes on, the full data for any of the five articles, including those they first-authored. As a consequence, they took a passive role in validating their articles, even while they took an active role in defending them.

Most how-to-improve-science lists include open data policies. Having more eyes on the data reduces the survival rate for honest errors. Having more eyes on the data reduces the opportunity for fabrication or falsification. But many authors are reluctant to share data publicly, and sometimes there are legitimate privacy concerns or externally imposed restrictions. Sharing data with coauthors, however, should be uncontroversial and feasible. Yet without institutional support, coauthors may feel uncomfortable requesting data. For example, once irregularities were identified in their articles, Dr. Stewart’s coauthors were reluctant to press him for the data, probably because of concerns related to friendship and loyalty. (There certainly is no scientific justification for refraining from requesting data.) Therefore, one recommended reform is that, short of an open data policy, journals should at least require authors submitting articles to confirm that all of their coauthors have a copy of the data.

---

67. Email from Dr. Johnson to me, June 6, 2019.  
68. Email from Dr. Mears to Dr. Sterett, May 30, 2019.
Editors and COPE

None of the editors followed COPE’s guidelines when alerted to the irregularities in Dr. Stewart’s articles. One editor seemingly tried to coordinate a collective response of ignoring the allegations, even though she recognized their potential seriousness. At two journals, the editors sought the whistleblower’s identity. I believe that it is possible that one or more of the editors would have revealed the whistleblower’s identity had they discovered it. For instance, email correspondence reveals that Dr. Johnson kept Dr. Stewart up to date on what his co-editors knew and were doing, even after officially recusing himself from the matter. Not a single editor started an investigation in response to the anonymous allegations. Dr. Linders explained her reluctance to take those allegations seriously:

At this point, especially since the person complaining would not come forward, I assumed this was something along the lines of the scientific version of complaints about ‘fake news’ (now ‘fake science’). If you cannot verify the credibility of the source, how can you trust the information?  

It appears two journals’ editors ignored COPE guidelines because they were unfamiliar with them. Once they learned about the guidelines from a publisher’s representative, they seemed committed to following them. For example, Dr. Sterett wrote to Dr. Linders: “I think the [COPE] flow chart I sent separately is far and away the most valuable document, except for the point that journals need policies.” One recommendation, then, is to require all editors to review COPE guidelines before taking on editorial responsibilities, and to follow them if they receive allegations about an article published in their journal.

At Criminology, what seems to have driven how the co-editors responded was sympathy for some of the authors and a low opinion of critics. Connections between the co-editors and authors are likely to blame; Dr. Johnson, the lead author of one article, was a co-editor, and Dr. Stewart, the lead author of the other, was to become a co-editor. The obvious recommendation is to avoid such conflicts of interest. When authors of questioned articles have relationships (professional or personal) with editors, journals should use independent investigators to investigate scientific irregularities.

Before closing, let me emphasize that many journals do require authors to post their data publicly, and some, like the American Journal of Political Science, go so far as to replicate reported results before publishing articles (link). But even at journals that do not, editors can still request authors’ data if concerns about

69. Email from Dr. Linders to Mr. Blong and four others, August 1, 2019.
70. Email from Dr. Sterett to Dr. Linders, September 27, 2019.
findings emerge (see Miyakawa 2020). The Stewart scandal took place over five months, and it required considerable time and effort from the editors involved. The editors corresponded extensively with each other and with other parties, including the authors, reporters, officials of their academic societies, publisher representatives, and university administrators. And the *Criminology* co-editors wrote multiple public statements about the steps they were taking to address the problems. Why did they not simply ask Dr. Stewart for his data? It would have saved a lot of time. Is there any good reason for editors not to verify data when someone, much less a coauthor, provides credible evidence of potential fraud?

The titles and abstracts of the five retracted articles

Here are the titles and abstracts of the five articles, as well as the Google Scholar citation counts as of February 27, 2019:

“Ethnic Threat and Social Control: Examining Public Support for Judicial Use of Ethnicity in Punishment” (Johnson et al. 2011, *Criminology*, citations: 80)

Research on social inequality in punishment has focused for a long time on the complex relationship among race, ethnicity, and criminal sentencing, with a particular interest in the theoretical importance that group threat plays in the exercise of social control in society. Prior research typically relies on aggregate measures of group threat and focuses on racial rather than on ethnic group composition. The current study uses data from a nationally representative sample of U.S. residents to investigate the influence of more proximate and diverse measures of ethnic group threat, examining public support for the judicial use of ethnic considerations in sentencing. Findings indicate that both aggregate and perceptual measures of threat influence popular support for ethnic disparity in punishment and that individual perceptions of criminal and economic threat are particularly important. Moreover, we find that perceived threat is conditioned by aggregate group threat contexts. Findings are discussed in relation to the growing Hispanic population in the rapidly changing demographic structure of U.S. society.


This article examines the legacy of lynchings on contemporary whites’ views of blacks as criminal threats. To this end, it draws on prior literature on racial animus to demonstrate the sustained influence of lynching on contemporary America. We
hypothesize that one long-standing legacy of lynchings is its influence in shaping views about blacks as criminals and, in particular, as a group that poses a criminal threat to whites. In addition, we hypothesize that this effect will be greater among whites who live in areas in America where socioeconomic disadvantage and political conservatism are greater. Results of multilevel analyses of lynchings and survey data on whites’ views toward blacks support the hypotheses. In turn, they underscore the salience of understanding historical forces, including the legacy of lynchings that influence contemporary views of blacks, criminals, and punishment policies.

“The Social Context of Latino Threat and Punitive Latino Sentiment”
(Stewart et al. 2015, Social Problems, citations: 46)
Prior research on the racial threat perspective and social control typically relies on aggregate-level demographic measures and focuses on racial, rather than on Latino group, composition. This predominant focus in research on racial threat and social control makes it unclear whether the assumed linkages are confined to one subordinate group or whether other groups, such as Latinos, are viewed as threatening and elicit heightened social control reactions as well. In the current study, we use data from the Punitive Attitudes Toward Hispanic (PATH) Study, a national sample of U.S. residents to investigate the influence of macro- and micro-level measures of Latino group threat on punitive Latino sentiment. More specifically, we use multilevel models to detect direct and interactive relationships between Latino presence and perceived Latino threat on punitive Latino sentiment. The findings show that Latino population growth and perceived Latino criminal and economic threat significantly predict punitive Latino sentiment. Additionally, multiplicative models suggest that the effect of perceived criminal threat on punitive Latino sentiment is most pronounced in settings that have experienced recent growth in the size of the Latino population.

“Lynchings, Racial Threat, and Whites’ Punitive Views Toward Blacks”
(Stewart et al. 2018, Criminology, citations: 8)
Disparities in historical and contemporary punishment of Blacks have been well documented. Racial threat has been proffered as a theoretical explanation for this phenomenon. In an effort to understand the factors that influence punishment and racial divides in America, we draw on racial threat theory and prior scholarship to test three hypotheses. First, Black punitive sentiment among Whites will be greater among those who reside in areas where lynching was more common. Second, heightened Black punitive sentiment among Whites in areas with more pronounced legacies of lynching will be partially mediated by Whites’ perceptions of Blacks’ criminality and of Black-on-White violence in these areas. Third, the
impact of lynching on Black punitive sentiment will be amplified by Whites’ perceptions of Blacks as criminals and as threatening more generally. We find partial support for these hypotheses. The results indicate that lynchings are associated with punitive sentiment toward Black offenders, and these relationships are partially mediated by perceptions of Blacks as criminals and as threats to Whites. In addition, the effects of lynchings on Black punitiveness are amplified among White respondents who view Blacks as a threat to Whites. These results highlight the salience of historical context for understanding contemporary views about punishment.

“The Social Context of Criminal Threat, Victim Race, and Punitive Black and Latino Sentiment”

(Stewart, Johnson et al. 2019, Social Problems, citations: 2)

A well-established body of research focuses on the relationship between criminal threat and the exercise of formal social control, and a largely separate literature examines the effects of victim race in criminal punishment. Despite their close association, few attempts have been made to integrate these related lines of empirical inquiry in the sociology of punishment. In this article, we address this issue by examining relationships among criminal threat, victim race, and punitive sentiment toward black and Latino defendants. We analyze nationally representative survey data that include both subjective and objective measures of criminal threat, and we incorporate unique information on victim/offender dyads to test research questions about the role victim race plays in the formation of anti-black and anti-Latino sentiment in the criminal justice system. The results indicate that both subjective perceptions of criminal threat and minority population growth are significantly related to punitiveness among whites, and that punitive sentiment is enhanced in situations that involve minority offenders and white victims. Moreover, we show that aggregate indicators of racial threat strongly condition the effect of victim race on punitive attitudes. Implications of these findings are discussed in relation to racial group threat theories and current perspectives on the exercise of state-sponsored social control.

Data, code, and documentation

All data, code, and documentation related to this research is available from the journal website (link).
References


Johnson, Brian D., Janet Lauritsen, David McDowall, and Jody Miller. 2019. Statement from the Co-Editors of *Criminology*. September 28. American Society of Criminology (Columbus, Ohio). Link


McDowall, David, Charis Kubrin, and Jody Miller. 2020. Editor’s Corner. The Criminologist (American Society of Criminology) 45(1, Jan.–Feb.): 9. Link


Justin T. Pickett is an associate professor in the School of Criminal Justice at the State University of New York at Albany. He received his Ph.D. in criminology in 2011 from the College of Criminology and Criminal Justice at Florida State University, where he was Eric Stewart’s teaching assistant. He is the 2015 recipient of the American Society of Criminology’s Ruth Shonle Cavan Young Scholar Award. His research interests include public opinion, survey research methods, theories of punishment, and police-community relations. His email address is jpickett@albany.edu.


About the Author

Justin T. Pickett is an associate professor in the School of Criminal Justice at the State University of New York at Albany. He received his Ph.D. in criminology in 2011 from the College of Criminology and Criminal Justice at Florida State University, where he was Eric Stewart’s teaching assistant. He is the 2015 recipient of the American Society of Criminology’s Ruth Shonle Cavan Young Scholars Award. His research interests include public opinion, survey research methods, theories of punishment, and police-community relations. His email address is jpickett@albany.edu.

Go to archive of Character Issues section
Go to March 2020 issue

Discuss this article at Journaltalk: https://journaltalk.net/articles/6005/
Captive of One’s Own Theory: Joan Robinson and Maoist China

Evan W. Osborne

HOSTILE OBSERVERS (INCLUDING MANY PROFESSIONAL CHINA WATCHERS) LIKE TO DISCRIMINATE THE REPORTS OF VISITORS WHO, THEY MAINTAIN, MUST’VE BEEN SHOWN AROUND. THERE IS A GREAT DEEP OF RELUCTANCE, BOTH IN THE SOVIET UNION AND IN THE WEST, TO BELIEVE WHAT SYMPATHETIC VISITORS TO CHINA REPORT. IS IT POSSIBLE TO CARRY OUT INDUSTRIALIZATION WITHOUT SQUEEZING AND DRAGOONING THE PEASANTRY? HOW CAN THERE BE DISCIPLINE IN A FACTORY WHERE THE WORKERS ARE FREE TO CRITICIZE THE BOSS? HOW CAN THERE BE INCENTIVES TO WORK WITHOUT INEQUALITY? HOW CAN THERE BE SOCIALISM WITH GRASS-ROOTS DEMOCRACY? HOW CAN A BACKWARD COUNTRY DEVELOP BY ITS OWN EFFORTS, WITHOUT BENEFIT OF FOREIGN AID AND FOREIGN ADVISERS?

Yet there are certain large facts which no one can deny. Frequent predictions of breakdown, famine, and chaos have proved false.

—Joan Robinson (1970b, 9)

Joan Robinson (1903–1983) began studying economics at Cambridge in 1922, and after graduating she would go on to write numerous major works in microeconomics, macroeconomics, the history of economic thought, and development economics. But in the last three decades of her life, she wrote very favorably of both the economies and the broader societies under two of the more ghastly regimes of the 20th century—Kim Il Sung’s North Korea and especially Mao Zedong’s China. Given that much information was available to anyone proposing to evaluate the economics and ethics of the Chinese regime, that she wrote what she did seems difficult to understand.

Robinson’s long, distinguished record has prompted many evaluations of her work. Most evaluations pay little or no attention to her China writings. When they
do, her complete and completely confident mischaracterization of what happened there is said to have been only temporary, and something she repudiated late in her life. But this paper argues that Robinson’s extreme views on China, even during its darkest post-1949 moments, were not a late-career descent into eccentricity, but followed directly from how she saw the problem of economic development, and the comprehensive superiority of Chinese communism in comparison to both liberalism and to Soviet rule since Stalin, and they betrayed an irresponsible gullibility about what she was told on her visits to the country. This paper uses Robinson’s own economic thinking to explain how she came to hold these views, and why she was so reluctant to disavow them. It also documents the failure of scholars on Robinson to grapple with this legacy, which was often expressed with complete certainty rather than scholarly caution.

Joan Robinson before her interest in China

Robinson (née Maurice) enrolled at Cambridge in 1922, and in 1925 passed the economics tripos. In the same year she married Austin Robinson, who himself became an accomplished economist and who later served as an associate editor of the *Economic Journal* while John Maynard Keynes was the editor. After that, she followed her husband while he worked in India; after independence, she would go there again and discuss economic planning with scholars there. Upon returning to England in 1928, she sought a job at Cambridge, but initially the only offer was to her husband, in 1929. In 1934 she was finally appointed an assistant lecturer after the publication of her first major work, *The Economics of Imperfect Competition* (see Harcourt and Kerr 2009, 4).

The period between the World War I armistice and 1930 in the UK had been very different from the ‘roaring’ 1920s in the United States. In the UK unemployment soared after the surviving soldiers returned home, and both monetary and fiscal policy were subsequently unaccommodating, as were trade unions who resisted wage cuts. The government also tried to maintain an unrealistically high value of sterling. Unemployment gradually drifted lower over the course of the 1920s after the initial huge postwar spike, but it was still over 7 percent by the time of the 1929 worldwide market crash.

Although macroeconomics was not yet a universally acknowledged sub-discipline, the British economists of the day could not help but think of the sustained postwar stagnation as something that traditional liberal economics could not explain. Especially after 1929, the standard liberal recipe of waiting for self-correction was insufficient if not counterproductive for addressing the crisis that faced first Britain and then much of the world.
Cambridge during the 1920s was a leader in this nascent discipline of macroeconomics, and its scholarship was more and more marked by skepticism of free-market processes. Robinson’s views fit in this environment well. Shortly after the publication in 1930 of the first version of Keynes’s *A Treatise on Money*, which Keynes had become somewhat dissatisfied with, Robinson and several other Cambridge faculty formed a group subsequently known as the ‘Cambridge Circus’ to discuss the claims in the work (Aslanbeigui and Oakes 2002). Richard Kahn, James Meade, and Piero Sraffa, in addition to Joan and Austin Robinson, discussed new ways of thinking about such problems as stagnation and large economic fluctuations, both of which had been known during the Industrial Revolution but which were now front and center. These discussions in 1930 and 1931 had some influence on Keynes’s *General Theory* (1936). Robinson would spend the rest of her career at Cambridge and become one of the department’s most influential thinkers.

**How Joan Robinson saw the world**

Robinson’s scholarly output fills seven volumes (Robinson 2001), and she wrote hundreds of essays designed for the broader public in venues from major periodicals to modest Cambridge student journals. When it comes to how she thought about China, several themes evolve over the years. They illuminate why she admired the Chinese economic and social model, even as it played out ever more disastrously during the time she was writing about it, as evidenced both by much newer scholarship and by the substantial deconstruction of the model beginning a few years after Mao’s death.

Throughout Robinson’s career her focus was usually on society or the economy as a whole. This is so despite *The Economics of Imperfect Competition* (Robinson 1969/1933) being what we would now call ‘microeconomic,’ and indeed despite its having much influence on how modern microeconomic textbooks are arranged. Subsequent to that work her focus even with respect to the question of monopoly and competition was how microeconomic structure affected the performance of the economy as a whole.

Robinson’s background led her to believe that the way to think about the economy was as a relation, generally a conflict, among different economic classes (see, e.g., Robinson 1947/1937). The key actors were not individuals but the collectives of labor, capitalists, and so on. In her framework there was little need to specify different firms, types of workers, or individual industries. Simple aggregates said to describe the interactions of millions of people—the unemployment rate, inflation rate and so on—were the variables of interest. ‘Labor’ suffers from high unemployment or not, ‘the economy’ grows or not, etc. Other than the very broad
categories of agriculture and industry, no attention was paid to sectoral shifts or
to individual microeconomic relations. This of course is not unusual in macro-
economic scholarship, but it would influence how she thought about China’s
economic challenges.

The idea of universal economic law applicable to all places in all times was
something Robinson for a time rejected. Today we might say that in general she
believed every society at any point in time was path-dependent. The choices made
in the past determined the choices available in the present. Each society in its
particular situation had to choose its own policy regime. Later, however, she would
nonetheless assert that the China model in particular was actually widely general-
izable (Robinson 1973; 1970d).

Thus, she was a profound critic of liberal economics. While her work on
Karl Marx (Robinson 1966/1942) is seen as having made key links between Marx’s
Capital and the economics of the era in which she wrote (see Gram and Walsh
1983), Robinson always had a complicated relationship with Marxism. A goal of
Robinson (1966/1942) was to deal with Marx’s problematic assumption of the
labor theory of value while preserving what was valuable about Capital as a whole.
Her views on political economy as a parade of progress among giants, Marx
included, is found in a work produced during her time in India called “Marx,
Marshall and Keynes” (Robinson 1978/1955). But Robinson always abjured the
label of ‘Marxist’ for herself, preferring to think of herself as a left-wing Keynesian.
She had frequent debates with neoclassical economists, notably Paul Samuelson.
Robinson (1966/1942) pronounced The General Theory and what followed to be a
more fruitful critique of capitalism than Marxism. Robinson criticized capitalism as
having a host of flaws, including the creation and toleration of unacceptable levels
of poverty, excessive consumption of luxuries by the rich—a particular problem
for developing countries—and perverse incentives. As an example of perverse
incentives, she cited the World War II practice of Brazilian companies destroying
output or capital stock to prevent prices of output from falling (Robinson 1943, 6).
She credited capitalism with being potentially self-regulating in the short run and
with creating no end of technological advances, but these features hardly redeemed
it. She wrote so enthusiastically about post-1949 China in part because of what she
saw as capitalism’s ethical deficiencies.

