Econ Journal Watch
Scholarly Comments on Academic Economics
Volume 21, Issue 1, March 2024

IN MEMORIAM

Editor’s Notes: Acknowledgments, April 2022 through March 2024

COMMENTS

McKinsey’s Diversity Matters/Delivers/Wins Results Revisited
Jeremiah Green and John R. M. Hand

Global Non-Linear Effect of Temperature on Economic Production: Comment on Burke, Hsiang, and Miguel
David Barker

Rejoinder on Ergodicity Economics
Matthew C. Ford and John A. Kay

Counter-Reply to Naumenko on the Soviet Famine in Ukraine in 1933
Mark Tauger

ECONOMICS IN PRACTICE

What Is the False Discovery Rate in Empirical Research?
Tom Engsted

Revisiting Hypothesis Testing with the Sharpe Ratio
Michael Christopher O’Connor
## CHARACTER ISSUES

Classical Liberalism in Russia  
*Paul Robinson*  
156–180

Classical Liberal Think Tanks in Greece, 1974–2024  
*Constantinos Saravakos, Georgios Archontas, and Chris Loukas*  
181–211

## WATCHPAD

*Trygve Hoff’s Appeal to Ragnar Frisch: Four Letters from 1941*

Foreword  
*Hannes H. Gissurarson*  
212–218

The Hoff-Frisch Letters  
*Trygve J. B. Hoff and Ragnar Frisch*  
219–234

Foreword to “Christianity Changes the Conditions of Government”  
*Daniel B. Klein*  
235–238

Christianity Changes the Conditions of Government  
*Numa Denis Fustel de Coulanges*  
239–250
In Memoriam
Editor’s Notes: Acknowledgments,
April 2022 through March 2024

We are grateful to our institutional home and friend the Fraser Institute, Canada’s leading think tank, in particular Jason Clemens, Chris Howey, Cheryl Rutledge, and Venia Tan. For generous support we thank the John William Pope Foundation, Gerry Ohrstrom (through Donors Trust), and Warren Lammert. 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.

EJW extends a special thanks to David Barker and Sarah Richardson for hosting a gathering of several EJW people and others in Carmel, California, in February 2024, and to Stephen Walker for his help in organizing the gathering.

I would like to extend thanks to co-editors David Barker, Brendan Beare, Andrew Gelman, Robert Kaestner, George Selgin, and Larry White, editorial advisor Jane Shaw Stroup, web designer and master John Stephens, proofreader Jon Murphy, and my chief partner in the project, managing editor Jason Briggeman.

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 April 2022 through March 2024

Alberto Abadie Massachusetts Institute of Technology
Joelle Abramowitz University of Michigan
James Bailey University of Providence
Donald Boudreaux George Mason University
Mykola Bunyk L’viv Regional Institute of Public Administration
Patrick Button Tulane University
Glenn Diesen University of Southeastern Norway
Tom Engsted Aarhus University
Andrew Farrant Dickinson College
Joseph Ferrie Northwestern University
Price Fishback University of Arizona
Andrew Friedson  
J. Arch Getty  
Warren Gibson  
Adam Gurri  
Björn Hasselgren  
Garett Jones  
Robert Kaestner (2x)  
Jukka Kekkonen  
Bohdan Klić  
Arnold Kling (2x)  
Leo Krasnozhon  
Lars-Folke Landgrén  
Timothy Loughran  
Guy Madison  
Brett Matsumoto  
Deirdre McCloskey  
Tom Mroz (2x)  
Steven Nafziger  
Radu Nechita  
Olga Nicoara  
Steve Pav  
Maarten Prak  
Aldo Rustichini  
Georgios Sideras  
Slavisa Tasic  
Michael Thom  
Alexis Toda  
Arne Vanhoyweghen  
Stephen Walker (3x)  
Tim Weithers  
Alex Young  
Konstantin Zhukov