Robinson was a persistent critic of Samuelsonian ‘neoclassical’ economics.
Its reliance on ‘equilibrium’ as the core of economic analysis struck her as mis-
guided. By the time the second edition of The Economics of Imperfect Competition was
published she regarded the idea of markets moving toward their neoclassical
equilibrium, which had been part of the analysis of the first edition of the book,
to be a “shameless fudge” (Robinson 1969/1933, vi). There was no reason to
think that the price and output adjustments that might occur in a liberal economy
should actually approximate that equilibrium. There would frequently be, she and other followers of Keynes thought, effective demand failures. She ridiculed macroeconomic equilibrium with characteristic mordancy, writing that “it is necessary to recognise that the classical doctrine does not exclude starvation from the mechanism by which equilibrium tends to be established” (Robinson 1946, 102). This failure to have a tendency to move toward such a fictional equilibrium indicated that a major economic problem was that the economy as a whole was subject to systematic dysfunction. Whereas much liberal thought had recommended government inaction in the face of macroeconomic ‘disequilibrium,’ believing that eventually (in the case of sustained high unemployment) wages and employment would adjust, Robinson believed that dysfunction was the norm. In another classic work and one devoted primarily to development economics, *The Accumulation of Capital*, she described the conditions necessary in liberal economics for a macroeconomic “golden age” to be sustained:

When technical progress is neutral, and proceeding steadily, without any change in the time pattern of production, population growing…at a steady rate and accumulation going on fast enough to supply productive capacity for all available labour, the rate of profit tends to be constant and the level of real wages to rise with output per man … no internal contradictions in the system … if entrepreneurs have faith in the future and desire to accumulate at the same proportional rate as they have been doing in the past, there is no impediment to prevent them [and] the system develops smoothly [with output and the stock of capital (valued in terms of commodities) growing at a rate compounded of the rate of increase in the labour force and the rate of increase in output per worker. (Robinson 1969/1956, 99)

But this golden age was a theoretical curiosity only, a “mythical state of affairs not likely to obtain in any actual economy” (ibid.) The actual state of affairs could easily be a tragic one of sustained high unemployment. Indeed, a recurring theme in her analysis was that the output market could be in equilibrium even as there was substantial unemployment and yet firms saw no need to add capital. This idea had been a key element in Keynes’s *General Theory*. It should be added that in contrast to postwar American Keynesians, she did not simplistically believe that these sorts of economic fluctuations could easily be managed by proper fiscal and monetary policy, but that in liberal economies they were inevitable and substantial.

Economic development depends critically on the generation and proper allocation of the ‘surplus’ first posited in classical economics, and subsequently depicted as capitalist exploitation by Marx. Robinson believed that the critical matter in modernization and economic development was capital accumulation. Robinson (1943) had given some credit to capitalism for having discovered modern
technology and industry, but she contended that the task for developing countries not yet shackled by capitalism was to make sure that agriculture could create enough surplus value to enable the state-directed building of modern industry while also distributing it fairly. This was still her thinking in the book on economic development she published late in her life (Robinson 1981/1979). In her analysis of development she averred that capitalism had generated the surplus and used it to industrialize, but only by making many mistakes and creating injustice along the way. After decolonization she hoped that the new developing economies could do things better, by taking advantage of technological and organizational discoveries made earlier in capitalist societies but then restructuring their economies in such a way that they could generate the surplus without the injustice.

An implication of Robinson’s way of seeing the world was that whether accumulation would happen under capitalism was not certain. Postulated growth paths were just that, with no evidence guaranteeing that actual accumulation and growth would occur in the hypothesized manner. Social institutions, long-term failures of effective demand, and other factors could prevent development, rendering it a perilous process. While writing about China she became more and more convinced that its system under Mao for deciding on the allocation of resources was superior to the happenstance and distorted sort of allocation that happened under other existing economic systems.

Critically, given what happened in China as she was writing about it, Robinson thought that economic policy had ethical dimensions beyond the merely ‘economic,’ and it was important to be explicit about them. While one could make predictions about the effects of various economic policies and arrangements, not to scrutinize the deeper ethics beyond efficiency of the outcomes yielded was unjustifiable. This was a lifelong theme of her work, and it was central in her Economic Philosophy (Robinson 1962a). Even Freedom and Necessity (Robinson 1970c), a substantial work in endogenous economic history, emphasized ethical thinking (to which economic history sometimes gives short shrift). In particular, there was a way to engineer policy to make poor countries wealthier. Doing so was an urgent task for economists, and this urgency motivated her to explore the question of development. That poverty could persist amid so much potential wealth was shameful, the more so because wealthier nations were now plowing so much of their own potentially investable surplus into military expenditure.

### The China that Joan Robinson wrote about

only invited foreign guests whom it thought could strengthen its image at home and abroad (Lovell 2015). She did not speak or read the language, but in addition to being familiar with Marxist economics, she made it a point to follow official Chinese accounts of deliberations on economic plans and pay attention to the Chinese government’s descriptions of who the contemporaneously perceived enemies of those plans were, both inside and outside the country. During the 34 years from the CCP coming to power in 1949 and her death in 1983, there were numerous events that were or could have been known to her, although contemporaneous accounts of them both inside and outside China did not reflect with certainty how awful we know now them to have been. Major events are discussed in the following subsections.

1947–1951: Land reform and mass murder of ‘landlords’

Begun even before the Communist Party of China (CCP) came to power nationwide in 1949, and based on the precedent set in areas captured by it in the 1920s and 1930s, it consisted of CCP agents taking private landholdings, from very small to quite large, from those asserted by these agents to be landlords. The land was then given to those deemed peasants. The post-1949 confiscation process involved bitter denunciations and severe physical abuse of ‘enemies’ of the peasants. Ultimately, perhaps two million ‘landlords’ were killed nationwide (Dikötter 2013, xii; Dillon 2010, 290). By the completion of the process, roughly 40 percent of agricultural land had been transferred to peasants. Robinson clearly knew about the campaign since she wrote about it frequently and admiringly in a scholarly sense. Robinson (1966) minimized the violence, and in later years she compared it favorably to processes in India, Egypt, and Latin America (Robinson 1981/1979, 52–54).

1949–1955: Nationalization

The CCP during this time took over all substantial private property. Buildings, especially factories, belonging before 1949 either to the Nationalist government (which had sometimes seized them after 1945 from the Japanese) and foreigners were simply taken. Sometimes in the latter case foreigners were held hostage until the process was completed (Thompson 1979). CCP cadres were inserted early on into the staff of both Chinese- and foreign-owned factories, and

---

2. Robinson (1966) writes: “Some landlords were executed, but the proportion seems to have been very small—only those convicted of an exceptional number of murders. The great majority were absorbed into the labor force.” Dikötter (2013) presents voluminous evidence to the contrary.
the freedom of their previous owners to manage the business was gradually restricted. As for domestic firms that had been privately owned in 1949, after a few years of tolerance because of the skills their owners had, all were nationalized quickly after Mao proposed it in 1955. That policy, if not necessarily the details, would have been known to Robinson by her second visit.

1953–1957: The first five-year plan

Heavily influenced by earlier such plans from the Soviet Union, it envisioned a process with industrialization at its core. The CCP discussed the planning process extensively both internally and with what they saw as foreign friends, including Robinson from her very first visit, and so she would have known the general pattern, to the extent that CCP officials spoke truthfully to her.

1953–1958: Collectivization of agriculture

Land reform had been meant as a temporary measure, although peasants were not told this. Shortly after its completion, the peasants were sometimes persuaded and sometimes herded one step at a time into ever larger collective farms, culminating in large communes, initially (although many would eventually fall apart) consisting of thousands of farmers and CCP managers. Robinson not only knew but as we will see frequently wrote favorably about the broad outline of this process.

1956–1959: Hundred Flowers Campaign and aftermath

For a short time beginning in 1956, Mao encouraged the Chinese people to participate in frank discussion of China’s problems. Starting in late 1957, many who spoke out were condemned to imprisonment and labor camps. By the mid-1970s, even communist activists in the West were acknowledging at least the quick reversal of intellectual tolerance, if not necessarily accepting the view that it was designed to lure Mao’s opponents out for elimination.3

3. For an example of Western radical acknowledgment of the oppressive nature of the post-Hundred Flowers response, see British and Irish Communist Organisation (1977, 13–15). More mainstream historians are divided on this question. Spence (2012/1991, 514) says that the desire to draw out hidden enemies was “merely” part of the motivation for the campaign. Dillon (2010) and Dikötter (2010) depict the campaign as a backlash against unexpectedly fierce criticism arising from the sudden freeing of expression. Chang and Halliday (2005) say that the campaign was designed to draw critics out.
1958–1962: Great Leap Forward (GLF) and associated famine

Launched at the instigation of Mao in 1958, the GLF was largely designed to accelerate China’s industrialization and its transformation into a true communist economy by relying on moral suasion and the willpower of the peasantry and working class to do what communism required rather than importing costly, advanced technology. During this period there was a catastrophic famine throughout much of the country. Scholarship now generally places the death toll in the low tens of millions. To what extent the famine was caused by weather and to what extent by government policy has been debated by scholars, but whereas Robinson always attributed the death toll (which she consistently minimized) to weather, policy is now thought to have been far more important. Descriptions of scope, causes, and effects are given by Jisheng Yang (2012), Frank Dikötter (2010), and Jasper Becker (1996).


After several years of tension, in 1960 the Soviet Union withdrew its economic, technical and military advisers from China, along with most of its engineering plans and even some physical capital. The decision resulted from an

---

4. Much other empirical research has also been done on causes. Meng, Qian, and Yared (2015) report that the government policy of forcing communes to provide food to the urban areas before feeding themselves must occupy a primary role. Clement (2012) argues that inadequate foodstuff production and flawed distribution were to blame. Bramall (2011) finds that weather had no effect in a cross-sectional analysis. He suggests that differences in the responses of local cadres (i.e., local CCP leaders) were decisive. The toll of the famine as a function of the adherence of local cadres to Mao Zedong thought was also emphasized by Kung and Chen (2011). Houser, Sands, and Xiao (2009) find that weather and government policy both mattered in cross-sectional death rates, but policy mattered more. Li and Yang (2005) report that the majority of the deaths were due to local differences in policies requiring the communes to supply grain to the center and to divert resources from agriculture to industry. They also find that radicalism of local cadres mattered. What unites all this empirical work is that weather, namely drought, was at best a secondary cause.
ideological split that had developed after Nikita Khrushchev’s secret de-Stalinization speech in 1956. The divorce quickly became global news.


The Cultural Revolution was begun by Mao after his political position was weakened substantially after the GLF. Its nominal justification was to purify Chinese communism by eliminating what Robinson (1970a), in conformity with CCP terminology, termed ‘rightists’ (右派)—as in right rather than left. In the end the damage caused by the Cultural Revolution to Chinese cultural heritage and human capital, as well as to general social cohesion, was devastating. The bulk of the damage occurred during the first four years, with roughly two years of chaos as violent groups called Red Guards (红卫兵), acting either independently or as a tool of one communist official or another, destroyed irreplaceable pieces of cultural heritage (temples, books, etc.), followed by two years of substantial military effort to bring this emerging power under control. The total death toll was probably less than in the anti-landlord campaign in the early 1950s, but sometimes the Cultural Revolution resembled incipient civil war. Relying on microdata from around the country, Andrew Walder (2019, 189) estimates the death toll at 1.6 million. Dikötter notes that “many more lives were ruined through endless denunciations, false confessions, struggle meetings and persecution campaigns” (2016, xviii). For individually focused accounts of the cruelty and devastation, see Lian Xi (2018) and Nien Cheng (2010/1987). Robinson wrote in some depth and with great confidence about the Cultural Revolution—which hereafter I shall abbreviate as ‘CulRev.’

1976: Arrest of Gang of Four

Within weeks after Mao’s death in September 1976, the so-called Gang of Four (四人幫) and many of their underlings were arrested, and at their trial in 1981 many of the charges involved stoking the disastrous chaos of the CulRev.

1978: Beginning of economic reform

At this time discussion and proposals for what would later be seen as merely tentative economic reform began to circulate among the CCP leadership, even as

5. In both Chinese terminology and Robinson’s thinking, ‘rightists’ usually referred to those who sympathized with a Soviet approach to development. The campaign beginning in June 1957 that ended the Hundred Flowers initiative was called the Anti-Rightist Campaign (反右运动).
actual unauthorized reform was underway. By the time of Robinson’s last published writings on China, it was clear that the first steps had been taken in China to move in a more economically liberal direction, a process that has continued to this point for over 40 years.

**Joan Robinson’s writings on China**

Robinson wrote short monographs on China and one full-length book (Robinson 1970a). Her views on the Chinese economy and society would also appear frequently either in chapters of edited scholarly books, in reviews in scholarly journals of books written by others on other themes, or in her own books. Many of her works specifically on China appeared in outlets designed for the educated public, including the portion of that public interested in China. G. C. Harcourt and Prue Kerr (2009) have divided her thinking on China into three phases. According to them, before 1963 she had assumed the Chinese surplus-accumulation problem was similar to that in the early Soviet Union. In the second phase, from 1963 until after Mao’s death in 1976, she “took a sharp turn to the left” (Harcourt and Prue 2009, 145), i.e., she endorsed Mao’s policies. In the last phase, she walked back what were, in hindsight, her naïve views. In the treatment that follows, I structure the discussion around Robinson’s visits to China.

**Visits in 1953 and 1957**

According to a comprehensive list of Robinson’s writings on China (Harcourt and Kerr 2009), her first such publication was released in 1954. Published initially by a small publisher in Cambridge known as Student’s Booksellers, it was later included in Robinson’s *Reports from China: 1953–1976* (1977/1953). It was written after her 1953 trip, and set the table for much of what would follow for the next two-plus decades. The trip occurred because she was invited to join 16 British businessmen who were going to attempt to do business with the new government, and she wrote of improvements that were immediately visible with respect to the squalor, vice, and corruption of the China that certainly existed before 1949.

Robinson (1977/1953) particularly credited the CCP—based on what they told her and allowed her to see—with eliminating the longstanding practices of selling children, forced marriage, and difficulty in obtaining divorce. She lauded a much greater provision of social services and the order of and cleanliness on the streets. In fact, significant progress had been made on some of these problems, especially with respect to women’s rights, for several decades before 1949, although the disappearance of both the homeless and drug users from the streets can be
attributed to draconian measures by the new government. During this time, it seemed clear to Robinson that education, and especially literacy, was much better, urban crime was now almost nonexistent, the non-Han Chinese minorities were being effectively integrated, land reform was said to be going well (although some landlords were a little bitter), and prisoners were being reformed through self-criticism. She also believed, correctly and unlike the peasantry who now owned the land, that land redistribution would not last long, being merely a transitory phase toward collectivization, which she would be in favor of throughout her writings on China.

Robinson was invited to come again in 1957, and a big part of this trip was her presentation of three scholarly lectures (Harcourt and Kerr 2009, 144–151). The visit immediately followed a visit to the Soviet Union, where she had engaged in similar conversations with economists there. The first lecture, given on September 4, emphasized her by now firmly held belief that the accumulation of capital via intelligent investment of the agricultural surplus was the key task in economic development, along with the already widely accepted view that Stalin had made mistakes. In the words of Pervez Tahir, “In the process of raising capital per head, Joan Robinson expected China to learn from the Soviet mistakes so as to minimize human costs” (quoted in Harcourt and Kerr 2009, 157). Perhaps the most important of these costs were the unjust burdens placed on the peasantry and working class. She particularly recognized the unfairness of peasants having to work land of different quality if they were paid by time worked, or by output produced. She proposed taxes assessed collectively on more geographically fortunate communes, and transfers given to those operating in less-productive physical settings.

In the second lecture, on September 6, she discussed the need to adjust ratios between labor and other factors in all productive environments to make sure that no one fell below a minimum threshold of consumption. Yes, short-term growth and the speed of the transition to socialism would suffer, but the economy would not suffer unacceptable Soviet-style moral costs.

In the third lecture, delivered September 9, she said that Marx was theoretically analyzing rapid accumulation under capitalism, and so the theoretical framework did not suit a mostly pre-capitalist society like China’s. She outlined what Tahir, Harcourt, and Kerr (2002, 274) call “legal, market and moral” approaches that could be used to propel an economy. The strength of the Chinese model, as

6. As this essay was being prepared and circulated, the author learned of a new book on Joan Robinson and China, authored by Tahir (2019). Tahir is a longtime admirer of Robinson’s work, and some of his other work on her is cited in this paper. The author was unable to obtain a copy of the book by the time the revised version of this paper was submitted.
she saw it then, was that the apparatus for accumulation was being constructed in conjunction with a moral code appropriate to that apparatus. The Chinese model would avoid the problems faced in capitalist societies, where the moral code emphasized the propriety of making as much money as possible, with the result that distribution of income and the goods and services it could buy became unacceptable. Already there was a hint of why she would have such a favorable view of the GLF and CulRev as they were happening, namely that the Chinese were discovering a method of promoting Robinsonian accumulation that was both economically and morally preferable to capitalism and to the Soviet model.

This attitude is in conspicuous contrast to her skepticism almost from the beginning of Indian economic planning, which in her view neglected the fundamental moral questions of the country’s distribution of wealth and its rigid social structure (Robinson 1962b; 1976b). Robinson felt India was in thrall to its traditions, and economic planning, when grafted onto a society with these traditions and funneled through parliamentary democracy, could not work. Robinson (1962a, 113) contrasted India’s struggles in carrying out a transformation of its economic structure with China, where “a violent reversal of ideas has opened the way for rapid changes in technology and in the social forms appropriate to exploiting them.” That she was enthralled by the Chinese model, despite this violent reversal, is indicated by the fact that despite having lived in India before independence and visiting to consult with scholars on development plans shortly after independence, she wrote far more about China.

Visit in 1963

Whereas before her 1963 visit, Robinson had seen China through the prism of capital accumulation and looked at the economy in terms of its ability to collect and allocate the surplus for that task, she gradually began to see the country now as in the process of seeking through experimentation a novel, better, and more appropriate way for China to accomplish this. She came to see the CCP commune system as the crowning experiment, a belief she would hold for the rest of her life. She indicated having visited “a dozen communes in four provinces” on the 1963 trip (Robinson 1973, 213).

In an account of the visit, Robinson (1977/1964) confidently said things that seem incredible now. She asserted that during the GLF famine, which she now referred to as “the bad years,” the “rationing system worked” (Robinson 1977/1964, 41). Rations had been calculated and honored based on proper criteria such as body type, age, etc. The Chinese reacted, she indicates, with immense “civic morality” (ibid.), this morality fully cultivated a mere 10 years after 1949. She praised a redirection of resources from previously prioritized heavy industry to
agriculture and light industry during the years of harvest failure, but such had not in fact happened. The opposite was true: although there is no reason to think the CCP told her so, agricultural produce was diverted from the starving communes to the cities and, in exchange for factory equipment to facilitate rapid industrialization, exported to the Soviet Union and Eastern Europe. These policies significantly exacerbated the famine (Meng, Qian, and Yared 2015; Yang 2012; Zhou 2012; Dikötter 2010; Li and Yang 2005; Chang and Halliday 2005). Rather than civic morality, the famine is represented in the newer, substantially more evidence-based scholarship by its viciousness, both in the macro sense in the policies of diverting commune production, and in the micro sense of the violence and other cruelties routinely inflicted by cadres in the countryside (Yang 2012; Zhou 2012).

Based only on accounts that were related to her by CCP officials, Robinson (1977/1964, 45) said confidently the GLF overall and the communes specifically were a success, pointing to construction and other investment, plus the GLF’s utility as a solution for a tradeoff inherent in agriculture. Family farms, she thought, were too small to take advantage of specialization and had higher transport costs, while big farms incurred high bureaucratic costs, were not conducive to worker discipline, and could not simultaneously take advantage of moral and monetary incentives to the extent that small farms could. Supervised communal production, in other words, struck a proper balance. She wrote that “it was the existence of communes which made it possible for the authorities to see the country through” (ibid.), in contrast to the harrowing accounts of life on them that have emerged in recent historical scholarship. As evidence, she raised the fact that Chinese famines previously would kill 10 million or more people, which is true. But private social organization, and any spontaneous, liberal improvement in agricultural productivity, were not available because the CCP had destroyed these networks after 1949. These networks and a potential for building new ones during crises had already been in evidence in China’s response to famine in 1921 (Fuller 2013). The CCP’s regimentation of society, notably its elimination of freedom to migrate to cities where famine-relief efforts had in the past been substantial, means that the GLF famine can be judged to have been substantially and unjustifiably worse than the 1921 famine. Robinson acknowledged that during the GLF there were mistakes, but using her method of seeing everything in macroeconomic terms described the mistakes as “overinvestment…which puts the economy into an unbalanced position,” early prioritization of heavy industry, and the government’s initial terminological mistake in calling the communes “communist” rather than “socialist” (Robinson 1977/1964, 46–47). She also praised the democracy of the

7. The hukou (户口) system had in 1954 eliminated freedom of movement by tying people to specific residences.
communes, but accounts have emerged of horrific abuses including beatings, deaths, and sexual abuse, for perceived malingering, hoarding, or, sometimes, merely violating the arbitrary will of cadres (see Yang 2012; Zhou 2012). Robinson’s tragically erroneous claims continued in a postscript added after the CulRev (Robinson 1977/1964, 53–55). In it she described the toleration of individual efforts to grow food during the GLF as “many concessions [that] had to be made to individualistic sentiment among the peasants; some communes actually disintegrated into private household cultivation” (ibid., 53). In contrast, Dikötter (2010, esp. Ch. 23) has described these efforts, to the extent they were successful despite official opposition, as critical to preventing the famine from being even worse.