Authors replying to comments in EJW, published September 2022–March 2024

<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexander Adamou</td>
<td>VesselsValue</td>
</tr>
<tr>
<td>Scott Barkowski</td>
<td>Clemson University</td>
</tr>
<tr>
<td>Dominik Baumann</td>
<td>Aalto University</td>
</tr>
<tr>
<td>Colm Connnaughton</td>
<td>University of Warwick</td>
</tr>
<tr>
<td>Vincent Ginis</td>
<td>Vrije Universiteit Brussel</td>
</tr>
<tr>
<td>Oliver Hulme</td>
<td>Copenhagen University</td>
</tr>
<tr>
<td>Allen Kachalia</td>
<td>Johns Hopkins Medicine</td>
</tr>
<tr>
<td>Natalya Naumenko</td>
<td>George Mason University</td>
</tr>
</tbody>
</table>

(article link)
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>(article link)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michelle M. Mello</td>
<td>Stanford University School of Medicine</td>
<td></td>
</tr>
<tr>
<td>Joanne Song McLaughlin</td>
<td>University at Buffalo</td>
<td></td>
</tr>
<tr>
<td>Victor Novack</td>
<td>Soroka University Medical Center</td>
<td></td>
</tr>
<tr>
<td>Jan Ott</td>
<td>Erasmus University</td>
<td></td>
</tr>
<tr>
<td>Ole Peters</td>
<td>London Mathematical Laboratory</td>
<td></td>
</tr>
<tr>
<td>James Price</td>
<td>University of Warwick</td>
<td></td>
</tr>
<tr>
<td>Simon Steinkamp</td>
<td>Danish Research Centre for Magnetic Resonance</td>
<td></td>
</tr>
<tr>
<td>Benjamin Skjold</td>
<td>London Mathematical Laboratory</td>
<td></td>
</tr>
<tr>
<td>Arne Vanhoyweghen</td>
<td>Vrije Universiteit Brussel</td>
<td></td>
</tr>
<tr>
<td>Bert Verbruggen</td>
<td>Vrije Universiteit Brussel</td>
<td></td>
</tr>
<tr>
<td>Maayan Yitshak-Sade</td>
<td>Icahn School of Medicine at Mount Sinai, NY</td>
<td></td>
</tr>
</tbody>
</table>
Consultants, business leaders, and activists often promote the view that a strong and settled business case exists behind the normative view that firms should increase the racial/ethnic diversity of their employees. A highly influential piece of evidence in support of this view comes from McKinsey & Co., which in a series of studies in 2015 (“Diversity Matters”), 2018 (“Delivering through Diversity”), 2020 (“Diversity Wins: How Inclusion Matters”), and 2023 (“Diversity Matters Even More: The Case for Holistic Impact”) reports finding statistically significant positive relations between the industry-adjusted earnings before interest and taxes as a percentage of revenues (EBIT margin) of global McKinsey-chosen sets of large public firms and the racial/ethnic diversity of their executives. Exhibit 7 from McKinsey’s 2020 study (p. 20), reproduced in our Figure 1, summarizes their 2015, 2018, and 2020 studies’ results.

Dame Vivian Hunt, McKinsey’s managing partner in the UK and Ireland at the time and a coauthor on all four of McKinsey’s studies, crystalizes McKinsey’s view that greater racial/ethnic diversity in a firm’s executive team drives better firm financial performance: “What our data shows is that companies that have more diverse leadership teams are more successful. And so the leading companies in our datasets are pursuing diversity because it’s a business imperative and driving real business results” (link).