Overall, Robinson’s attitude toward the famine now was that it was not nearly as bad as Western propaganda claimed, and that it was caused by a combination of weather problems and to a lesser degree planners’ mistakes—relocation of labor from agriculture to industry, starting too many projects at once in 1958, and too much enthusiasm for producing steel in the countryside. But the evidence uncovered by contemporary historians on the GLF and the famine is far too critical of both policy disasters and cruelty from Mao on down to justify her anodyne and even laudatory verdict (Robinson 1977/1964; 1980/1964), which she never meaningfully retreated from.

Visit in 1967

The visit that yielded the most troubling turn in her thinking was in 1967, during the early stages of the CulRev. It resulted in the most thorough work based strictly on her visits, The Cultural Revolution in China (Robinson 1970a). The trip took place in the fall of 1967. While it is now believed that Mao had been plotting the CulRev for some time as a means to counter the strong opposition he encountered in the highest circles of the CCP after the GLF, several opening shots of the CulRev proper have been identified. The first was in November 1965, when an article by Yao Wenyuan, later one of the Gang of Four, was published in two Shanghai newspapers. It criticized the performance several years earlier of a play by Wu Han called “The Dismissal of Hai Rui” (海瑞罢官). The play lionized a Ming dynasty official who spoke the truth not just to power but to the emperor himself about the extent of corruption among lower officials, and who was thus removed from office. The play was latched onto by Mao as an implicit attack on the Chairman himself. On May 25, 1966, a CCP secretary and teacher in the philosophy
department of Beijing University named Nie Yuanzi placed a big-character poster on a wall specifically identifying several Beijing party officials, charging them with seeking to bureaucratically colonize and therefore defeat what Nie called, based on a term already in use, the “Cultural Revolution” (文化革命). On June 1 the People’s Daily printed a now-famous editorial, “Sweep Away All Monsters and Demons!” (横扫一切牛鬼蛇神), which contended that while the bourgeoisie and exploitive classes had indeed had their property taken after 1949, the ideological struggle was still ongoing, and these classes were biding their time throughout China, waiting to pounce. Quoting Mao, it said the struggle between capitalism and socialism was unfinished. The next day the text of Nie’s big-character poster was repeated in the same paper. Shorn of many of the details, the basic account in Robinson (1968) is similar to this account.

From there the fanaticism and violence in China escalated rapidly. Teachers, including professors, especially in Beijing, were among the first targets—humiliated, tortured, and sometimes killed by their students. Different factions in government at all levels struggled for power. The CCP Central Committee in early August 1966 issued guiding principles for the CulRev, indicating that those who criticized the masses (indicating ‘rightists’ in the government) should not be trusted. The document is one of several Robinson chose to include in The Cultural Revolution in China (1970a, 71–80). On August 18, perhaps one million young people gathered to hear Mao pronounce in favor of continuing revolution, while his deputy Lin Biao argued for the destruction of everything old. Groups of students in middle and high schools then began to form what soon became known as the Red Guards, and they extended the cruelty down the educational ladder to teachers at almost all levels and across the country. People with unsatisfactory class backgrounds, real or imagined, also became targets.

By the end of 1966 the new revolutionary ethos had spread to factories in many large cities. Some workers, often younger, literally seized the means of production, taking over production facilities in objection to management structure and how workers were compensated. Several state media organizations in Shanghai and elsewhere were among them, to the applause of Mao. As 1967 unfolded, military leaders expressed alarm about the rapidly deteriorating situation, and in consequence social order dissolved even more into violence, as young people in the thrill of Mao’s radicalism, along with military units, fought each other and among themselves. Many factories and shipyards across the nation were now hardly functioning, with employees often not bothering to show up. Violence against individuals seemingly randomly targeted as class enemies on the street was common. On August 1, 1967, the radical magazine Red Flag ran an editorial urging the new revolutionaries to “take firm hold of the gun” (无产阶级必须牢牢掌握枪杆子) and drive counterrevolutionaries out of the army. Over several months in
1967 several embassies were attacked, culminating on August 22 with the burning of the British Embassy and physical attacks on several people in it, though no one was killed. In September, fearing civil war, Mao, his wife Jiang Qing, and other senior CCP leaders who had promoted the recent agitation temporarily changed position, with Mao personally touring the parts of the country that had seen the most violence and urging peace.

The above narrative describes the China in the grips of the CulRev that Robinson herself traveled to in November 1967. Violence was already rampant and would get worse until mid-1968, when the military and radical CCP leadership reached a modus vivendi. By then millions of class enemies and perceived enemies of Mao had been killed or otherwise purged (Chang and Halliday 2005). But while there was a general perception of chaos prominent in contemporary Western media coverage, Robinson dismissed that perception, and in her subsequent work she favorably described new developments in the economy, especially in the communes (Robinson 1981/1979; 1975; 1973).

Robinson’s views now transitioned to Harcourt and Kerr’s (2009) second phase, where she became as they see it naively radical. But her views about Mao during the CulRev were a natural evolution of her previous views, so perhaps we have naïveté compounded. In The Cultural Revolution in China (1970a), her primary goal was to explain and justify the reasoning behind the CulRev. Originally published in 1969, the only change in the 1970 version was the addition of a postscript, taking account of things that had happened since her trip in 1967. It reported remarks by Mao and others, including Marshal Lin Biao, who supported Mao against possible attack from the left on the Chairman as insufficiently pure, and who would die in 1971 in a plane crash over Mongolia. In presenting the statements she received in translation Robinson expressed no doubt whatsoever about their truth, and was equally credulous about quotations in official (there was no other kind then) Chinese media statements, including quotations from ordinary people therein. The book acknowledged but minimized the turmoil between 1966 and 1970, applauded a good 1968 harvest and indicated that “economic development seems to have been running on” (Robinson 1970a, 152).

The book proper had an introductory chapter of roughly 40 pages and then a report by an anonymous leftist participant in the 1967 turmoil in Shanghai, followed by several other documents then already known in the West that had promoted what became the takeover of production facilities in Shanghai and elsewhere, interspersed occasionally with Robinson’s analysis. The supplemental documents supported her interpretation of events, in which she agreed with and added to Mao’s and the CCP’s then-current theoretical argument for the need for the CulRev, which was that, as Soviet history indicated, socialism did not eliminate class conflict between the proletariat on the one hand and the exploitive classes on
the other. In particular, this fight was continuing in the form of socialist bureaucrats forming their own dominant class and passing on membership in this class to their children. Chinese ‘rightists,’ Soviet-style bureaucrats running the party in the factories, defended themselves by saying that they were needed to guide the uneducated masses. But Robinson was having none of that, and was sympathetic with the idea of the need to sweep this new ruling class out of leadership positions—in other words, of a continuation of the 1949 revolution. She believed that the number of people who needed to be scrutinized was small, but they had to be purged. She applauded overall the role played by students, including children, in eliminating the old order, minimizing the violence by saying that it was “not in the rule book, but it broke out from time to time” (Robinson 1970a, 25). As for the Red Guards, she excused them, including their “melodramatic and sometimes farcical aspects” (ibid., 26). Among these aspects were students’ attacks on teachers and principals in Beijing in August 1966, in which perhaps hundreds died, over a year before Robinson’s first CulRev-era visit to China. Robinson dismissed the violence overall by saying that “perhaps they [rightists] are still wondering what hit them” (Robinson 1970a, 27).

When discussing economics, The Cultural Revolution in China (Robinson 1970a, 34–39) did not describe in detail how the communes (were supposed to) operate now, although Robinson (1973) did so later. The book did discuss how essential the communes had been during the GLF, and Robinson wrote glowingly of record harvests since the famine’s end. All of this reflected her view of the CulRev as necessary to escape the shackles of the new, Soviet-style ruling class. Referring to CCP documents, which she took at face value, she emphasized the need to get transportation moving again and the avoidance of “interfering with production” (Robinson 1970a, 23) and the need (quoting Mao) to “grasp revolution and stimulate production” by “reducing red tape and reducing the ratio of administrative personnel to production workers” (ibid., 34). She believed that factories now stripped of administrative deadwood and fired by moral purpose would finish the accumulation of a surplus. She again dismissed the violence and destruction launched by the Red Guards and extended her blindness to other conflicts during the CulRev, saying they amounted to “bickering between rebel

9. The stories of some of these victims, who otherwise might have been forgotten to history, were recorded by Wang (2001).

10. So too, in the postscript written for the 1970 edition, Robinson evinced little curiosity about the fate of the ‘rightist’ Liu Shaoqi, Mao’s previous number-two purged in 1967 and previously when in power an advocate for policies to alleviate the GLF famine. She wrote only that “nothing has been said about the fate of the man, as opposed to the symbolic figure” (Robinson 1970a, 150). Liu the man had in fact been arrested in 1967 and would die in prison, emaciated, denied medical treatment—except, when needed for his trial, for Type 2 diabetes, yet while strapped to his cell bed (Chang and Halliday 2005, 524; Dikötter 2016, 235).
groups” (ibid.).

Robinson (1970a) also admired the “Agriculture should learn from Dazhai” (or Tachai) propaganda drive (农业学大寨). Dazhai was a village whose brigade—a collective farm smaller than a commune—was assigned to work low-quality land. Mao in 1963 admired how the communist spirit of the brigade seemed to lead to dramatic increases in production. Robinson transmitted the CCP’s account of the brigade’s experience without skepticism (Robinson 1970a, 36–37). But the nationwide efforts to emulate Dazhai included its techniques and engineering, even when the land elsewhere was not suitable. But Robinson felt vindicated, using Dazhai as a jumping-off point for discussion of the theoretical socialist rent proposed in her 1957 lectures and analyzed in Harcourt and Kerr (2009: 148). This was the use of transfers to equalize the geographical inequalities facing brigades, communes or other organizational units. Having modified her view now, she said the fact that the Dazhai brigade refused such payments was evidence in favor of the worth of moral values in moving beyond mere ‘economism’ (经济主义), an ideological target during the CulRev.

So Robinson was a strong believer in the CulRev, at the time, and in both of its official goals—sweeping away the remaining ‘rightist’ and bourgeois threats to China, in particular removing the new bureaucratic ruling class from CCP organizations, and the reorganizing of agricultural and industrial production in a fully cooperative way uncontaminated by monetary incentives. She characterized the CulRev as not only essential, but in fact having already been won, its violence negligible and the work of a small number of extremists (Robinson (1970a, 20). China was now building its economy “in a genuinely democratic manner” (ibid., 42). She pronounced Western media reports that were available to her as deceitful, misrepresenting the revolution “as mere chaos and disintegration,” and added confidently that “[t]o the historian of the future it will appear as the first example of a new kind of class war—a revolt of the new proletariat of workers and socialist enterprises and peasants turned to commune members against the incipient new class of organization men in the Communist Party” (ibid., 28). Having long said she was not a Marxist, she was now a Maoist.

11. Other Western scholarly praise of Dazhai came from Maxwell (1975), who largely agreed with Robinson’s perspective although he did not cite her work. In work Robinson surely would not have agreed with, Marshall (1979) attributed the increase in agricultural production there mainly to mere addition of resources by a government eager to see the campaign succeed, a sort of neoclassical production-function perspective. The brigade was depicted as causing substantial environmental damage by Shapiro (2001) and Dikötter (2016, 219–231). The Dazhai practice of ruling by fear through public arbitration of assertions of inadequate work by brigade members (自报公议) was emphasized by Kueh (2008, 26).
Robinson’s views in the aftermath of the Cultural Revolution

Robinson had rejected contemporaneous reporting in the Western media of chaos and conflict, and wrote that China was sweeping out the old and bringing in the new, with an eye toward building a better socialism (Robinson 1970c). Harcourt and Kerr (2009) described what followed as the third phase of her Chinese scholarship: a chastened abandonment of a mistaken enthusiasm for Maoist radicalism. But a detailed analysis reveals that after a full-fledged endorsement in The Cultural Revolution in China (Robinson 1970a) the scales only gradually began to fall from her eyes and, relative to the scale of the catastrophe of Mao’s rule, only modestly at that. A useful way of presenting her evolution is to investigate a series of articles she wrote for China Now and Monthly Review. The former was a monthly publication of the Society for Anglo-Chinese Understanding (SACU), an organization founded in 1965 by Robinson, the historian of Chinese science Joseph Needham, and the British diplomat Derek Bryan, who had served in both the Republic of China and the People’s Republic of China. From the society’s founding until the opening of China to broader foreign contact in the early 1980s it was one of the prime avenues for the British leftist elite to travel to China and engage the country’s leaders. China Now ran from 1970 to 1995 before being replaced by another publication. Monthly Review was and is a left journal of political thought published in New York City.

In her first article for China Now, Robinson (1970d) was fully invested. Despite her earlier emphasis on the importance of historical contingency, in this piece she argued that the Chinese model generalized to any “poor peasant society” (Robinson 1970d, 5). Her argument was built around a commentary by John Kenneth Galbraith (1969) and especially Prasanta Mahalanobis’s (1969) review of Gunnar Myrdal’s Asian Drama: An Inquiry into the Poverty of Nations. Robinson agreed with Mahalanobis that the key problem for a “poor peasant society” was the mobilization of an agricultural surplus, which was very difficult under capitalist farming because of the incentive to hoard grain in years of good harvest and sometimes to restrain crop production to increase profits. But the CulRev, Robinson wrote, was just the hammer for this nail; in a China of communes and planning this obstacle to generating the surplus would not exist. A commune could produce food rationally, without needing to pay attention to the artificial limits of “family ownership or tenancy” (Robinson 1970d, 5). The lesson from the unsuccessful (as she saw it) Indian experience, Robinson argued, was the importance of “freedom
from private property in the means of production” (ibid.). She quoted approvingly 1956 remarks by Mao arguing for heavy industry, but only in its place. She characterized the GLF merely as “a period of shortages and hardship” (ibid., 6), with bad weather conditions handled much more effectively than in the old China, the only mistake being that peasants had been amalgamated into communes too quickly for them to acclimate to the new environment. The communes themselves “were invaluable in nursing the country through the bad years” (ibid.). Conspicuously and dramatically unlike the cruelty of forced grain procurement later emphasized in the revisionist literature, she wrote: “The manner in which the surplus is collected has removed the burden from the poorest villages and put it where it can be least painfully borne” (ibid.). The strength of Chinese socialism was in its flexibility and open communication between “the masses and the leadership,” and the absence in them of what she called Stalin’s “tyrannous and unsuccessful” collectivization (ibid., 7). Now, socialist China and CulRev China specifically were in her view “the prescription for development” (ibid., 8). Jung Chang and Jon Halliday (2005, 434–435), in contrast, liken the communes to slave plantations.

In a China Now article written after an additional visit in 1972, Robinson told a tale that was much the same. This time she only mentioned Mao once, in the very last sentence: “In all these endeavors the unique characteristic of Chairman Mao’s leadership has been to trust the people, and the people are not letting him down” (Robinson 1974, 3). Her admiration for CCP economic policies, which she characterized as giving workers freedom to experiment and solve their problems, Dazhai-style, remained. She described the history of China since 1949 as “startlingly original and yet founded in common sense” (ibid., 2). The agricultural surplus had before 1949 been spent on “rent, usury and plunder,” but now much of it was being redistributed to the peasantry. As for the CCP’s management of the entire economy and society, she still remained amazed: “As the national plan of accumulation proceeds, the area of modern development gradually spreads, but growth of the great cities is checked and production is diffused over the interior, taking industry to the population, instead of dragging the population into the centers of industry. The communes have carried out heroic feats of investment in improving and creating cultivable land, turning what used to be ‘seasonal unemployment’ into a source of wealth” (ibid.). Clearly she was still a firm believer in both the China model, including communes, and in Mao.

By 1978, Robinson did change her view at least of some Chinese leaders, if not of Mao nor of the centrality of state-directed surplus extraction. In China Now (Robinson 1978a), she reacted to an article by the French Marxist Charles Bettelheim (1978b) in Monthly Review in which he had announced his divorce both intellectually and in terms of activism from Chinese communism in the wake of new reforms, in particular resigning from the Franco-Chinese Friendship Associa-
tion. Robinson began by expressing some regret that Maoism did not work out as well for China as she had maintained it would: “I think we all had a lot of wind in our heads; it was hard to believe that, in a socialist country, policy could have been the sport of personal ambition and it was deflating to be told that the Cultural Revolution is over and that the new aim of policy is modernization. We know only too well what it is like to be modern” (Robinson 1978a, 4). In other words it was court politics, and not the policies themselves, that deserved attention when evaluating the CulRev. She did not say here whose “personal ambition” was in play.

Having written so many times in the past about democratic discussion in China, now she believed there was even more: “There is more frankness of discussion, both amongst Chinese and between Chinese and foreigners, than over the last 25 years” (ibid.). Bettelheim (1978a) had depicted China’s incipient economic reforms as a betrayal of Maoism, in particular the linking of wages to productivity. Robinson (1978a, 5) disputed this, saying the CCP was now merely tinkering with a system that had now been in place for years, although excesses had happened. She asserted that workplace democracy remained, because workers decided among themselves to whom the available new bonuses would be distributed. It was “nothing like a change over to an incentive wage system” (ibid.).

With regard to the nature of the CulRev itself, the chaotic element had by now metamorphosed in her account from “melodramatic and sometimes farcical” (Robinson 1970a, 26) to “the period of anarchy” (Robinson 1978a, 5). To prove that its negative aspects had now been eliminated, she pointed out that unlike during the CulRev the “buses are punctual” (ibid.) and there was politeness between drivers and passengers, unlike the brawling that also took place during the CulRev. Rather than attributing the brawling to the chaos of the CulRev itself, she strikingly attributed it to the demoralizing effect under CulRev wage procedures of drivers “having to examine what type of conduct deserves an award” (ibid.). Factories were now conceded as having during the CulRev fallen into “factional strife,” but this was due not to anything Mao unleashed by design but to the fact that in some places the Gang of Four “was in the ascendancy” (ibid., 6). She ended the essay with words of wisdom from Mao, and the impression left was that the madness of the CulRev happened in spite of rather than because of him (a view also expressed in Robinson 1976a).

The next year, Robinson (1979) wrote again of Bettelheim’s resignation, which had been immediately addressed in Robinson (1978a). She now described Bettelheim’s Maoist purity as “infantile leftism” (Robinson 1979b, 25). While

---

12. In that issue of *Monthly Review*, Robinson favorably reviewed one of Mao’s books (Robinson 1978b), but apart from confirming her unwavering loyalty to the Maoist project this piece does not directly add to the analysis here.
continuing to criticize China’s new emphasis on modernization, Robinson
admitted mistakes by China but was forgiving, describing the country as having to
choose among several very imperfect options. She cited approvingly an essay in
the same issue by Jerome Ch’en that while praising Mao Zedong added that Mao
“was no God, his teachings no dogma; to be cool-headed toward his teachings has
nothing in common with revisionism” (Ch’en 1979, 23), though Robinson (1979)
did not express similar sentiments herself.

The last thing Robinson ever wrote on China was a short introduction to
the 100th edition of China Now (Robinson 1982). In it, she did not mention any
way in which the CulRev went off the rails, but merely repeated without skepticism
a 1981 statement from the CCP Central Committee talking of how harvests had
increased continuously during those years. Of her visits to China after 1967 she
wrote that she “was always fighting off disillusionment until I could no longer
accept the obscurantism, the violence and the downright silliness of the last stages
of the Cultural Revolution” (Robinson 1982, 3). As for her acknowledged previous
enthusiasm for the CulRev, Robinson felt reaction to it in the West was at least as
objectionable—“it is absurd to talk of 10 wasted years (1966–76) for a great deal
of construction went on in that period” (ibid.). She thought that China’s economic
problems now emanated from its rapidly growing population—a view which she
and many others then held, not just about China but about many developing
countries—and doing something about the size of rural families would be difficult.
China’s one-child policy had only been mandated in 1979. In this very short piece,
she said nothing that expressed unambiguous admiration of Mao, and she
acknowledged to some degree that real, substantial damage was done during the
upheaval, though as the quote above indicates her analysis was still wrong about the
chaos of the first two years, including the time she was there.13 Still she believed the
CulRev to have been essential for China’s progress.

Robinson’s enthusiastic support for Mao’s project from 1949 to his death
was continuous, and in this she stood out. Foreign visits to China were very rare
prior to 1978, and until 1972 those by Americans almost unknown, so Robinson
was in a privileged position to explore what the truth was and then tell the outside
world what she saw. After Richard Nixon’s visit in 1972 to China, two groups
of American academics, self-proclaimed “radical political economists,”14 were al-

13. The historical record suggests that the greatest and most shocking CulRev violence, including the
sometimes-fatal attacks by students on teachers, actually took place in the first two years, including while
Robinson was there in 1967. In an effort to control it, senior CCP leadership in December 1968 adopted
the policy of sending youth down to the villages to learn from the peasants (上山下乡), after the army
in 1967 had been asked by CCP leadership to reestablish control over the anarchy brought about by Mao’s
deputized agents of chaos, a task accomplished by mid-1968 (Walder 2019; Dikötter 2016).
14. The group called itself the First Friendship Delegation of American Radical Political Economists, and
lowed to visit China. While these American economists praised China’s achievements, some also reported that the Chinese economists they met with were only willing to speak in Marxist abstractions rather than discuss the details of economic planning, and many felt that a re-education facility they visited was “outright disturbing” and “simply shocking” (Weber and Semieniuk 2019, 24–26). Another useful comparison to Robinson is Amartya Sen, who in the year Robinson died had (considering the full spectrum of economic ideology) some similar views to a younger Robinson on the weaknesses of liberal economics and the importance of an active state. Nonetheless, even as Sen (1983) praised Mao’s China for its achievements in education and health, he noted that China’s lack of democracy was a serious drawback. Tom Buchanan (2012, 187) says, in comparison to other British leftists, Robinson provided “[t]he most sustained and immediate defence of the Cultural Revolution.”

**Economists on Robinson on China**

Robinson was one of the English-speaking world’s most influential economists, and so it is not surprising that numerous retrospectives of her work have emerged since her death. Her writings on China, both scholarly and non-scholarly, have been little considered in this work, despite the extensive amount of evaluation of her works on development economics and on the legacy of Marxism. The Chinese strand in her scholarship, visible from the 1950s, is either not mentioned or mentioned only cursorily in the wide-ranging assessments of her work by Marjorie Turner (1989) and Thanos Skouras (1981). In their reflection on Robinson’s work in the *Journal of Economic Literature*, Thomas Gram and Vivian Walsh (1983) do not discuss her China writings even though, as in the two works cited above, there is analysis of the contributions of her work to how economists subsequently saw Marxist economics and capital (Robinson 1966/1942).


15. Aslanbeigui and Oakes (2009) do not mention China at all either, although their book focuses on her work in macroeconomics and industrial structure rather than development.
is not raised in any of the papers in another edited volume published by Routledge, *The Joan Robinson Legacy* (Rima 1991), and only trivially in three of the papers in yet another Routledge volume, *The Economics of Joan Robinson* (Marcuzzo et al. 1996), which is comprehensive in the sense that it covers her work on macroeconomics, Marx, development economics, and capital theory. One chapter in the latter book, authored by Siro Lombardini, at one point speaks of how Robinson was “interested in the cultural and social-institutional conditions for development, as shown by her interest in the developments in China” (Lombardini 1996, 137). In Lombardini’s concluding paragraph, without going into any detail and citing only conversations he had had with her over the years, he says that Robinson believed that China had already succeeded in industrializing, and would achieve a superior alternative to capitalism “after a difficult transition” (ibid., 144). In the same volume Bertram Schefold (1996, 314) merely likens the “symbiosis” of the ancient Mediterranean world to “that of Red China and Hong Kong a few years ago.” And lastly, Harcourt, whose contribution had the theme of “Robinson’s changes of mind,” described her simply as being “both philosophical and practical about China” (1996, 324).

Among the papers collected in *Joan Robinson’s Economics: A Centennial Celebration*, a volume published by Elgar and edited by Bill Gibson (2005), the paper by Robert Blecker (2005, 343–345) is the only one to look at the success of East Asian economies, including China, in terms of the analysis of what he calls the “new mercantilism” in Robinson (1947/1937). A paper by Harcourt (2005, 26–28) places her views on China in the larger context of her views on development, and notes without criticism her confidence in the agricultural communes established during the GLF.16

Julian Gewirtz (2017), in a study of the influence of Western economists in China, briefly characterizes Robinson’s work as a substantial failure on her part. But not even Gewirtz examines in any detail how Robinson came to believe that policies under Mao were the engine of China’s ongoing progress, policies now widely thought to have been catastrophic.

In a 31-page biographical work, James Cicarelli and Julianne Cicarelli (1996) only mention Robinson’s China work once, and without citation, but they assert:

> As for comforting the afflicted, Robinson published many papers in the 1960s sympathetic to the economic progress being made in socialist countries, particularly China. This sort of attention from one of the world’s foremost economists certainly did nothing to tarnish the image of these emerging socialist nations and probably modified to some extent the negative portrayals frequently appearing in the Western press. Robinson’s infatuation with China

was revealed in many of her writings. She also penned works supportive of Cuba and published a ringing endorsement of North Korea’s so-called ‘economic miracle.’

Her unflinching support for socialist causes, coming as it did during the midst of the Cold War, was courageous; Robinson paid a high price for her convictions. Western economists were prone to dismiss the value of her work, believing it was all tinted ‘red.’ This made it easy, and respectable, to ignore her even when she had something of substance to say. In the end, Robinson’s affection for socialism, which had its roots in her liberal upbringing, was one of the prime factors that rendered her ineligible for a Nobel Prize. (Cicarelli and Cicarelli 1996, 20)

It is difficult to verify the extent to which Robinson’s writings on China were effective in altering the “negative portrayals” of the country in the Western media. But to the extent they did, that cannot be considered a victory for truth.

In all of the work evaluating or memorializing Robinson I have only been able to find three cases in which the author(s) explicitly acknowledge its problematic nature, although in each case very briefly. One is the aforementioned Gewirtz (2017, 56–57), whose background is in history rather than economics and who is not particularly an admirer of Robinson. The second example is the first three sentences of Tahir, Harcourt, and Kerr (2002):

Joan Robinson always admitted to a leaven of advocacy in her writings on China because she thought they could do something to offset when she perceived to be the hostility of most other scholars and commentators writing on China. It is true that some of her writings and assessments were far too partial and uncritical, especially during the period of ‘Cultural Revolution’ when Mao’s influence was at its greatest and the spirit of the (radical) age was a yearning for cult figures and the immediate establishment of Utopias. But if we look at the whole body of her writings on China from the early 1950s to the early 1980s (she died in August 1983) we get a more balanced view. (Tahir, Harcourt, and Kerr 2002, 267)

Neither citations nor further analysis of impartial or insufficiently critical writing is offered.

Thirdly, George R. Feiwel, a longtime heterodox economist at the University of Tennessee, wrote in the 1989 edited volume Joan Robinson and Modern Economic Theory:

Someone once said that there are things about which one ought to write a great deal or nothing at all. Joan’s writings on Mao’s China are a case in point. Why she was attracted to and fascinated by the possibilities of an alternate social economic design is fairly clear, but why she wore blinkers when she looked
at China is something of an enigma. The puritan in Joan was attracted to the aims of the Cultural Revolution: combat egoism and eschew privilege. One wonders how it is possible that Joan Robinson, the realist, did not perceive the human and material costs of the undertaking in practice. (Feiwel 1989, 94)

Feiwel did not elaborate.

In contrast, in the same 1989 volume—and in tune with the overwhelming majority of the reaction to Robinson’s work on China after her death—Irma Adelman and David Sunding (1989), in a 21-page assessment of Robinson’s work in development economics, devote over three pages to her writings on China. It is mostly a direct evaluation of the academic work, but at one point they write:

As Robinson herself admitted in a postscript to the preface of Aspects of Development and Underdevelopment (Robinson, 1981 [1979]), the evidence now coming out of Chinese statistics indicates that her conceptions of China, formed largely during the Cultural Revolution, were idealized. (Adelman and Sunding 1989, 716)

Robinson (1981/1979) was a reprint, and her new postscript in its entirety reads:

News which has come out of China since the death of Mao shows that some of the allusions in what follows to the success of Chinese agriculture were overoptimistic; all the same, the level of production and standard of nutrition favor comparably with those of the third world. (Robinson 1981/1979, x)

This remark was made after Robinson in the original 1979 preface had written, “I do not think that anyone would deny that the Chinese method of organizing a highly labour-intensive agriculture is more successful than any in the so-called free world” (Robinson 1981/1979, ix). Robinson’s 1981 brief addition to the 1979 preface, and Adelman and Sunding’s (1989) evaluation of it, are at best underwhelming as critical reflection. Referring later not to what Robinson had recently learned but to her entire history in China, Adelman and Sunding continue that Robinson’s “exposure to the Chinese case, and her disillusionment with Fel’dmanite17 development in East European socialist countries, seem to have had a humanizing influence on her development theory“ (Adelman and Sunding 1989, 714–715).

One might hope for a lengthier and more nuanced interpretation from the work of Harcourt and Kerr (2009), who have written extensively on Robinson. They do present the most comprehensive account of the evolution of her views

17. Grigorii Alexandrovich Fel’dman was a Soviet economist. From 1925 to 1931 he worked in Gosplan, the Soviet economic planning agency.
on China over her career, and when they wrote had far more knowledge about what transpired under Mao available to them. However, much of the section on China, which comes from a chapter on development, uses (verbatim, often) the same language as Tahir, Harcourt, and Kerr (2002). Evaluating Robinson on China, they write:

After Mao’s death in 1976, she discovered, to her horror, that the Chinese had not told the truth even to trusting analysts. This discovery marked the beginning of her third phase. As more information became available in post-Mao China, she looked back at her previous writings and put some of the record straight. It was a period of self-criticism; she admitted to having been starry-eyed about the decade of the ‘Cultural Revolution’ and she returned to supporting Rightist economic reform. Her story was not plausible, even when she was not misled by the Chinese, for sometimes she did not follow the logic of her own argument. Nevertheless it is possible to salvage from her thinking about and enthusiasm for economic development in China, a set of ideas that differ little from the views of those dubbed the Rightists in the so-called two-line struggle of Mao’s China. As this set of ideas is now, on the whole, dominant in China itself, it is, as we said, relevant and timely to consider them in our discussion of Joan Robinson’s approach and analysis. (Harcourt and Kerr 2009, 145–146)

There is an endnote in the original after the word “reform,” but it does not support an assertion that Robinson meaningfully recanted her views. It provides no citations of published work, instead merely recounting a conversation between the authors and Peter Nolan, the topic of which was Joan Robinson’s view that capitalist production had a moral foundation insufficient to achieve “full human potential” (Harcourt and Kerr’s words; Harcourt and Kerr 2009, 240–241). Exaggerated as it is as an account of Robinson’s change of heart, it adds to the impression that the extent of her retreat on China has been overstated by the small number of Robinson scholars who have chosen to raise it.

**Conclusion**

Intellectuals of the left have a spotty record of critical reflection when it comes to evaluation of leftist totalitarian regimes. Frequent stubbornness with respect to Stalin has been noted by William L. O’Neill (1982) and Amity Shlaes (2007), and with respect to China after 1949 (although without mentioning Robinson) by Paul Hollander (1981, 278–346). O’Neill (1982, 367) argues that much of the support for Stalin even as his crimes became known was motivated
by a belief that when making an omelette eggs have to be broken. O’Neill and Hollander note that dismissal of media reports about cruelties in communist societies was driven by skepticism of liberal societies and a belief that the media in them were servants of business and the government. So it was with Robinson, who was almost until the end in truly irresponsible denial that horrible things happened under Mao, and that Mao was to blame for any of it.

Whatever else may be said about China in 1966, clearly it was not liberal, nor was it bureaucratic in quite the Soviet style. From the first time Robinson went to China until the end of her life, she saw it as praiseworthy, and unique among communist nations. Robinson was guided in her admiration for China by two principles in which she, along with many others, resolutely believed. First, liberal economics was a system with severe moral and economic drawbacks. Second, she thought that the core initial task of any economic system—reaping and redeploying the investable surplus above basic needs generated not by cooperative interaction between entrepreneurs and workers when it was mutually beneficial, but merely by workers themselves, especially agricultural workers—was something that could be done better than had happened under liberal capitalism, with its waste and instability, and Soviet communism, with its blunderbuss implementation and its cruelty. Socialism and communism held promise because they did not have to appeal to greed as a motivation to induce people to work hard enough to generate the surplus, nor did they have to create the surplus inefficiently as entrepreneurs would.

From her first trip to China if not before, Robinson thought that the country, as born anew in 1949, offered the possibility of something better even than socialism or communism as they then existed. By substituting Mao’s vision of a just society and how to build it for the capitalist incentives that had made poor countries so wretched, brutal, and corrupt before, such countries could reach a modern standard of living without capitalism’s excesses and distortion of human nature. She continued to write about China to nearly the end of her life in 1983, and maintained throughout that through various experiments the Chinese were developing the right way. It is admissible that in the last three or four years of her life she began to sense that she had overstated the achievements of the CulRev, whose dark side went far beyond mere mischief. But she still thought these achievements were many, and never recanted views about the post-1949 Chinese experience more broadly, in particular not with respect to the GLF. And she never doubted in print that Mao and the CCP saw the country’s problem properly. Any serious deviations that occurred were the responsibility of renegade party leadership—first ‘rightists’ like Liu Shaoqi, who had gone Soviet, and then eventually people like the Gang of Four, who were driven by personal ambition.

Starting in 1978, the Chinese prioritized modernization, to the chagrin of
Robinson, and the CCP slowly began to see the value of the chance to make money in motivating the Chinese to go out and develop the country. For long after the GLF, many Chinese in the countryside had remained desperately malnourished, despite the reports of continually increasing harvests in which Robinson placed such stock. A catalyst for reform occurred in 1978 when farmers in the village of Xiaogang in Anhui province secretly agreed to divide parts of their commune, keeping or selling whatever they could produce on their own above their state obligations, and to raise each other’s children if some of them were arrested or executed.  

News of the success of the farmers’ experiment came to the attention of first other farmers and then local authorities, who approved similar experiments elsewhere. After that, economic liberalization proceeded (and still proceeds) in stages. State-owned enterprises were first given some management and pricing autonomy, and then small private businesses were tolerated, and then special economic zones were created along the coast. Many state-owned enterprises were then at least partially and often completely privatized, the volume of merchandise trade and financial contact with foreign firms grew substantially, and financial markets were established. The ability of Chinese to live and work where they like, while still not perfect, has become much freer as the CCP’s *hukou* system of tying people to particular places of residence has been progressively relaxed. Today’s China is in no sense fully liberalized, and it has its share of problems, for example substantial corruption and nearly unprecedented pollution. But there is effectively no one in China today advocating a return to how it used to be. Even government statements now refer to the CulRev as ‘ten years of turmoil’ (十年动乱) or the ‘ten-year catastrophe’ (十年浩劫). The balance between statism and voluntarism is something Chinese authorities are still working out, and in that they are to some degree mindful of the views of the newly empowered, significantly freer (at least in an economic sense) Chinese population. And according to government data they have moved substantially in the direction of the liberal economics, the ‘capitalism,’ that so concerned Robinson. Without question the right to earn a living as one thinks best is a foundational human right, and China’s economic reforms for that

---

18. The farmers are now honored in a museum in the village.
19. According to reforms announced in 2019, restrictions on moving to and living in all but the country’s thirteen largest cities are to be eliminated, and even in those cities restrictions on obtaining residence permits are to be loosened. Some children of the migrants, however, still do not have equal access to school.
20. The degree to which China is ‘capitalist’ is a matter of judgment. According to data from the National Bureau of Statistics of China (2018, Table 4.3), in 1995 private-firm employees were 7.4 percent of all employment, while in 2017 they constituted 61.3 percent. The trend in annual investment has been similar. But the CCP still plays a role similar to those of governments in other postwar East Asian societies in guiding the general industrial direction of the economy, even as most production relevant to daily life occurs in the nonminimally private sector, substantially though not entirely in response to market incentives.
reason alone deserve praise. But while China has made tremendous progress in this regard since 1978, its hostility to political rights has undoubtedly grown in the years since 1989, with the CCP under Xi Jinping no exception. But of course after 1949 there was no political liberalism whatsoever during any of the years when Robinson was admiring Mao’s China, with the possible exception of the brief Hundred Flowers period.

Western reporters were almost entirely unknown in China during this time, and in any event would not have been permitted to travel, engage, and freely observe China. But the GLF famine was reported at least in broad terms in the Western media during and immediately after it. Refugees who made it to Hong Kong told of what was transpiring. Also during the CulRev there were reports in Western media of Red Guard fanaticism and something approaching civil war. But Robinson chose not to believe such reports, dismissing them as substantially overstated, even as propaganda against a better social model. Her defense of Mao and his China was longstanding, comprehensive, and uncompromising. It was until the very end also unreflective, and then only very modestly so.

Robinson had a responsibility to not accept at face value the story as it was told to her by CCP-authorized persons, many of whom she came to know well, during her visits to China, nor to place such blind faith in what she acknowledged (approvingly) to be radical social experimentation. No Western or expatriate Chinese historian today contends that the CulRev and the GLF were anything short of disastrous for the Chinese people. Robinson interpreted Mao and the CCP’s actions in the most charitable ways possible. The gap between what she was allowed to see and what transpired all around was vast, in both the purely economic and broader moral senses she emphasized. Many in the West, especially in the circles of power, were deeply hostile to the international communist movement, and so one can forgive suspicions about the strongest claims made against communist nations. But when push came to shove Robinson seemed to believe completely the information handed to her by officials in a totalitarian state, where independent media, freedom of movement, and other basic features of a free society did not exist.