1. Texas A&M University, College Station, TX 77843.
2. University of North Carolina, Chapel Hill, NC 27599.
3. See Eswaran 2019; Holger 2019; Kim 2018; Lorenzo and Reeves 2018; Lorenzo et al. 2017; 2018; Richard, Triana, and Li 2021; Wittenberg 2017. Different or contrary views include those of Edmans 2018; Ely and Thomas 2020; Levitt 2021.
4. In the References section these McKinsey studies are listed under their authors’ names: Hunt et al. 2015; 2018; Dixon-Fyle et al. 2020; 2023.
Given the business and societal importance of determining whether greater racial/ethnic diversity in corporate executives does or does not correlate in a statistically reliable way with higher firm financial performance, the goal of our paper is to revisit McKinsey’s results by means of a quasi-replication. We do so by applying McKinsey’s empirical testing approach to those firms that were in the S&P 500® Index on 12/31/19. We choose 12/31/19 so as to keep our quasi-replication within the same pre-covid time window as McKinsey’s 2015, 2018, and 2020 studies. We focus on the large public US companies in the S&P 500® for two reasons. First, McKinsey would not provide us with their detailed datasets, nor the names of the firms in their datasets, so we were unable to directly replicate or investigate their analyses. Second, in their first study, McKinsey reports finding a statistically significant positive relation between the racial/ethnic diversity of the executive teams of 186 large public firms in US + Canada with annual revenues of at least $1.5 billion (hereafter, ‘US firms’) and the likelihood that these firms display financial outperformance (McKinsey 2015, 3–4). Since the firms in the S&P 500® are large US public companies, they likely match well with McKinsey’s set of US firms, even if we cannot observe which firms are in McKinsey’s datasets.

5. We speculate that one reason why McKinsey may have chosen not to disclose the names of the firms in their datasets may be that the firms involved are McKinsey clients, and if so, this prevents McKinsey from sharing the data because of confidentiality agreements with those clients.

6. McKinsey also studies large companies in the Asia-Pacific region, Continental Europe, Latin America, Sub-Saharan Africa, and the UK. We focus on the US, as racial and ethnic diversity is an ongoing and currently politically and socially important issue in the US and collecting racial and ethnic background data
To use their 2015 study as the representative example, McKinsey measures the racial/ethnic diversity of executives in US firms using a Herfindahl-Hirschman index applied to eight racial/ethnic groups, where race/ethnicity is judged by McKinsey researchers using the photos and names of the executives found on the firms’ 2014 websites. The eight racial/ethnic groups used were African ancestry, European ancestry, Near Eastern, East Asian, South Asian, Latino, Native American, and Other. McKinsey defines financial outperformance as a firm’s EBIT margin during the years 2010–2013 minus the firm’s national industry median EBIT margin over the same period, and they compare the likelihood of financial outperformance in the top vs. bottom quartiles of their US firms ranked on the degree of executive racial/ethnic diversity. As reproduced below from Exhibit 6 of their 2015 study, McKinsey reports that 61 percent of US firms in the top McKinsey quartile of executive racial/ethnic diversity measured in 2014 exhibited financial outperformance in 2010–2013, versus 41 percent in the bottom quartile. The difference of 20 percent is statistically significant based on the $z$-statistic of 2.0 ($p$-value = 0.05).

Figure 2. Reproduction of Exhibit 6 from McKinsey’s 2015 study (p. 6)

7. McKinsey does not report a $z$-statistic in their exhibit 6. Our calculation assumes there are 47 firms in each of the top and bottom quartiles, leading to our estimation of a $z$-statistic equal to 20% , $\left\{ \frac{61\times39\%}{47} + (41\times59\%}{47} \right\}^{\frac{1}{2}} = 2.0$ per the Bernoulli-based test of the difference between two proportions.
In contrast to McKinsey’s results, the key finding of our study is that we observe no statistically significant difference between the likelihood of financial outperformance as measured by the industry-adjusted EBIT margin of S&P 500® firms during the years 2015–2019 in the top vs. bottom quartiles of S&P 500® firms ranked on McKinsey’s executive racial/ethnic diversity metric measured in mid-2020. Instead, we find that 54.0 percent of S&P 500® firms in the top executive race/ethnicity-ranked quartile have a positive industry-adjusted EBIT margin vs. 51.2 percent in the bottom quartile, with the $z$-statistic on the difference of 2.8 percent being a not statistically significant 0.5 ($p$-value = 0.65).