Robinson only visited North Korea once and only wrote one specific piece

21. It is common for revisionist history to generate criticism by scholars who produced the existing history, and some of the work cited here is of this kind. The criticism that emerged did not deal with any of the claims central to this article. Regarding Chang and Halliday (2005), a collection of mostly critical reviews released soon after the book’s release is presented by Benton and Chun (2010). The criticisms almost unanimously accept the catastrophic effect of the GLF and CulRev, and are full of the language of the vast, tragic, or criminal suffering of the Chinese people under Mao. Other criticism of Chang and Halliday (2005) by some historians of China is found in Fenby (2005), along with their response. On Dikötter (2010), see Ó Gráda (2011).
about it (Robinson 1980/1965). But her misreading of that country’s likely future was similar to her analysis of China before 1979, and looks equally irresponsible in hindsight. Citing the state of music and the arts, industrialization, electrification in particular, and very equal distribution of income, she described North Korea as a “nation without poverty” (Robinson 1980/1965, 208) whose experience meant that “all the economic miracles of the postwar world are put in the shade by these achievements” (ibid.). Later in this essay she even described how “every service is building up capacity so as to be able to rush aid to the south as soon as communications are opened up” (ibid., 213). South Korea was governed dictatorially and was very impoverished then, although its economic miracle had already begun, and obviously Robinson’s analysis proved to be nearly as wrong as it is possible to imagine.

So too it was with China, although with much more of Robinson’s theoretical scaffolding, and for much longer. The magnitude of her error, the contrast between the China she saw, the China that actually was during that time, and the China that has developed after Mao, especially given that her errors all followed from assumptions that she brought to her analysis, should be noted in future analysis of her work. In the end, Robinson’s writing on China did the Chinese people no favors, and with regard to China she became an example of something she had long criticized: a prisoner of her own models, no matter what that the reality was.

References


British and Irish Communist Organisation. 1977. The Politics of Revolutionary China. Bel-
fast: British and Irish Communist Organisation. Link


Hollander, Paul. 1981. *Political Pilgrims: Travels of Western Intellectuals to the Soviet Union, China


Robinson, Joan. 1943. *Private Enterprise or Public Control. Unless We Plan Now*. London: The
Association for Education in Citizenship.


About the Author

Evan Osborne is professor of economics, Wright State University. He reads and writes Chinese, and his current interests include Chinese history from the late Qing to the present. He has done work on ethnic conflict and more general social conflict, and on the economics of art, empirical analysis of litigation, development economics, and various topics in sports economics. He is married and the father of two children. His email address is evan.osborne@wright.edu.

Go to archive of Character Issues section
Go to March 2020 issue

Discuss this article at Journaltalk:
https://journaltalk.net/articles/6006/
Edward Leamer
Deserves a Nobel Prize
for Improving Argumentation
That Uses Statistics

Arnold Kling¹

LINK TO ABSTRACT

This book could be classified under the heading of “metastatistics.” Statistics is the theory of inferences ideally drawn from data. Metastatistics is the theory of inferences actually drawn from data… Metastatistics analyzes how the researcher’s motives and opinions influence his choice of model and his choice of data… it also deals with the social mechanism by which information is transmitted among individuals.
—Edward Leamer (1978, v)

Edward Leamer deserves the Nobel Prize in Economic Sciences for launching the movement to examine critically the uses of statistical methods in empirical research. The movement has had repercussions that go beyond econometrics. It has affected medicine and epidemiology, where John P. A. Ioannidis has been a leading figure in pointing out methodological failures (Ioannidis 2005; 2016; Begley and Ioannidis 2015). It has impelled psychology and behavioral economics to confront what has become known as the ‘replication crisis’ (Camerer et al. 2018).

Leamer noted that economists usually work with data that is observational, not experimental. With observational data especially the proper specification of the analysis is uncertain. To deal with mis-specification risk, researchers employ what Leamer (1978) termed “specification searches.” Leamer rigorously analyzed

¹. Mercatus Center at George Mason University, Arlington, VA 22201.
the pitfalls of this process, eventually leading economists to rethink their approach to doing empirical work.

Suppose that you wish to compare the performance of charter schools with public schools by looking at test scores for students in each type of school. How do you control for the way that test scores are affected by the characteristics of the students enrolled in the schools, not just the differences in the approaches to education taken by the two types of schools?

Before the publication of his 1978 book Specification Searches and his 1983 article “Let’s Take the Con Out of Econometrics” in the American Economic Review—two works which as of February 2020 together had nearly 6,000 Google Scholar citations (link)—the standard method was multiple regression. The investigator would have estimated a regression equation to explain student test scores using the school type (charter or public) while also including average income of parents, average education of parents, and other auxiliary variables. The goal of including the auxiliary variables is to account for differences in student characteristics, so that those factors do not contaminate the estimate of the effect of school type on test scores.

The originality of Leamer’s contribution is that he articulated a previously under-emphasized conflict between the mathematical derivations of inferential statistics in regression (i.e., significance levels and confidence intervals) and the routine practice of applied econometricians. The inferential statistics, as reported by computer programs, are applicable to an experimental setting. But most economic research is conducted on data generated by uncontrolled economic activity, not by a controlled experiment. In practice, the econometrician tries a large number of regressions before reporting the ones that he or she thinks are most informative. Some econometricians, most notably David F. Hendry and his collaborators in the United Kingdom and elsewhere, embraced the use of specification searches, and developed progressively more complicated algorithms to automate them (see, e.g., Krolzig and Hendry 2001). Leamer took the opposite position on specification searches. He forced the econometric theorists to deal with the reality of econometric practice, and he forced applied researchers to face up to the corruption of their reported inferential statistics. Once the power of Leamer’s critique sunk in, top journals became far less inclined to accept empirical work of the sort that had routinely been published before he wrote.

After Leamer pointed out problems inherent in the multiple-regression approach and the inevitable specification searching that it involves, economists have turned to quasi-experimental methods, as surveyed by Joshua Angrist and Jörn-Steffen Pischke (2010). To examine charter school performance, a researcher might look for instances in which admission to the charter school was determined by lottery. The researcher would compare students who attended the charter
school with public-school students who had entered the lottery but were not selected. The assumption would be that there are no systematic differences between students who won the lottery and students who lost the lottery, other than the type of school that these students wound up attending.

The significance of what Angrist and Pischke (2010) termed the “credibility revolution in empirical economics” can be seen in the John Bates Clark Medal awards given to researchers who participated in that revolution. Between 1995 and 2015, of the fourteen Clark Medal winners, by my estimate at least seven—David Card, Steven Levitt, Esther Duflo, Amy Finkelstein, Raj Chetty, Matthew Gentzkow, and Roland Fryer—are known for their empirical work using research designs intended to avoid the regression problems that Leamer highlighted. Duflo shared the Nobel Prize in 2019 for work that, atypically in economics, involved conducting field experiments rather than finding quasi-experiments in existing data.

Leamer has somewhat different views for how best to proceed with non-experimental data. Indeed, he has criticized some of the currently popular research methods (Leamer 2010). In his book *Macroeconomic Patterns and Stories* (Leamer 2009) he demonstrates an idiosyncratic approach to the challenges of finding convincing, robust patterns in macroeconomic data.

Here I explain Leamer’s critique, as spelled out in his 1978 book and his 1983 AER article. I review Leamer’s 2009 book on empirical macroeconomics, and I summarize his doubts about the ‘credibility revolution.’

Leamer’s long career also includes research in international trade and in regional economic forecasting. I omit that work from this essay.

## Specification searches

The issue that drew Leamer’s attention was this: In theory the multiple-regression approach means using the computer to estimate the one equation that the investigator knows best characterizes the data; in practice however investigators do not start out knowing the best way to characterize the data and they undertake an iterative, trial-and-error process in which they use the computer to estimate many different equations utilizing the same data. The trials in such a process are what Leamer (1978) termed “specification searches.”

Econometricians use specification searches because they did not conduct the process that generated their data. Therefore, they need to explore their data in order to learn more about the behavior involved. As Leamer writes, “an experiment defines a model. When a specification search occurs, the researcher reveals that he does not think an experiment was conducted” (Leamer 1978, 2).

With an experiment, the investigator controls how the data are generated.
The statistical analysis can be based on that a priori knowledge. With observational data—meaning data that are generated in the world by a process that the investigator does not control—the investigator is uncertain about what model should be used to describe the data. The iterative, trial-and-error process is used to discover the model that the investigator believes is most appropriate.

Leamer argues that with observational data specification searches are inevitable and necessary. But from the standpoint of statistical theory, what practitioners do is simply bad. They violate what Leamer (1978, 4) terms “the Axiom of Correct Specification.”

The Axiom of Correct Specification
(a) The set of explanatory variables that are thought to determine (linearly) the dependent variable must be
(1) unique,
(2) complete,
(3) small in number, and
(4) observable.
(b) Other determinants of the dependent variable must have a probability distribution with at most a few unknown parameters.
(c) All unknown parameters must be constant.

If this axiom were, in fact, accepted, we would find one equation estimated for every phenomenon… Quite the contrary, we are literally deluged with regression equations, all offering to “explain” the same event… “Believers” use ad hoc techniques to search for specification, throwing out insignificant variables here and there, for example, but they continue to regard the end result of such a methodology to be identical to the end result obtained in the experimental sciences (or at least cynically to act that way). Believers report the summary statistics from the \( n \)th equation as if the other \( n - 1 \) were not tried, as if the \( n \)th equation defined a controlled experiment.

At the other extreme are agnostics, who gladly admit the irrelevance of classical inference. … Agnostics may discount any statistical result until it has been employed in a prediction outside the data period. (Leamer 1978, 4)

Leamer argues that the practicing econometrician’s exploration of alternative specifications is a proper adaptation to the situation one confronts in a non-experimental setting. Given that you did not design the process that generated the data, you must engage in exploratory analysis. Otherwise, you are likely to overlook important information. In his AER paper, Leamer writes

A study of the anomalies of the data is what I have called “Sherlock Holmes” inference, since Holmes turns statistical inference on its head: “It is a capital mistake to theorize before you have all the evidence. It biases the judgements.” (Leamer 1983, 40)
In *Specification Searches*, Leamer writes

Sherlock solves the case by weaving together all the bits of evidence into a plausible story. He would think it indeed preposterous if anyone suggested that he should construct a function indicating the probability of the particular evidence at hand for all possible hypotheses and then assign prior probabilities to the hypotheses... Sherlock Holmes procedures are an essential feature of scientific learning. But when models are instigated by the data, the traditional theories of inference are, regrettably, invalidated. (Leamer 1978, 286–287)

The Sherlock Holmes method of using the data to generate hypotheses is a key motivation for specification searches. It is the subject of an entire chapter of Leamer's book and is the main form of specification search discussed in the AER piece. In the article, Leamer goes on to write:

[A] theory constructed before seeing the facts can be disastrously inappropriate and psychologically difficult to discard. But if theories are constructed after having studied the data, it is difficult to establish by how much, if at all, the data favor the data-instigated hypothesis. For example, suppose I think that a certain coefficient ought to be positive, and my reaction to the anomalous result of a negative estimate is to find another variable to include in the equation so that the estimate is positive. Have I found evidence that the coefficient is positive? (Leamer 1983, 40)

Consider the hypothetical example of comparing charter schools with public schools. Suppose that in the first regression that we run, controlling for the income and educational attainment of parents, we find a small, insignificant advantage for public schools. We are disappointed by this result, and we explore the data further. We consider that ethnicity may have a causal effect that is not picked up by parents’ income and education. We add to the equation variables that will indicate whether a student is African-American or Hispanic or Asian. In this equation, charter schools deliver significantly higher test scores, and this is the result that we report.

From the perspective of the Axiom of Correct Specification, we should never have tried the second regression. But it is neither realistic nor wise to suggest that only one equation can be tried with the data. What Leamer points out is that the specification search is not wrong in itself. What’s misleading is the reporting of the results. Rather than proposing to end the practice of specification searches, Leamer tries to come up with better ways to undertake reporting of the process. As we will see, the economics profession took the problem posed by specification searching seriously, but it gravitated toward using different research designs rather than different reporting methods.

One approach that allows specification searches and still provides for valid
statistical inference is to randomly divide the data into two (not necessarily equal-size) samples. The investigation sample is used multiple times for trial-and-error specification searches, arriving at the investigator’s preferred model. The second sample is ‘held back’ and used only once, to provide a statistically valid assessment of the preferred model. The results that the investigator reports would be based entirely on the results from the ‘holdback sample.’ This approach is most readily applied in situations involving ‘big data’ and machine learning, in which there is sufficient data to provide both a large investigation sample and a large holdback sample. It has been scarcely practiced in economic research. It is not likely that there would be enough data available to apply this method to the problem of comparing test scores at charter schools with those at public schools.

In *Specification Searches*, Leamer suggested a different approach in a Bayesian spirit. He suggested that each time you try a new specification, you should give weight to the previous specification in the sense of a Bayesian prior. If I first looked at test scores across schools without controlling for ethnicity, then I behaved as if I had a prior belief that the coefficients on these variables were likely to be near zero. Subsequently, when I include these variables, I should use a Bayesian prior distribution that incorporates that prior belief. The result will be coefficients somewhere in between what the data find and the prior of zero.

But such an approach involves many practical problems. For one thing, two different researchers, using the same data, could arrive at different coefficients on variables based on the order in which equations are estimated. Suppose I include only X in my first specification and you include only Z in your first specification, but then both of us estimate a second equation that includes both X and Z. When we report coefficients, my coefficient on X will be larger than yours and my coefficient on Z will be smaller than yours.

In later writing Leamer advocated instead for a more conscientious approach, which might be termed comprehensive sensitivity analysis. It was common practice for economists to undertake ad hoc sensitivity analysis. The investigator might have attempted, say, 100 different specifications, and then selected three or four of these to report in the published paper, implicitly claiming that these reported specifications cover the range of reasonable inferences from the data. In his AER paper, Leamer argued that this type of sensitivity analysis is inadequate.

The defect of this style is that the coverage of assumptions is infinitesimal, in fact a zero volume set in the space of assumptions. What is needed instead is a more complete, but still economical way to report the mapping of assumptions into inferences. (Leamer 1983, 38)
Leamer proposed a method, sometimes called “extreme bounds analysis,” in which one should attempt every plausible specification and then report for the parameter of interest the range of consequent values. If the range is narrow, then inference about the parameter is robust. If not, then any inference is fragile. He described it thusly:

Include in the equation the treatment variable and a single linear combination of the additive controls. Then find the linear combination of controls that provides the greatest estimated treatment effect and the linear combination that provides the smallest estimated treatment effect. That corresponds to the range of estimates that can be obtained when it is known that the controls are doubtful. (Leamer 2010, 37)

Leamer’s recommendation was criticized by Michael McAleer, Adrian Pagan, and Paul Anthony Volker (1985) and was not widely adopted by practitioners. Instead, what eventually transpired was what Angrist and Pischke called the ‘credibility revolution.’ Economists looked for quasi-experimental circumstances embedded in observational data. Recall our example of comparing students who won the lottery to get into charter schools with students who lost such a lottery. If the lottery is truly random, then there should be no systematic differences between winners and losers in terms of parental income, parental educational attainment, ethnicity, or other variables that might affect outcomes. The investigator can justify omitting all such variables from the study, thus avoiding the problem of specification searches. In less ideal situations where assignment is random conditional on controls, the selection of such controls is at least limited to those that are plausibly related to assignment, mitigating the impact of search.

In effect, much of the profession decided that arguments about empirical values could not be settled by any method that included specification searches. Specification searches create too many opportunities for the investigator to present unreliable results, either intentionally or accidentally, through choice of specification. In practice the parameters estimated using these conventional regression techniques on observational data were likely to be fragile, whether or not they were reported as such. Economists have come to express more confidence in the robustness of estimates obtained from quasi-experimental methods.

**Leamer the macroeconomist**

“The bad news is that I am not a macroeconomist and thus cannot claim an expert’s knowledge of the theory of the field. The good news is that I am not a macroeconomist, and I do not carry the heavy intellectual baggage that most
macroeconomists lug around.” So writes Leamer in the preface to Macroeconomic Patterns and Stories (2009, v–vi). That book was positioned as a textbook—it was subtitled “A Guide for MBAs.” But in fact it is a treatise that displays an original macroeconomic perspective that should have been a strong candidate for replacing conventional approaches in both teaching and research. The book deserves to be studied and its ideas debated within the highest reaches of the macroeconomic profession.

Leamer opens with a powerful epistemological statement:

You may want to substitute the more familiar scientific words “theory and evidence” for “patterns and stories.” Do not do that. With the phrase “theory and evidence” come hidden stow-away after-the-fact myths about how we learn and how much we can learn. The words “theory and evidence” suggest an incessant march toward a level of scientific certitude that cannot be attained in the study of the complex, self-organizing human system that we call the economy. The words “patterns and stories” much more accurately convey our level of knowledge, now, and in the future as well. It is literature, not science. (Leamer 2009, 3)

Perhaps it would have been better to equate macroeconomic pattern-seeking and storytelling with the study of history. As with history, macroeconomic events take place in sequence. We have only one sequence of events to study, with many possible explanatory variables.

The difficulty with macroeconometrics became apparent to me when I first read Specification Searches. This was in 1982, at a time when empirical macroeconomists were just coming to grips with problems caused by very high serial correlation. Because macroeconomic time series tend to have high serial correlation, when they are measured in levels all variables appear to be highly correlated. But when the serial correlation is treated by measuring the variables in terms of change from quarter to quarter, the correlation often drops to near zero. Equations uncorrected for serial correlation are severely biased. Equations properly corrected tend to be uninformative. Either way, coefficients in structural macroeconometric models are fragile in the extreme.

The ‘credibility revolution’ largely bypassed macroeconomics. Angrist and Pischke (2010, 6) wrote that in empirical macroeconomics “progress—by our lights—is less dramatic.” In my view, this is an understatement. In fact, no one has been able to revive the project of estimating structural macroeconometric models (with a consumption function, an investment function, a Phillips Curve, a money-demand function, and so on) using quasi-experimental methods. Leamer writes:

2. I further discuss the problems with empirical macroeconomics in Kling (2011; 2012).
Finally, I think that Angrist and Pischke are way too optimistic about the prospects for an experimental approach to macroeconomics. Our understanding of causal effects in macroeconomics is virtually nil, and will remain so. (Leamer 2010, 44)

Conventional macroeconomists deal in equations, such as a Phillips Curve or a money-demand function. Because macroeconomic data are quantitative, expressing macroeconomic theory in terms of equations may seem like a plausible approach. But because macroeconomic data reflect a single historical unfolding of events, the use of equations is not really justified. Just as a historian seeking to understand the rise and fall of empires or the outcomes of wars would not hope to gain much insight by writing down a mathematical model with a few equations, macroeconomists who bury themselves in equations are mostly wasting their time.

In *Macroeconomic Patterns and Stories*, Leamer eschews the usual approach of treating the economy as a system of equations. Instead, he examines the patterns in the data and then invites the reader to consider what sorts of stories might explain those patterns.

For each important macroeconomic variable, Leamer first examines how it is measured. He explains how statistical agencies arrive at numbers for Gross Domestic Product, employment, inflation, and so on. He then plots the data in terms of levels, quarterly changes, and moving averages of quarterly changes. Finally, he plots the data relative to quarters that the National Bureau of Economic Research has decided by committee vote were cyclical peaks.