Because our key finding differs from that of McKinsey and does not provide empirical support for McKinsey’s interpretation that greater racial/ethnic diversity in a firm’s executives “is a business imperative that drives real business results,” we expand our tests and critiques in several directions:

- We extend beyond the likelihood of financial outperformance per se by calculating the mean levels of industry-adjusted EBIT margin in the top and bottom executive racial/ethnic diversity-ranked quartiles. Here too, we find a not statistically significant difference between the top and bottom diversity quartiles. The mean industry-adjusted EBIT margin in the top racial/ethnic diversity quartile is 1.9 percent vs. 0.8 percent in the bottom quartile ($t$-statistic on the 1.1 percent difference in means = 0.9).
- We relax McKinsey’s focus on only the top and bottom executive racial/ethnic diversity quartiles. Letting $DIV_{McK8}$ denote the degree of executive racial/ethnic diversity measured using eight racial/ethnic groups per McKinsey (2015) and using data on all S&P 500® firms, not just a subset, we find that $DIV_{McK8}$ is uncorrelated with the likelihood that a firm’s industry-adjusted EBIT margin is positive (Pearson correlation = 0.02, $t$-statistic = 0.5) and is also uncorrelated with firms’ industry-adjusted EBIT margin (Pearson correlation = 0.02, $t$-statistic = 0.5).
- Since in their 2018 and 2020 studies McKinsey uses a maximum of five racial/ethnic groups within a given geography to measure executive racial/ethnic diversity, including the US, we also repeat all our tests using $DIV_{McK5}$ instead of $DIV_{McK8}$. We find correlations using $DIV_{McK5}$ that are almost uniformly even closer to zero than with

---

8. We place all $n = 127$ S&P 500® firms with zero executive racial/ethnic diversity in the bottom quartile and only the bottom quartile; the $n = 124$ firms with the highest executive racial/ethnic diversity are in the top quartile.
We evaluate five other measures of firm financial performance: sales growth, gross margin, return on assets (ROA), return on equity (ROE), and total shareholder return (TSR), all on an industry-adjusted basis. For each, we repeat the tests described above based on industry-adjusted EBIT margin. This yielded 40 non-independent $z$-statistics or $t$-statistics testing the null hypothesis that there is no relation between firm financial performance and McKinsey’s metric for executive racial/ethnic diversity in US S&P 500® firms. Of the 40 test statistics, 37 are insignificant, one is reliably positive and two are reliably negative.

- We highlight the fact that even if our results had agreed with those of McKinsey, McKinsey’s interpretation of their results, namely that US publicly traded firms can deliver improved financial performance if they increase the racial/ethnic diversity of their executives, is flawed because their tests are structured so as to evaluate the opposite direction of causality, namely that higher firm financial performance leads to or causes greater executive racial/ethnic diversity.

In conclusion, our results indicate that despite the imprimatur often given to McKinsey’s 2015, 2018, 2020, and 2023 studies, McKinsey’s studies neither conceptually (in terms of the correct direction of causality) nor empirically (in terms of their set of large US public firms) support the argument that large US public firms can expect on average to deliver improved financial performance if they increase the racial/ethnic diversity of their executives.

Data and metrics

Firms and executives

We gathered data on the race, ethnicity, and other characteristics of all executives for all firms that were in the S&P 500® Index at 12/31/19. We follow the website-disclosure approach of McKinsey (2015) by defining an executive as any individual who is publicly disclosed by a firm to be on its leadership team,

10. A full description of the executive characteristics that were coded is provided in the Appendix. Many of the reported items do not pertain to this study but may be relevant to other research questions and projects.
most often on the firm’s website. In the infrequent cases in which we found no executives on the firm’s website, we took a firm’s executives to be the employees listed on the firm’s Bloomberg or Yahoo! Finance profile page, else the firm’s annual report, else we judged them from among the employees on its Comparably.com page.\textsuperscript{11} Primarily from each firm’s website, we then tracked down and, where present, captured in a screenshot the face photo of each executive, together with her or his first and last names.\textsuperscript{12} The Appendix presents raw firm and executive data items for two example firms.

Table 1. Waterfall criteria applied in arriving at those S&P 500\textsuperset{\textregistered} firms that were publicly traded on US stock exchanges at 12/31/19 and for which at least one named executive was found on the firm’s website, or the firm’s Yahoo! Finance profile page, or the firm’s Bloomberg profile page, or the firm’s Annual Report, or on comparably.com

<table>
<thead>
<tr>
<th>Step</th>
<th>Waterfall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. # firms in S&amp;P 500\textsuperset{\textregistered} Index (SP) at 12/31/2019</td>
<td>500</td>
</tr>
<tr>
<td>less: 2. # SP firms with no website or no executive/s on firm’s website</td>
<td>(9)</td>
</tr>
<tr>
<td>plus: 3. # firms of the n = 9 SP in Step 2 where \geq 1 executive was found on Yahoo! Finance, Bloomberg, Annual Report, or comparably.com</td>
<td>6</td>
</tr>
<tr>
<td>= # SP firms with \geq 1 named executive</td>
<td>497</td>
</tr>
<tr>
<td>less: 4. # RS firms in the n = 497 above where no executive photo could be found</td>
<td>0</td>
</tr>
<tr>
<td>= # SP firms with \geq 1 executive with a face photo</td>
<td>497</td>
</tr>
</tbody>
</table>

Note: Executives are defined as employees whose names are disclosed on the firm’s website as part of the firm’s executive, leadership, and/or management teams, or in its set of officers.

Table 1 presents our data availability waterfall. Based on our definition of an executive and the availability of individual data items, we arrived at 497 S&P 500\textsuperset{\textregistered} firms for which we were able to identify at least one named executive. In Table 2 we present descriptive statistics on the industry composition and selected financial characteristics of our sample firms at the most recent fiscal year-end on or prior to 12/31/19. Panel A reveals that in terms of the Fama-French 12-industry classification, S&P 500\textsuperset{\textregistered} firms are spread out, being most concentrated in Finance (20 percent) and Business Equipment (17 percent) and least concentrated in Consumer Durables (2 percent) and Telephone and Television Transmission (2 percent).\textsuperscript{13} Panels B and C present descriptive statistics on key firm financial

\textsuperscript{11} Yahoo! Finance’s profile page lists up to five executives. Bloomberg’s profile page typically lists 3–10 executives. Comparably.com lists up to 50+ people who work for the firm, only some of whom we judged to be executives.
\textsuperscript{12} The bulk of the capturing of executive names and photos took place June 10–August 5, 2020. For documentation and authentication purposes, we saved all screenshots of executives in a separate Word + PDF file for each firm.
\textsuperscript{13} McKinsey includes seven industries in their studies: Finance, Insurance, and Professional Services;
characteristics either at 12/31/19 or at the most recent fiscal year-end before 12/31/19, and show that S&P 500® firms are large from a capital market and accounting perspective and usually in strong financial positions.

Table 2. Descriptive statistics on the industry composition and selected firm financial characteristics at 12/31/19 or for the fiscal year ended on or before 12/31/19 for firms in the S&P 500® Index

<table>
<thead>
<tr>
<th>Panel A. Fama-French industry</th>
<th>Panel B. Firm financial characteristics ($ millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fama-French 12 Industry</td>
<td></td>
</tr>
<tr>
<td>Business Equipment</td>
<td>86</td>
</tr>
<tr>
<td>Chemicals and Allied Products</td>
<td>21</td>
</tr>
<tr>
<td>Consumer Durables</td>
<td>10</td>
</tr>
<tr>
<td>Consumer Nondurables</td>
<td>31</td>
</tr>
<tr>
<td>Finance</td>
<td>102</td>
</tr>
<tr>
<td>Healthcare, Medical Equipment, and Drugs</td>
<td>41</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>45</td>
</tr>
<tr>
<td>OIl, Gas, and Coal Extraction and Products</td>
<td>22</td>
</tr>
<tr>
<td>Other</td>
<td>57</td>
</tr>
<tr>
<td>Telephone and Television Transmission</td>
<td>11</td>
</tr>
<tr>
<td>Utilities</td>
<td>30</td>
</tr>
<tr>
<td>Wholesale, Retail, and Some Services</td>
<td>44</td>
</tr>
<tr>
<td>Market cap</td>
<td>$9,842 $20,646 $102,130</td>
</tr>
<tr>
<td>Total assets</td>
<td>$3,693 $22,684 $153,219</td>
</tr>
<tr>
<td>Total liabilities</td>
<td>$914 $14,563 $143,789</td>
</tr>
<tr>
<td>Total equity</td>
<td>$1,653 $6,732 $25,067</td>
</tr>
<tr>
<td>Revenue</td>
<td>$1,630 $6,611 $28,563</td>
</tr>
<tr>
<td>COGS</td>
<td>$438 $3,420 $17,342</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>$0 $0 $729</td>
</tr>
<tr>
<td>EBIT</td>
<td>$296 $1,395 $3,949</td>
</tr>
<tr>
<td>Net income</td>
<td>$175 $912 $3,050</td>
</tr>
<tr>
<td>CFOPS</td>
<td>$359 $1,697 $6,539</td>
</tr>
<tr>
<td>CAPEX</td>
<td>$0 $195 $3,218</td>
</tr>
<tr>
<td>TSR</td>
<td>-4% 31% 64%</td>
</tr>
</tbody>
</table>