At each step in his analysis, Leamer finds phenomena that are often overlooked by even the leading practitioners of macroeconomics. For example, macroeconomic data is routinely adjusted for regular seasonal fluctuations, and these adjustments are large. The unadjusted numbers show a drop from the pre-Christmas to the post-Christmas season that is much steeper than we see during bad recessions. As Leamer puts it:

> Though on average, nominal GDP has grown at the rate of 9.2 percent, the January–March first-quarter production has been down compared with the fourth quarter at the annualized rate of about −20%! After working so hard to get those gifts ready for the holiday, Santa and his elves huddle around the fireplaces in the winter and do not get much done. (Leamer 2009, 37–38)

Leamer also notes that the extent of seasonality has been declining over time. The decline points to another challenge overlooked by many macroeconomists: the structure of the economy keeps evolving. The share of workers employed in service industries has increased relative to that in goods-producing industries. The use of computers for tracking and forecasting has reduced the magnitude of inventory
fluctuations. General price inflation has become less variable cyclically, particularly when compared with secular divergence involving large price increases in higher education and health care with declining prices in computers and communications.

Looking at data on changes in employment using both monthly changes and three-month moving averages of monthly changes, Leamer discovered a phenomenon that to my knowledge no one else had previously found: momentum. That is, when the three-month increase in payrolls is well below average, the following month’s increase is very likely also to be well below average. But for this momentum to become established, it takes three months of below-average employment growth.

Another device Leamer uses is a chart of recessions and recoveries relative to the dating established by the NBER cyclical dating committee. This plot allows a direct side-by-side comparison of the ten recessions since 1948. The horizontal axis is the number of quarters before or after the NBER peak. To the left are quarters before the recessions begin; to the right are recession quarters. What is displayed is the percentage difference between employment and the employment level at the peak. (Leamer 2009, 111–112)

What one sees in these data is that the early postwar fluctuations were short and sharp. Employment fell dramatically during recessions but recovered quickly afterward. Periods of growth were interrupted relatively soon by another recession. More recently, growth episodes have been milder but longer lasting. This pattern also holds for the period following the severe recession of 2008–2009, which took place subsequent to the appearance of Macroeconomic Patterns and Stories. To me, this suggests that the phenomenon of recession and recovery is now very different from what it was decades ago. Recessions used to be dominated by manufacturing layoffs, with workers returning quickly to their previous jobs on the assembly line once excess inventories had been worked off. In the more recent recessions, inventory corrections have not played an important role. Workers have lost jobs to which they were never going to return, and the process of creating new jobs during a recovery has proceeded relatively slowly.

The goal of Leamer’s extensive empirical analysis is not to arrive at a set of equations to describe the economy. Instead, he arrives at stories. One of his key chapters is entitled “Idleness Stories,” in which he tries to explain how it is that the processes of market adjustment fail to ensure continuous full employment. He notes that employment is not just a momentary state. It is a long-term relationship.

This gives us a good unemployment story. In recessions with limited cash flows, facing difficulties servicing debts and making payrolls, some employers simply go out of business, ending forever their relationships with employees.
Other employers may continue to operate, but reluctantly sever their relationships with some employees. (Leamer 2009, 146)

In the subsequent chapter, Leamer tells “Cycle Stories,” attempting to explain how fluctuations in business conditions occur. His stories are based on the fact that the greatest cyclical volatility is in manufacturing and construction.

Overall, Leamer demonstrates the virtues of Sherlock Holmes analysis, of carefully examining the data in order to develop hypotheses. As a result, his exploration opens up different perspectives not found in conventional macroeconomics. If economists would read *Macroeconomic Patterns and Stories* with an open mind, without commitment to conventional theoretical frameworks, I am convinced that they would come away with many new insights. In my case, reading the book influenced me to think much more in terms of structural variation in the economy and to think less in terms of treating the economy as if it were one big factory producing GDP.

**Leamer on the shortcomings of natural experiments**


As Angrist and Pischke persuasively argue, either purposefully randomized experiments or accidentally randomized “natural” experiments can be extremely helpful, but Angrist and Pischke seem to me to overstate the potential benefits of the approach. (Leamer 2010, 32)

One problem Leamer notes is the issue of whether results found in a narrow quasi-experimental setting can be generalized to other contexts.

For example, how does [David] Card’s (1990) study of the effect on the Miami labor market of the Mariel boatlift of 125,000 Cuban refugees in 1980 inform us of the effects of a 2000 mile wall along the southern border of the United States? (Leamer 2010, 33)

Another example of the problem of extrapolating findings in one setting to a different context that Leamer points to is the behavior of mortgage-related securities during the financial crisis. Bond rating agencies and other quantitative analysts made predictions about the performance of these securities based on
historical norms for housing markets that did not apply in 2006–2008.

Because of the extrapolation problem, Leamer writes:

I thus stand by the view in my 1983 essay that econometric theory promises more than it can deliver, because it requires a complete commitment to assumptions that are actually only half-heartedly maintained. The only way to create credible inferences with doubtful assumptions is to perform a sensitivity analysis that separates the fragile inferences from the sturdy ones: those that depend substantially on the doubtful assumptions and those that do not. Since I wrote my “con in econometrics” challenge much progress has been made in economic theory and in econometric theory and in experimental design, but there has been little progress technically or procedurally on this subject of sensitivity analyses in econometrics. Most authors still support their conclusions with the results implied by several models, and they leave the rest of us wondering how hard they had to work to find their favorite outcomes and how sure we have to be about the instrumental variables assumptions with accidentally randomized treatments and about the extent of the experimental bias with purposefully randomized treatments. It’s like a court of law in which we hear only the experts on the plaintiff’s side, but are wise enough to know that there are abundant arguments for the defense. (Leamer 2010, 36)

Today, we know that the ‘credibility revolution’ was followed a few years later by the ‘replication crisis.’ That would seem to vindicate Leamer.

Friedrich Hayek titled his 1974 lecture accepting the Nobel Prize “The Pretence of Knowledge.” He wrote:

Unlike the position that exists in the physical sciences, in economics and other disciplines that deal with essentially complex phenomena, the aspects of the events to be accounted for about which we can get quantitative data are necessarily limited and may not include the important ones. While in the physical sciences it is generally assumed, probably with good reason, that any important factor which determines the observed events will itself be directly observable and measurable, in the study of such complex phenomena as the market, which depend on the actions of many individuals, all the circumstances which will determine the outcome of a process, for reasons which I shall explain later, will hardly ever be fully known or measurable. (Hayek 1992/1974)

Along similar lines, Leamer writes:

Let’s face it. The evolving, innovating, self-organizing, self-healing human system we call the economy is not well described by a fictional “data-generating process.” The point of the sensitivity analyses that I have been
advocating begins with the admission that the historical data are compatible with countless alternative data-generating models. If there is one, the best we can do is to get close; we are never going to know it. (Leamer 2010, 38)

Edward Leamer emphasizes to economists the uncertainty that we face and the humility with which we should make our claims. The profession has benefited from partially listening to Leamer concerning the weaknesses of multiple-regression methods. It would benefit even more by taking all of his methodological suggestions to heart.

References


About the Author

Arnold Kling is a member of the Financial Markets Working Group of the Mercatus Center at George Mason University. He is the author of seven books, most recently The Three Languages of Politics and Memoirs of a Would-Be Macroeconomist. He writes at Askblog (link). His email is arnold@arnoldkling.com.

Discuss this article at Journaltalk: https://journaltalk.net/articles/6007/
It Will Soon Be 1984…

Ingemar Ståhl’s pioneering 1979 proposal
to abolish cash to strengthen the Swedish welfare state

Ingemar Ståhl
translated by Alan Harkess
with an Afterword by Lars Jonung

It is clearly evident from recent developments that government authorities in many respects have failed to gain full control over the citizens of our country. This is especially notable in relation to the problems that have arisen from the spread of so-called grey and black markets and the associated increase in the number of transactions that have evaded taxation.

This is not just an efficiency problem in the sense that citizens—partly influenced by the tax evasion propaganda from the right-wing bourgeoisie and the ruthless attacks on society’s ambitions to construct a gentle People’s Home (folkhem) based on long-term government planning—are indulging in activities of dubious economic benefits to society as a whole. It also presents a sizable problem of social justice, equity, and possibly even of gender equality.

Trusting the tax authorities

It is above all a question of creating trustworthy relations between citizens and tax inspectors. The latter should form a group with whom citizens should be

---

1. Originally published in Svenska Dagbladet, December 31, 1979. The use of boldface is as in the original article. All footnotes are by Lars Jonung and are new in this republication. This article also will be included as a chapter in a forthcoming book in English about Ingemar Ståhl (Jonung and Jonung 2020).

2. The term folkhem, the people’s home, is a metaphor for the Swedish welfare state, prominent in the ideology of the Social Democratic Party since the interwar years. Core elements include ambitious social protection and policies to reduce inequality, a government-run pension system, universal health services, and free access to higher education.
able to discuss their economic problems and seek advice with great openness and trust.

It is of particular importance that the tax inspectors will have the opportunity to both explain and to persuade less solidaristic citizens—undoubtedly a diminishing group—of the necessity of surrendering most of their income to a society which by means of long-range government planning founded on solidarity has a greater capacity to achieve an efficient and equitable use of that income than would have been possible if individuals had been left to use it for themselves.

However, the Swedish tax system has an obvious flaw or perhaps just a blemish. This relates to the failure to register substantial numbers of transactions. Accordingly it is still possible for an individual to conduct a transaction in the black economy without it reaching the eyes and ears of our society. A few very simple steps could be introduced to deal with this obvious anomaly. It is a question of making minor administrative adjustments to the existing system rather than introducing a series of new measures.

We all have a personal identification number

Each citizen in Sweden is assigned a personal identification number (personnummer) by the tax authorities. Sweden has gained substantial international respect for the leading role that it has played in this field. All registered companies already have their own organization number (organisationsnummer). It is in fact only non-profit organizations, foundations and public authorities that are not required to have an organization number. At most, there are around 100,000 legal persons that do not possess such numbers for identification. It would hardly be a major administrative exercise to put an end to this anomalous situation.

The central principle underlying the proposal that I will outline here, which has undoubtedly occurred to numerous responsible citizens, is that every transaction should be recorded, specifying the seller, the buyer, and the nature of the transaction.

Hence it is a matter of providing a computer record of for example, my personal identification number (38 06 02-0859), the organization number of the

---

3. In public debate in Sweden, the term ‘society’ (samhället) is often used as a synonym for the state or the government. Ståhl adopts this usage here although as an economist he was keen to make a clear distinction between society and government.

4. Sweden introduced in 1947 a system of personal identification numbers covering every citizen. It was the first country in the world to do so. This measure was due to a change in the system of taxation. The Social Security numbers used in the United States are similar to the Swedish system of personal identification numbers although they do not cover every U.S. citizen.
local supermarket (56 78 34-8215) and a code that denotes the nature of the transaction (commodity group 1106, plain bread, weight 0.8 kilo, value 4.20 SEK). All of this information is currently available.

To simplify matters, it will be advantageous if every citizen is provided with a citizen debit card based on his or her personal identification number. When used for purchasing, the citizen’s personal identification number will be scanned automatically. On grounds of rationalization, most of the cash registers in the supermarkets will be equipped with computers and automatic scanners by 1984.

Once a month, large transaction units like major shops will submit electronic data to the national transaction register of the National Tax Board (Riksskatteverket). Smaller transaction units will not have to meet these monthly requirements. It will be sufficient for them to hand in manual lists once a quarter. Normally sellers have the legal responsibility for registering transactions. Compulsory registration of transactions is already the case for financial institutions such as commercial banks and to some extent the tax authorities make use of this opportunity.

The registration department of the National Tax Board will consequently be able to carry out monthly computer calculations that will provide information on the expenditures, incomes and changes in net balances for every natural and legal person in Sweden.

Free Saturday sweets!

Naturally, this will involve a relatively large volume of computer data. However, it does not seem to be overwhelming or unmanageable, not least since there are numerous advantages, as we shall see below. In total, it may be estimated that between 30 and 40 billion transactions per year are carried out by natural persons (ten transactions per day per citizen) and approximately the same number by legal persons. It would seem appropriate that small purchases such as evening newspapers and Saturday sweets that cost less than ten crowns would not have to be registered.

As a result of this change, which is actually nothing other than the rationalization of an existing system, the tax authorities will gain a complete overview of not just the incomes of individuals but also of their expenditures and changes in their holdings of assets.

5. The personal identification number of Ingmar Ståhl reveals that he is a male person born on June 2, 1938 in the city of Stockholm.
6. Ståhl proposes here the idea of medborgarköpkort, literally a citizen purchasing card. Today, this would amount to a debit card issued to every citizen.
It will be practically impossible to make purchases in the black-market economy since income and expenditure have to balance, allowing for changes in net wealth. In the same way, a stricter control of the value-added taxation will also be enforced.

A new form of taxation

The system also opens up the opportunity for a whole range of new theoretically interesting forms of taxation. A progressive expenditure tax becomes a practical possibility. The purchase of foreign currency or indeed all types of purchases abroad could become subject to value-added taxation in Sweden. Capital-gains taxation could be made more stringent and be extended to include personal goods and belongings.

It also raises the possibility of working with designated tax rates as part of the foreign exchange rate policy where taxes on imports and exports can be levied and changed to contribute to current-account balance.

Purchases made by foreigners in Sweden need not give rise to any problems. The individual seller would in principle take note of the purchaser’s passport number which would automatically harmonize with the national transaction register. Many foreigners will almost certainly be so fascinated by this rational Swedish system that they will apply for temporary citizen debit cards.

A certain amount of leakage due to minor cash purchases would, however, have to be accepted. The expenditure deficit due to such purchases could be made managed by means of a tax-free allowance, for example equivalent to the purchase of a newspaper once a day or a bar of chocolate once a week.

Higher quality of life

The benefits of the proposed system will however not just be a matter of improvements to the tax system. There will also be a whole range of positive effects that will be disseminated throughout society. First and foremost, there is the widespread sense of satisfaction, tantamount to an increase in the quality of life, arising from the improved cooperation between public authorities and citizens. These improvements will in turn help society to fulfill its objectives in relation to social justice and the idea of the Swedish folkhem.

---

7. A progressive expenditure tax, that is a tax on consumption, was the subject of debate in Sweden in the second half of the 1970’s.
Another equally notable effect is that the periodic reports from the national transaction register will allow citizens to improve their economic planning. In the long run, it is conceivable that further improvements in the quality of life could be achieved by allowing the National Board for Consumer Policies (Konsumentverket) to carry out statistical analyses of the data in the national transaction register in order to identify those individuals who could be said to have inappropriate “consumption profiles” from a social standpoint.

This could give rise to extensive outreach and preventive programs. In many cases, a conversation between the household and the National Board for Consumer Policies or the local municipal consumer board should be sufficient. In more deviant and problematical cases, society may have to intervene with more intensive consumer care measures.

Submission of a consumption plan

After some experience of the system, it would seem appropriate that each household every year submits a consumption plan. Once the plan has been approved by the National Board for Consumer Policies (Konsumentverket)—following the possible intervention of a consumer adviser or the municipal consumer planning board—the plan can be implemented. Automatic comparisons between outcomes as registered in the national transaction register and the approved consumption plans will indicate at an early stage where discrepancies arise and where appropriate measures should be taken. It is obvious that the national transaction register provides markedly increased opportunities to further the long-term planning of consumption and by extension, also of production.

It is also possible to improve consumer planning in particular designated areas. For example, the National Board of Health and Welfare (Socialstyrelsen) has for a long time taken responsibility for the Swedish population’s consumption of bread. The Board has nevertheless lacked the technical means to carry out comprehensive follow-up studies. The national transaction register will be able to deal with this shortcoming. Consequently, there will be no difficulty in attaching a separate code to bread transactions and presenting them separately. Then it will be possible to monitor whether the consumption goal of six to eight slices of bread per day is actually being fulfilled.8

---

8. In 1976, the marketing organization of the Swedish bakery association (Brödinstitutet) launched a campaign based on the slogan “The National Board of Health and Welfare wants us to eat 6–8 slices of bread a day.” The campaign gave rise to a heated debate about the involvement of the state in the life of citizens.
The proposed system is also able to ensure that consumers who over- or under-consume will be automatically requested to consult their district health care center, where they will be examined and advised by a team of specialists. (The composition of the team may consist, as it suits the case, of a specialist in primary care, a specialist in metabolic disorders, a medical welfare counsellor, a dietary specialist, a district nurse and a trade union representative from the Municipal Workers Union (Kommunalarbetarförbundet)).

The state-run pharmacies in Sweden are already in the process of introducing this type of transaction register. The possibilities of expansion into new areas will allow us to obtain rapidly information about the various side effects of consumption.

In the same way that the 1970s became the decade of the working environment, the 1980s could under the influence of the measures recommended here become a decade characterized by issues related to the consumption environment.

Solving the mystery of milk

The national transaction register opens up unimagined opportunities for research in the social and behavioral sciences. We will be able to compare consumption patterns in different groups and regions. Who drinks milk and why? This question will finally be answered.

The costly, time-consuming compilation of many of our official statistics could be replaced by simple, routine data processing in the national transaction register.

Deductions from the National Supplementary Pension System (ATP)

Tax evasion and property crime will at last become much easier to deal with as a result of the proposed system of a national transaction register. In principle, three types of crime are feasible: an expenditure deficit, an expenditure surplus and non-registered transactions. An expenditure deficit may arise from consumption abroad or black-market purchases (thereby avoiding the payment of value-added tax or at worst using non-taxed labor).

The submission of a travel plan backed up by appropriate documentation should help to deal with the problem of foreign consumption. Swedish tax could
thereby be levied immediately. The presumption made in relation to other types of non-recorded consumption is that they are a criminal offense. In this context, a tax surcharge can be automatically imposed electronically. The charge could be deducted from the individual’s capital holdings or ultimately from his or her supplementary pension (ATP).9

A recorded expenditure surplus arises when an individual has received untaxed foreign income or is in receipt of so-called black income. In this case, a Swedish tax or a tax surcharge will be imposed in the same way as with an expenditure deficit. The only difference will be that the relevant taxes will take the form of payroll charges and income tax.

**More police, of course!**

Transactions not recorded with the national register are obviously a serious crime. They arise principally when individuals exchange goods or services without a registered payment. Once such trade is discovered, it would have to be punished severely. It is anticipated that this would lead to a certain increase in police surveillance, opening of correspondence and the tapping of telephones. Home visits carried out by specially selected supervisors from the transactions register might also be used.

It must be emphasized however that these control measures will not affect honest, loyal citizens. They are solely directed at actual or suspected tax evaders. There is no reason why this type of measures which are in the interest of society as a whole should not be compatible with the principles of a state governed by law. It is not the aim of the legal system to protect criminals who sabotage the institutional foundations of society.

**Cash is criminal evidence**

As a result of the steady expansion of electronic surveillance systems, it can be expected that cashless transactions will increasingly become the norm. The purchase of newspapers, sweets and chocolate will continue to exist as exceptions in the little “free circle” outside the transaction registration system. This will also mean that a large cash holding in itself will give rise to suspicion of criminal acts or of criminal intentions. Solidaristic behavior by the citizens will quickly reveal any cash-rich saboteurs of the system of registered transactions.

---

9. The ATP pension was introduced in 1960 as a part of the state-run pension system.
Other forms of anti-social behavior can also be more readily suppressed by the transaction register system. Criminals who have committed thefts may be more easily detected by irregularities in their patterns of consumption. It is also likely that it would become easier to trace drug addicts. Hence additional research efforts might well be able to identify the income and expenditure profiles that are strongly correlated with various types of criminal behavior. The systematic scanning of individual income and expenditure profiles would also be able to make a marked contribution to the rationalization of police work.