Panel C. Annual financial performance over the years 2015–2019, both raw and Fama-French 12-industry median-adjusted

<table>
<thead>
<tr>
<th>Raw: Not industry-adjusted</th>
<th>5th percentile</th>
<th>Median</th>
<th>95th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBIT margin %</td>
<td>3% 17% 45%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue growth</td>
<td>-12% 6% 33%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gross margin %</td>
<td>13% 42% 88%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROA</td>
<td>-1% 5% 19%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROE</td>
<td>-23% 14% 67%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSR</td>
<td>-32% 12% 57%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw less median of FF12 Industry</td>
<td>5th percentile</td>
<td>Median</td>
<td>95th percentile</td>
</tr>
<tr>
<td>EBIT margin %</td>
<td>-17% 0% 23%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue growth</td>
<td>-15% 0% 26%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gross margin %</td>
<td>-28% 0% 39%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROA</td>
<td>-8% 0% 11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROE</td>
<td>-37% 0% 49%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSR</td>
<td>-38% 0% 41%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

McKinsey’s primary measure of firm financial performance is the firm’s average annual EBIT margin less the national median EBIT margin for the firm’s industry. The annual periods that McKinsey uses in its averaging are 2010–2013

Heavy Industry; Healthcare and Pharmaceuticals; Telecom, Media, and Technology; Consumer Goods and Retail; Transportation, Logistics, and Tourism; and Energy and Basic Materials (e.g., McKinsey 2015, 2 n.1). We use the Fama-French 12-industry classification as a balance between McKinsey’s seven industries on one end of the spectrum and the 83 2-digit Standard Industrial Classification (SIC) industry groups on the other end of the spectrum (setting aside the 416 3-digit SIC industry groups and 1,004 4-digit SIC industry groups).

14. There are three exceptions to McKinsey’s focus on EBIT margin. First, in their 2018 and 2020 studies, McKinsey reports using average ROE for financial companies in place of EBIT margin. Second, in their
in their 2015 study, 2011–2015 in their 2018 study, and 2014–2018 in their 2020 study. We follow McKinsey by making EBIT margin our primary measure of firm performance, and defining annual financial data over 2015–2019 and industries according to Fama-French’s 12-industry classification. However, as part of our subjecting McKinsey’s approaches to critique and stress testing, we also compute and evaluate five other measures of raw and industry median-adjusted firm financial performance: revenue growth, gross margin, ROA, ROE, and TSR. In panel C of Table 2 we report the 5th, 50th, and 95th percentiles of each performance measure, noting that while industry median–adjusted firm financial performance is zero for all six measures, there is substantial variation across S&P 500® firms within each measure.