**Goodbye to illicit distilling!**

The heated debate on keeping a register of the purchase of alcohol would be automatically resolved. At the same time, there would be a certain check on illicit distilling and wine production. A household that purchases bread, sugar, and yeast would immediately raise strong suspicions.

The issues of ‘tax exhaustion’ and tax evasion need no longer act as a brake on the continued growth of the welfare society. What the Khmer Rouge was unable to achieve because of their technological backwardness will be fulfilled by the technical ingenuity and cultural prowess of the Swedish people over the coming five-year period. We have already taken the decisive step by adopting modern computer techniques and personal identification numbers.

As is evident the transaction register system will provide us with a wide range of substantial advantages. It should thus be introduced without further delay. Already 1984 seems too late.

---

The introduction to Ståhl’s original article in Svenska Dagbladet stated that “We must create confidential relationships between the public and the tax inspectors! By 1984 every economic transaction should be registered and the tax authorities would then finally be able to have complete surveillance of everybody.”

The article included three illustrations. The first was a photo of a hand holding a citizen debit card in front of a supermarket cashier with the text: “With a few administrative changes we

10. Ståhl is referring here to the policy of Pol Pot, leader of the Khmer Rouge, of abolishing money after taking power in Cambodia in 1975.
can get rid of the obvious anomaly that people can consume whatever they want. A citizen debit card should be introduced at the latest in 1984 in order that the State may finally obtain control over the individual. . . , as Ingemar Ståhl writes in a New Year satire.”

The second illustration shows a pen for scanning price tags with the text “The system of automatic scanning of price-tags which is already in use can easily become the basis of the transactions registration system. The point of the cashier’s pen will be the eye of the State on your consumption.”

The final illustration comprised six pictures of the citizen debit card issued by the Consumption Control Agency for “Ståhl, Ingemar 38 06 02-0859, Consumption area: Lund” with a photo of Ståhl.

About the Author

Ingemar Ståhl (1938–2014) was professor of economics at Lund University from 1971 to 2005. He was strongly engaged in public debate on a vast number of issues like industrial policies, rent control, defense, taxation, health care, and social policy, promoting a market-oriented approach. He introduced public choice, financial economics, and law and economics into the academic curriculum at Lund University. He proposed and designed the Swedish system of student financial support, introduced in 1965. He was a member of the Royal Academy of Sciences, serving on the Prize Committee.

→ read the afterword by Lars Jonung
Afterword to
“It Will Soon Be 1984…”

Lars Jonung

Ingemar Ståhl (1938–2014) started his career as an economist in the 1960s by working for a large number of government committees dealing with such issues as higher education, energy, defense, infrastructure investments and the indexation of loans. He proposed and designed the Swedish system of student financial support as an expert for a government committee headed by Olof Palme (who subsequently became prime minister of Sweden). In those days he had close ties with the Social Democratic Party.

Initially, Ståhl was an applied micro-economist well versed in Pigouvian welfare economics and cost-benefit analysis. He made an exceptionally swift academic career, being appointed as professor of economics at Lund University in 1971 at the age of 33. As a young professor in the 1970s, he was fascinated by the rapidly growing public sector in Sweden and the accompanying sharply rising marginal tax rates.

His experience of serving as policy advisor made him increasingly pessimistic about the practical use of welfare economics. Gradually, he adopted and endorsed a public choice approach. In his opinion, the economist at the university should abandon the role of the social engineer and instead serve as an “eye-opener” revealing the political incentives driving economic policies. He introduced public choice in Sweden, by summarizing the subject in popular articles.

He moved towards a market-friendly view. In his opinion, the welfare state had grown too large and should be reduced in size. However, he was pessimistic about the possibility of downsizing the public sector due to its popular support across all parties.

By 1979, Ståhl was a well-known economist in Sweden. On New Year’s Eve that year he published the article republished above, titled “It will soon be 1984…,” in Svenska Dagbladet, a leading newspaper. He acted in the Swedish tradition of preparing an end-of-year, crystal-ball account of what the future may hold.

In his obituary of Ståhl, presented to the Royal Academy of Sciences in 2014,

1. Lund University, 221 00 Lund, Sweden. I have benefitted from comments by Kevin Dowd, David Laidler, Kurt Schuler, Claes-Henric Siven and Geoffrey Wood.
2. For a portrait of Ståhl, see Jonung (2019) as well as Jonung and Jonung (2018; 2020).
Assar Lindbeck described Ståhl in the following way:

Ingeamar had an exceptional flair for seeing connections between different phenomena. Most prominently, he had a great capacity to view the problems of society from unconventional perspectives. Conversations with Ingeamar were for this reason unusually rewarding.3

Indeed, the 1979 article is a nice illustration of this characteristic of Ståhl. In addition to George Orwell’s 1984, Ståhl found most likely inspiration for his proposal of a cashless economy in the monetary economics of Knut Wicksell, the first economist who analyzed in depth a model for a cashless economy. In Interest and Prices, Wicksell (1898) discussed two extreme cases of monetary systems: the pure cash economy, without any credit facilities, and the pure credit economy, with no cash at all.4

Ståhl’s 1979 satire for abolishing the use of cash preceded by several decades the present discussion of reducing the use of cash, most notably Kenneth Rogoff’s 2016 book The Curse of Cash.5 There are similarities as well as differences between Ståhl’s plan and Rogoff’s. The main similarity is the aim to reduce tax evasion and criminal activities. This is the first of two major arguments for Rogoff’s recommendation that higher-denomination notes should be taken out of circulation as they are used primarily in the underground economy.

As his second reason, Rogoff argues that phasing out cash would allow central banks to use negative policy rates more successfully when framing monetary policy. The public would no longer be able to respond to negative rates by hoarding cash, making negative policy rates more efficient. For obvious reasons, Ståhl did not propose this argument in 1979, a world of high nominal interest rates. However, it would fit perfectly with his view that new types of policy measures can be adopted in a cashless economy such as a progressive expenditure tax, new forms of foreign exchange rate policies—and negative central bank interest rates.

Like Ståhl, Rogoff (2016, 94) also suggests that the government “provides all individuals the option of access to free basic-function debit card/smartphone accounts,” similar to Ståhl’s idea of a citizen debit card. Both envision what Rogoff (2016, 98) terms “universal financial inclusion.” They would also allow a “free circle”—the term used by Ståhl—of cash for minor purchases, at least initially. Ståhl recommends a free allowance of up to 10 Swedish crowns, roughly three

---

3. Personal communication with Assar Lindbeck.
4. For interpretations of Wicksell’s pure credit system see, i.a., Bordo and Jonung (1987), Jonung (1979) and Laidler (2006).
5. See Hummel (2017) for a critical review of Rogoff’s work with a response by Rogoff (2017); see also Dowd (2019).
to four dollars today, while Rogoff suggests “up to a few hundred dollars or equivalent, perhaps a bit more” (Rogoff 2016, 93). Ultimately, they envisage a cashless or almost cashless economy.\(^6\)

Both of them suggest a period of several years to reduce the use of cash. In Ståhl’s plan, at the most five years are sufficient to hit the deadline of 1984. Rogoff is more cautious, seeing the necessary time span as at least 10–15 years (Rogoff 2016, 92). Rogoff (2017, 166) suggests at least 50 years for the move to an economy with only coins as cash.

Ståhl satirically promotes the all-encompassing social welfare state. Rogoff’s approach focuses on the United States and is more limited in scope. The idea is not to improve the workings of the welfare state but to restrict criminal activities and to advance the efficiency of monetary policy by eliminating cash. Rogoff is not going as far as Ståhl’s central idea of making all transactions traceable and stored electronically by the tax authorities. He adopts a rather partial analysis, although he states “that the overall social benefits to phasing out currency are likely to outweigh the costs by a considerable margin” (Rogoff 2016, 8).

All in all, Ståhl presents a satirical warning of a cashless future where the government has unlimited capacity to monitor the activities of the individual citizen while Rogoff proposes a cashless future of less crime and tax evasion and extended effectiveness of the central bank. Both have the same goal—a better society—although they have different routes to this goal. Ståhl fears that society will become worse off by restricting the use of cash, whereas Rogoff imagines that it would be a better world with less cash.

The article by Ståhl did not induce any comments or debate in the months after it appeared judging from the database of all major Swedish newspapers at the Royal Library in Stockholm. But in the decades since, and particularly in recent years, Sweden has moved more rapidly towards a cashless or less-cash economy than any other country. The volume of cash in circulation has been cut in half since 2007. Many shops and restaurants refuse to accept cash. As cash is used less frequently, crimes linked to cash, such as robbery of banks and of cash-in-transport and taxi robbery, have fallen sharply in the 2010s. No cash-in-transport robbery was reported to the police in 2018. The number of robberies of shops has fallen dramatically.\(^7\)

---

6. Although Rogoff (2016) is careful in putting his ideas in the context of earlier contributions to monetary economics, surprisingly he pays no attention to Wicksell’s work on the pure credit system. In addition, Wicksell’s analysis in *Interest and Prices* is presently the intellectual foundation of inflation targeting, the common policy strategy of central banks today. When Rogoff (2016) recommends in the second part of his book the reduction of cash to strengthen the impact of negative central bank rates, he is implicitly adopting Wicksell’s monetary framework in phrasing his argument.

7. Statistics taken from the report *Payments in Sweden 2019* issued by the Riksbank. See also Rogoff (2016,
Sweden is likely to be the first nearly cashless society—not through the policy of banning cash as suggested by Ståhl, but by the rapid evolution of new transaction technologies and new payment systems that make cash an inferior alternative. It is a spontaneous process driven largely by market forces rather than direct policy interventions. As a strong proponent of market solutions, Ståhl would have appreciated this process.

In the future, with the general adoption of more advanced payment techniques, the curse of cash is likely to be less of a curse. The measures recommended by Rogoff may become superfluous if the United States follows the present Swedish route.

References


107) on Sweden’s move towards a less-cash economy.

8. The case for negative interest rates is weakened by the Swedish experience of such rates in the period 2015–2019. The Riksbank has abandoned this experiment for other reasons than cash preventing an efficient central bank policy.

About the Author

Lars Jonung is professor emeritus in economics at Lund University. He has served as professor at the Stockholm School of Economics, as research advisor at the European Commission, Brussels and as chairman of the Swedish Fiscal Policy Council. His research is focused on monetary and fiscal policies and on Swedish economic thought. His books include The Long-Run Behavior of the Velocity of Circulation: The International Evidence (with M. D. Bordo, 1987); The Political Economy of Price Controls: The Swedish Experience 1970–1983 (1990); and, as editor, The Stockholm School of Economics Revisited (1991); Bertil Ohlin: A Centennial Celebration, 1899–1999 (edited with R. Findlay and M. Lundahl 2002) and The Great Financial Crisis in Finland and Sweden (edited with J. Kiander and P. Vartia, 2009). His email address is lars.jonung@nek.lu.se.
Bentham Versus Blackstone

Gertrude Himmelfarb

Jeremy Bentham’s first published work, *A Fragment on Government*, appeared in 1776, a date memorable for other reasons. Even as a literary event it was eclipsed by such more notable works as *The Wealth of Nations* and *The Decline and Fall of the Roman Empire*. Yet there is a peculiar appropriateness in Bentham’s maiden appearance at that time. For his philosophy was one of several versions of the “new science of politics” that competed for dominion in the New World.

Bentham himself was properly impressed by the historical importance of his work. His copy of the *Fragment* bore the handwritten note: “This was the very first publication by which men at large were invited to break loose from the trammels of authority and ancestor-worship on the field of law.” When the book was reissued in 1838 as part of Bentham’s collected works, an admiring editor included that comment. It is an eminently quotable statement, and few of his biographers and commentators have been able to resist quoting it—or, more important, to resist accepting it at face value.

“The trammels of authority and ancestor-worship” refers to William Blackstone, whose *Commentaries on the Laws of England* was the authoritative exposition of English law. Bentham himself, as a young student of the law, had attended Blackstone’s lectures at Oxford in the 1760s. More than half a century later, he recalled that even as a sixteen-year-old he had listened to those lectures with “rebel ears,” although he had not had the audacity to think of “publishing my rebellion.” When he did get around to publishing it, in the *Fragment*, it created—again, as he later remembered it—a “sensation.” He published it anonymously, not out of any lack of “audacity,” but because he thought that secrecy would stimulate curiosity

---

and thus sales. The strategy, he claimed, succeeded. The book was attributed to various authors, all of them “of the very first class”: Lord Mansfield (Chief Justice of the King’s Bench), Lord Camden (formerly Lord Chancellor), and John Dunning (formerly Solicitor-General, better known by his later title, Lord Ashburton). The sensation, unfortunately, dissipated when the true authorship was revealed by a doting parent.

Such was Bentham’s account, written in 1822 and first published as a “Historical Preface” to the 1838 edition of the Fragment. About fifty pages in length (half as long as the Fragment itself), rambling, anecdotal, often abusive, at times almost incoherent, the essay would surely have been dismissed as the meanderings of a man in his dotage, had it come from any less a personage than Bentham. At the very least, its allegations would have been subjected to scrutiny. Is there any independent evidence that the book was attributed to those notables, that it produced a “sensation,” that Mansfield was “delighted” with it, while others greeted it with “alarm and displeasure”? Almost all of Bentham’s biographers, commentators, and editors have echoed his assertions, often without making it clear that they were his. Indeed, such facts as Bentham himself offers belie his claims. In the “Historical Preface” he reprinted the whole of one of the reviews (there were two in all), and explained that it would have stimulated controversy had a friend not chosen to reply to it, thus putting an end to the welcome publicity. But the review said almost nothing about the substance of the book, objecting rather to Bentham’s “peculiar” and “tedious” mode of argument and his “conceit” in presuming to argue with Blackstone—hardly a selling review. Nor was the complaint about the lack of advertisements and the difficulty of obtaining the book from the booksellers—as if “the author may have had reasons for introducing it as privately as possible”—evidence of its having created a sensation.

There is a more interesting sense in which Bentham’s account is self-contradictory. If the Fragment could have been attributed to such eminences of the legal establishment as Mansfield, Dunning, and Camden, if it could have been received warmly by Mansfield himself and (again according to Bentham) by such other Tories as Lord North and Samuel Johnson, and if these claims could have been accepted and perpetuated by generations of commentators—surely we ought to reconsider the conventional roles in which Bentham and Blackstone have been

---

4. Ibid., pp. 504, 526, 540.
5. Ibid., pp. 515, 517.
7. According to his friend and editor, John Bowring, the first edition consisted of five hundred copies. A year after publication, Bentham reported to his brother that the stock of one bookseller had been sold out, but he later discovered that a parcel of the books had been mislaid in the warehouse. See Correspondence of Jeremy Bentham, ed. Timothy L. S. Sprigge (London, 1968), II, 103, 148–9 (March 10, 1777).
cast. Perhaps Blackstone was not quite the sacred cow of the establishment and Bentham the lone iconoclast bravely defying the “trammels of authority and ancestor-worship.” Not the least of the curiosities of this affair is the fact that no one has thought to ask how so subversive a work could have been attributed to the very people whose views were being subverted.

One final oddity: Toward the end of the preface Bentham casually mentioned one other person to whom the book had been attributed, John Lind. Today that name is virtually unknown, but at the time it was far better known than Bentham’s. Lind was the unofficial minister in London of the king of Poland, a friend of North and Mansfield and an occasional writer and journalist. In 1774 he wrote a critique of Blackstone which he gave Bentham for his comments. Bentham was so taken with the idea that after editing Lind’s essay he decided to write his own, whereupon Lind good-naturedly turned over the subject to him. Although Bentham privately acknowledged that Lind’s was the “parent” work to which his own was “much indebted,” he failed to make such an acknowledgment in public. Instead, he accused Lind of plagiarizing from him—the charge involving several sentences in a letter by Lind on another subject published in a newspaper. Whatever else the story suggests about Bentham’s character, it confirms the fact that Bentham was not alone in challenging the authority of Blackstone and that the intellectual atmosphere of the time was hardly as repressive and conformist as Bentham might lead one to think.

*A Fragment on Government* is sometimes described as a commentary on Blackstone’s *Commentaries*. It is in fact a commentary on seven pages of the introduction to that four-volume work. And those seven pages, as Bentham noted, were a “digression,” casual reflections on the nature of sovereignty and the origins of society and government. Reading the few pages that inspired this impassioned critique, one is struck by how unprovocative they were even from Bentham’s point of view. If one did not know otherwise, one might attribute some of Blackstone’s sentences to Bentham himself: the pronouncement, for example, on the sovereignty of the legislature (“sovereignty and legislature are indeed convertible terms; one cannot subsist without the other”); or the absolute nature of sovereignty (“supreme, irresistible, absolute, uncontrolled”); or the requisites of sovereignty (“wisdom to discern the real interest of the community; goodness, to endeavour always to pursue that real interest; and strength, or power, to carry this knowledge into execution”).

---

8. For a more complete account of this episode, see Gertrude Himmelfarb, “Bentham Scholarship and the Bentham ‘Problem,’” *Journal of Modern History*, 1969, pp. 200–4. Lind’s manuscript (or part of it) is in the collection of Bentham papers at University College London (Mss. XCV: 1–28).

9. The *Fragment* should not be confused with Bentham’s *Comment on the Commentaries*, a more ambitious but incomplete work which was first published almost a century after his death. But even this much larger work dealt almost entirely with two sections of Blackstone’s introduction.
and intention into action”). 10 Even the account of the origins of society was less controversial than might be thought, since the “social contract” was described as a figurative expression which did not imply any actual historical contract or state of nature. 11 The only part of this section that was obviously objectionable to Bentham was the eulogy of “mixed government”: a form of government “so admirably tempered and compounded” that it preserved all the virtues of the three classical forms and combined them in a harmonious whole. 12

If Bentham made so much of this digression, it was because he was convinced that Blackstone was a “determined and persevering” enemy of reform and that “the interests of reformation, and through them the welfare of mankind, were inseparably connected with the downfall of his works.” 13 And if he spent less time on the substantive issues and more on logical flaws—inconsistency, imprecision, faulty reasoning—it was because he believed that moral deficiencies revealed themselves in rational deficiencies, that Blackstone’s “antipathy to reformation” expressed itself in “obscure and crooked reasoning.” 14

The irony is that Blackstone, whatever might be said of his antipathy to reformation, did accomplish a considerable reform, and precisely of the kind that Bentham himself was later to undertake (with far less success). The Commentaries, after all, was the first serious attempt to systematize and clarify the entire body of English law. Elsewhere Bentham paid grudging tribute to it on this account, admitting that, for all its faults, it was a “work of light, in comparison with the darkness which previously covered the whole face of the law.” 15 But in the Fragment the most he would concede was a certain felicity of style; and even this was cause for criticism since it served to make attractive “a work still more vicious in point of matter to the multitude of readers.” 16

Bentham’s most devoted admirers would be hard put to absolve him of “crooked reasoning” (still more to attribute to him any stylistic felicity). A crucial part of his argument centered on a phrase that appeared not in this section but in the final volume of the Commentaries: “everything is now as it should be.” 17 Bentham cited this as if it had been applied to the entire judicial and constitutional system of England. But in fact (as Bentham admitted at one point), it referred

10. William Blackstone, Commentaries on the Laws of England (Oxford, 1765–9), I, 46–9. This, the first edition, was the one used by Bentham (Fragment, p. 401). Unless otherwise noted, all references are to this edition.
11. Ibid., p. 47.
12. Ibid., p. 51.
14. Ibid.
16. Fragment, p. 413.
17. Ibid., pp. 400, 407; Blackstone, Commentaries, IV, 49. Bentham quoted this first as “everything as it should be,” and then as “everything is as it should be.”
only to the Church’s laws regarding heresy. On another occasion he deliberately omitted Blackstone’s qualification of a statement, and then justified that omission in a typically obtuse fashion: “When a sentiment is expressed, and whether from caution, or from confusion of ideas, a clause is put in by way of qualifying it that turns it into nothing, in this case if we would form a fair estimate of the tendency and probable effect of the whole passage, the way is, I take it, to consider it as if no such clause were there.” 18 Elsewhere he chose to disregard not only the qualification but the sentiment itself, the substance of Blackstone’s opinion, when it belied the view that Bentham ascribed to him. Thus when Blackstone was perverse enough to criticize a particular law and propose a reform, Bentham found these so out of keeping with his “general disposition” that “I can scarce bring myself to attribute them to our Author”; at best they were evidence of “an occasional, and as it were forced contribution, to the cause of reformation,” and therefore not to be taken seriously. 19

If Bentham sometimes refused to attribute to Blackstone what Blackstone actually said, at other times he attributed to him opinions and words he never uttered. He made great play, for example, with the “perfection” claimed for the social contract: the “perfect habit of obedience” presumed to exist in political society and “perfectly” absent in natural society; the “perfect state of nature” or “state of society perfectly natural” as against a “government in this sense perfect.” 20 From the repeated and italicized appearance of “perfect,” and the large objection Bentham took to it, one might suppose that the word was Blackstone’s. In fact, he never used it in this context. On the contrary, he went to some pains to make it clear that there never was any state of nature, that the very notion was “too wild to be seriously admitted.” 21 His discussion of the subject was deliberately tentative and qualified, quite the opposite of anything like the “perfect state of nature” Bentham ascribed to him.

Nor did Blackstone credit the British constitution or government with “perfection,” as Bentham claimed. Certainly in the seven pages under review, the government did not appear as “all-powerful + all-wise + all-honest = all-perfect.” 22 The only time Blackstone used the word “perfection” in this context was when he spoke of those qualities of government—wisdom, goodness, and power—“the perfection of which are among the attributes of Him who is emphatically styled the Supreme Being.” 23 The point of this statement was exactly the opposite of

18. Ibid., p. 409.
19. Ibid., p. 420.
21. Commentaries, I, 47.
22. Fragment, p. 472.
that imputed to him by Bentham. It was God alone who had those virtues in “perfection”; human beings and institutions had them only to an imperfect degree.24 Even Blackstone’s praise of mixed government was expressed in utilitarian and relative terms. He may have been mistaken or excessive in his praise, but not absurd or nonsensical.

In his preface Bentham announced that his task was essentially negative—to “overthrow” a work that was all the more “vicious” because it was so influential.25 Toward the end he confronted the obvious question: If Blackstone’s idea of the constitution was so thoroughly unsatisfactory, what was Bentham’s own idea of it? His answer was disdainful and dismissive. “I may have settled it with myself and not think it worth giving; but if ever I do think it worth giving, it will hardly be in the form of a comment on a digression stuffed into the belly of a definition.”26 One can imagine the ridicule with which he would have greeted such an evasion on Blackstone’s part.

The Fragment itself yields only a few clues by way of an answer. The most positive idea in the book, and its most obvious claim to distinction, was the principle of utility. The opening page of the preface enunciated the “fundamental axiom”: “It is the greatest happiness of the greatest number that is the measure of right and wrong.”27 But that idea was only occasionally invoked and not at all developed, and the axiom itself, as Bentham acknowledged, originated not with him but with Beccaria, Helvétius, and Priestley. More distinctive was Bentham’s assertion that utility was “the sole and all-sufficient reason for every point of practice whatsoever.”28 While the principle of utility was common enough, in that exclusive and absolute form it may well have been unique to him. Thus he praised Hume for demonstrating that “the foundations of all virtue are laid in utility,” but went on to

---

24. Elsewhere in the Commentaries, although not in the section analyzed in the Fragment, Blackstone did use language suggestive of “perfection.” But when he did so, he qualified it by saying that while the principles of the constitution approached perfection, the practice fell short of it, which was why it was so important to understand and respect the principles. The last paragraph of the work, which he himself referred to as a “panegyric,” concluded by speaking of the faults of the constitution:

Nor have its faults been concealed from view; for faults it hath, lest we should be tempted to think it of more than human structure: defects, chiefly arising from the decays of time, or the rage of unskilful improvements in later ages. To sustain, to repair, to beautify this noble pile, is a charge entrusted principally to the nobility, and such gentlemen of the kingdom, as are delegated by their country to parliament. The protection of the liberty of britain is a duty which they owe to themselves, who enjoy it, to their ancestors, who transmitted it down; and to their posterity, who will claim at their hands this, the best birthright, and noblest inheritance of mankind. (IV, 435)

25. Fragment, pp. 420, 413.
26. Ibid., p. 473.
27. Ibid., p. 393. Here, as elsewhere, I have eliminated most of the italics. Bentham used italics so indiscriminately that to reproduce them all gives an undue impression of emphasis.
28. Ibid., p. 448.
rebuke him for making “exceptions” to that principle. Bentham’s own principle, however, while absolute and unqualified, had little positive substance; it was not even related to the “greatest happiness” axiom he had enunciated earlier. Instead, it served an essentially negative, critical function. There is surely some truth in Bentham’s criticism of Blackstone: that by being an “expositor” of the law rather than a “censor” (critic), by giving reasons for the law as it is rather than as it ought to be, Blackstone was bestowing upon the law an implicit “approbation.” The converse could be said of Bentham, who assumed the exclusive role of “censor”; by making utility the sole basis of the law as it is and ignoring the reasons for the law as it is, Bentham implicitly illegitimized the existing body of law.

On one subject Bentham and Blackstone were in agreement: their opposition to American independence. Although the subject as such was not mentioned either in the Commentaries or in the Fragment (even in the later editions of those works), their views may be deduced from their discussion of other issues—and from what was not discussed. Blackstone’s opposition to independence, for example, could have been used by Bentham as an example of his incorrigible “antipathy to reformation”—had Bentham himself not been equally opposed to it. Blackstone’s attitude is not surprising: as a staunch defender of the British Constitution and a supporter of the Tory government, he was bound to resist the Revolution. Bentham, on the other hand, professedly independent, untrammeled by authority and unawed by the constitution, might have been expected to sympathize with the American cause. How could he do less than Edmund Burke?

Bentham did, in fact, do far less. At one point in the Fragment he seemed to suggest that his philosophy allowed for a great tolerance of revolution, in principle at least, than Blackstone’s. Blackstone’s idea of a social contract, Bentham argued, made “a necessity of submission”; having contracted to enter society and form a government, to exchange their wills for the will of the sovereign, the people could not reverse that decision. This might have been a telling argument against a Hobbesian contract, but not against the Lockean variety which Blackstone held to and which posited, in effect, two contracts, one establishing society and the other

29. Ibid., p. 440.
31. Again and again one awaits some reference in the Fragment to America: in the chapter on the “Formation of Government,” when Bentham alluded to the American Indians but not to the colonists; or when he wondered at what point the Dutch colonies, claiming independence from Spain, could be said to be in a state of rebellion and therefore in a state of nature; or in the following chapter, when he cited numerous cases of political conflict (between the Spaniards and Mexicans, Charlemagne and the Saxons, and other more obscure examples) without ever mentioning the obvious case of George III and the Americans. He even derided Blackstone’s use of the word “founders,” again without any reference to America. See ibid., p. 452.
32. Ibid., p. 481.
government, so that the latter could be overthrown without reverting to a state of nature. This was a small difficulty, however, compared with what was to follow. For having accused Blackstone of making a necessity of submission, Bentham then proceeded to accuse him of being “eager to excite men to disobedience,” and to do so upon “the most frivolous pretences,” indeed, upon “any pretence whatsoever.” This incitement to sedition Bentham found in the doctrines of natural and divine law. Blackstone had written that no human laws should be “suffered to contradict” the laws of nature and of revelation, and if any did so, “we are bound to transgress that human law.” The practical effect of this injunction, Bentham said, was to force resistance “as a point of duty,” to “impel a man, by the force of conscience, to rise up in arms against any law whatever that he happens not to like.”

In place of this “dangerous maxim” Bentham proposed the doctrine of utility, which permitted each man to calculate his own “juncture for resistance” against the “probable mischiefs of submission.” While this calculus might justify the resistance of any particular individual, it did not provide any “common sign” or “common signal” that could serve as a collective call to resistance. In the absence of such a common sign, there were no grounds for revolution, and the sovereign remained inviolate. Moreover, sovereignty itself was unlimited. “Unless such a sign then, which I think impossible, can be shown, the field, if anyone may say so, of the supreme governor’s authority, though not infinite, must unavoidably, I think, unless where limited by express convention, be allowed to be indefinite.”

Lest it be thought that “express convention” referred to a written constitution limiting the sovereign authority, Bentham explained in a footnote that what he had in mind was the case of a state which had submitted itself to the government of another. Whatever infelicities or ambiguities might be found in this passage (it was typical of Bentham’s mode of expression and reasoning), the burden of his argument was clear. Revolution was “impossible” and sovereignty was “indefinite.” The supreme body, the legislature, had no “assignable” or “certain” bounds; there was nothing “they cannot do,” nothing that was “illegal” or “void” or “exceeding their authority.”

Since sovereignty was absolute, the question of whether a government was

33. Ibid., pp. 482–3.
34. This is, in fact, the way some radicals read Blackstone. William Cobbett, for example, much preferred him to Bentham (even the later Bentham, who professed to be a radical). In 1818, at a time of social unrest and political agitation, Cobbett cited Blackstone (perhaps not quite accurately) on the right of resistance. “I say, therefore, upon this point, what judge blackstone says: and that is, that the right to resist oppression always exists, but that those who compose the nation at any given time must be left to judge for themselves when oppression has arrived at a pitch to justify the exercise of such right.” See William Cobbett, in Political Register, 1818. His Legacy to Labourers (1834) was also full of quotations from Blackstone.
35. Fragment, p. 484.
free or despotic depended not upon any “limitation of power” but upon “circumstances of a very different complexion”:

On the manner in which that whole mass of power, which, taken together, is supreme, is, in a free state, distributed among the several ranks of persons that are sharers in it—on the source from whence their titles to it are successively derived:—on the frequent and easy changes of condition between the governors and governed; whereby the interests of the one class are more or less indistinguishably blended with those of the other:—on the responsibility of the governors; or the right which a subject has of having the reasons publicly assigned and canvassed of every act of power that is exerted over him:—on the liberty of the press; or the security with which every man, be he of the one class or the other, may make known his complaints and remonstrances to the whole community:—on the liberty of public association; or the security with which malcontents may communicate their sentiments, concert their plans, and practise every mode of opposition short of actual revolt, before the executive power can be legally justified in disturbing them.  

This passage may seem to support the claim that Bentham was a liberal in the American tradition established by the Founding Fathers. But in the context of the book as a whole, that claim is dubious. Here, as in his later writings, Bentham insisted that it was important to provide for good government without in any way limiting the power of government; indeed, the legislature was required to have unlimited power in order to satisfy the principle of utility and achieve the greatest happiness of the greatest number. It is questionable whether this doctrine of an “omnicompetent legislature,” subject only to the kinds of conditions Bentham specified—frequent elections, publicity, a free press, and freedom of association—is “liberal” in the usual meaning of that word, and even more questionable whether it resembles anything remotely like the American mode of liberalism, which depends precisely upon the limitation of power.

The denial of any limitation on power is hardly consistent with the kind of government established in the wake of the American Revolution. Indeed, the denial of the possibility of revolution would seem to preclude any kind of American revolution. Yet Bentham later gave quite a different account of his early views. In the preface to the Fragment written half a century later, he contrasted the English government, “the least bad of all bad governments,” with that of the United States, “the first of all governments to which the epithet of good, in the positive sense of the word, could with propriety be attached.” He did not specify what was

37. Ibid., p. 485.
good about the American government, but since he never wavered in his belief that checks and balances, the separation of powers, judicial review, a bicameral legislature, and a bill of rights were unequivocally bad, the “epithet of good” must have been considerably qualified.

Bentham also later claimed that he had not opposed the American Revolution as such but had only objected to the ground on which the Americans had justified their revolution—the principle of natural right instead of utility. This is a plausible view of the matter, but not, as it happens, a true one. In 1776, when Lind was writing a pamphlet attacking the colonists, Bentham prepared an outline of the arguments that should be used against them. Making no mention of either utility or rights, he based his case entirely on the principle of sovereignty: the “power vested in the crown” which invalidated the American claim to independence. Nor did he alter his position after independence had been achieved. Five years after the end of the American war, having almost completed his *Principles of Morals and Legislation*, he prepared to send a copy to Benjamin Franklin, with a letter candidly expressing his disapproval of the revolution. “If any…of the ideas contained in it [the book] should be the means of adding to the prosperity of your country (since the unhappy distinction is now made) it will be some consolation for the misfortunes you have been a means of bringing upon mine.”

Even the reference in the *Fragment* to the liberty of the press—one of the very few times the word “liberty” appears in that book—turns out to be problematic in the light of Bentham’s other writings at the time. Among his manuscripts is a thirty-page document entitled “Plan for a Government Newspaper,” written soon after the publication of the *Fragment*. Provoked by the “malignant,” “virulent,” “incendiary” attacks on the government in the opposition press (attacks directed especially against Lord North’s policy on America), he recommended that the government establish its own newspaper to present its own point of view. The bulk of his proposal consisted of a series of “maneuvers” and “screens” designed to conceal the government’s ownership and control of the paper: a title containing some such word as “candid” or “impartial”; a price low enough to attract circulation but not so low as to arouse suspicion about the subsidy; a printer known to have been prosecuted under the libel laws; and occasional articles mildly critical of the government.

A government-owned newspaper was justified, Bentham argued, because “the business of conducting newspapers may be considered a very important branch of national education.” While a minister was not authorized to take that

41. Ibid.
business “out of improper hands,” he did have the power to “put it into hands that he thinks proper.” Bentham did not address himself to the propriety of secrecy and deception, the “maneuvers” and “screens” to prevent the public from knowing that the paper was owned and controlled by the government. Nor did he try to reconcile these stratagems with the principle of publicity he made so much of on other occasions, and which was one of his main securities against the abuse of power. Although nothing came of this proposal, it hardly inspires confidence in Bentham’s solicitude for the liberty of the press. And it reminds us once more how equivocal his liberalism was.

If Bentham was not as liberal and progressive as he is often made out to be, neither was Blackstone as benighted and reactionary. It is curious to find that only two years after his impassioned attack on Blackstone, Bentham endorsed a penal bill drafted by Blackstone and William Eden; the only changes he recommended were designed to make the regimen of the prisoners more arduous and to “augment the terror” of punishment. It was this bill, providing for prisons in the form of “Houses of Hard Labour,” that later inspired Bentham’s first reform proposal, the Panopticon; and his proposal was considerably harsher than Blackstone’s. Blackstone also anticipated Bentham in advocating reforms of the criminal law, the game laws, and the laws governing property and inheritance. Even on the subject of parliamentary reform, Blackstone was amenable to change: in the Commentaries he criticized the rotten boroughs and suggested that there might be a reason to favor a “more complete representation of the people.”

But it was not for these reasons that the Americans, both before and after the Revolution, read Blackstone so diligently and, for the most part (with the notable exception of Jefferson), so respectfully. Long before Bentham proposed to create a “science” of the law by codifying and systematizing it, Blackstone had done just that. He had, in fact, used that very word, referring to the Commentaries as an attempt to create a “science of the law.” Mill is often quoted as saying of Bentham: “He found the philosophy of law a chaos, he left it a science.” But the same can be said—indeed, has been said—about Blackstone. The nineteenth-century jurist James Fitzjames Stephen was not alone in claiming that it was Blackstone who “first rescued the law of England from chaos.”

---

43. Bentham Papers, University College London, Mss. CXLIX.
45. Commentaries, I, 172.
46. Ibid., I, 4, 27, 30.
Daniel Boorstin has argued that in making law accessible not only to lawyers but to all educated laymen, Blackstone “did more than any other writer in the English-speaking world to break down the lawyer’s monopoly of legal knowledge.” This itself, as Bentham pointed out on other occasions, had a democratic effect, democratizing the law by demystifying it, as we would now say. It also had the effect of instilling among Americans a respect for English common law and principles of government. This is what Burke had in mind when he pointed out, on the eve of the Revolution, that nearly as many copies of the Commentaries had been sold in America as in England. That legal training, he said, gave the Americans the habit of thinking in terms of principles rather than mere grievances. The principles, moreover, were those of the mother country. “They are therefore not only devoted to liberty, but to liberty according to English ideas, and on English principles.”

However much Blackstone himself opposed the American Revolution, he gave the revolutionists the heritage of parliamentary and legal institutions that has come down to us today as the “Anglo-American” tradition. That tradition has been much modified over the years, and in some respects (the principle of “one man one vote,” for example) in ways that might have displeased Blackstone and would certainly have pleased Bentham. But in other respects we are even further from Bentham today than we ever were, further not only from his idea of an “omnicompetent legislature” but also from his proposals for legal reforms: the abolition of the jury system, the elimination of legal procedures that impede the swift execution of justice, and changes in the rules of evidence to admit whatever evidence, however obtained and from whatever source, a judge might deem relevant (including the testimony of a wife against her husband or a lawyer against his client).

The conflict between Bentham and Blackstone reflected a profound difference of philosophy and disposition. What Blackstone and the Founding Fathers had in common, and what Bentham notably lacked, was a large tolerance for complexity. When Bentham quarreled with the idea of mixed government, he was not only opposing the particular “mix” favored by Blackstone; he was rejecting any kind of mix, any multiplicity of principles and institutions. His own “political and moral science” derived from a single principle: the “sole and all-sufficient” principle of utility. It was the singleness of that principle, as much as the principle itself, that he took to be an essential part of his science, just as it was the singleness

of the legislature (“omnicompetent” and unicameral) that he took to be an essential feature of a rational polity.

The Founding Fathers, on the other hand, like Montesquieu, believed simplicity to be a feature of despotism and complexity a condition of liberty. So far from relying on a single principle, their “science of politics” was deliberately based upon the “efficacy of various principles.” Even while establishing a new nation and a new regime—and a republic at that—they deliberately retained as many features of the British system of law and government as were compatible with republicanism. Indeed, it was precisely a new government, and a republican one, that they believed most in need of a plurality of principles and competing institutions, of all the means that resourceful men could devise so that “the excellencies of republican government may be retained and its imperfections lessened or avoided.” Thus in addition to the separation of powers, checks and balances, judicial review, and bicameral legislature, they introduced one additional principle: federalism.

It is little wonder, then, that Blackstone, not Bentham, was a guiding spirit in the early years of the Republic. “In the history of American institutions,” Boorstin has written, “no other book—except the Bible—has played so great a role as Blackstone’s *Commentaries on the Law of England.*” Bentham’s works, by contrast, were almost unknown in America during the formative years of the Republic and well into the nineteenth century. If today it is Bentham who more often engages our attention, who appears as the more “modern” and “relevant” thinker, that can only testify to a profound misunderstanding either of Bentham or of America.

References


---

53. Boorstin, p.i.

About the Author


Go to archive of Watchpad section
Go to March 2020 issue

Discuss this article at Journaltalk:
https://journaltalk.net/articles/6009/