Table 3. Descriptive statistics on the key non–race/ethnicity characteristics of the named executives with a face photo as of mid-2020 in the set of firms in the S&P® 500

Panel A. Number of executives per S&P® 500 firm

<table>
<thead>
<tr>
<th># executives per firm</th>
<th># execs</th>
<th>Min.</th>
<th>25%</th>
<th>Mean</th>
<th>75%</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7,246</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel B. Executive gender

<table>
<thead>
<tr>
<th>Male</th>
<th>Female</th>
<th>Total</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,533</td>
<td>1,713</td>
<td>7,246</td>
<td>76% 24% 100%</td>
</tr>
</tbody>
</table>

Panel C. Executives occupying Chief and Officer positions and executive presidential rank

<table>
<thead>
<tr>
<th>Chief or Officer (outright or Co-) position</th>
<th>C-Label</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEO</td>
<td>CEO</td>
<td>501</td>
</tr>
<tr>
<td>President</td>
<td>Pres</td>
<td>351</td>
</tr>
<tr>
<td>Chief Financial Officer</td>
<td>CFO</td>
<td>491</td>
</tr>
<tr>
<td>General Counsel or Chief Legal Officer</td>
<td>GC, CLO</td>
<td>452</td>
</tr>
<tr>
<td>Chief Operating Officer</td>
<td>COO</td>
<td>170</td>
</tr>
<tr>
<td>Corporate Secretary</td>
<td>CS</td>
<td>242</td>
</tr>
<tr>
<td>Chief Human Resources (or People) Officer</td>
<td>CHRO</td>
<td>228</td>
</tr>
<tr>
<td>Chief Information Officer</td>
<td>CIO</td>
<td>143</td>
</tr>
<tr>
<td>Chief Technology Officer</td>
<td>CTO</td>
<td>113</td>
</tr>
<tr>
<td>Chief Marketing Officer</td>
<td>CMO</td>
<td>87</td>
</tr>
<tr>
<td>Chief Accounting Officer</td>
<td>CACO</td>
<td>84</td>
</tr>
<tr>
<td>Chief Diversity/Equity/Inclusion Officer</td>
<td>CDEIO</td>
<td>19</td>
</tr>
<tr>
<td>Senior Executive Vice-President</td>
<td>SEVP</td>
<td>65</td>
</tr>
<tr>
<td>Executive Vice-President</td>
<td>EVP</td>
<td>1,686</td>
</tr>
<tr>
<td>Senior Vice-President</td>
<td>SVP</td>
<td>1,676</td>
</tr>
<tr>
<td>Vice-President</td>
<td>VP</td>
<td>1,162</td>
</tr>
</tbody>
</table>

Table 3 reports descriptive statistics for selected non-race/ethnicity characteristics of the 7,246 executives we identified in S&P 500® firms. Panel A indicates
that S&P 500® firms have on average 14.6 executives, and panel B shows that 76 percent (24 percent) of executives are male (female). Panel C presents the frequencies of different chief and officer-level positions. The most common executive positions are CEO, CFO, General Counsel, President, Corporate Secretary (often the same person as the General Counsel), Chief HR Officer, and COO. In terms of seniority, the most senior VP level (Senior EVPs plus EVPs) slightly outnumbers Senior VPs, who in turn outnumber VPs.

Executive judged race/ethnicity

In judging an executive’s race/ethnicity, we follow McKinsey (2015) by visually studying each executive’s photo and first and last names and classifying the executive into one of eight categories: African ancestry (aa), European ancestry (eur), Near Eastern (ne), East Asian (ea), South Asian (sa), Latino (lat), Native American (na), and Other (o). All race/ethnicity judgments were made by one coauthor. Because we stress test McKinsey’s results in part by ascertaining the effects of shrinking the number of racial/ethnic categories, we also separately place executives into the five race/ethnicity categories used by the National Center for Educational Statistics’ Integrated Postsecondary Education Data System (IPEDS). The five IPEDS race/ethnicity categories are American Indian/Alaska Native (aian), Asian/Pacific Islander (api), Black (b), Hispanic (h) and White (w), where given the allocation of Other (o) into Pacific Islander (pi) and Alaska Native (an), we set aian = na + an, api = ea + sa + pi, b = aa, h = lat, and w = eur + ne. The categories match closely with the race/ethnicity groups that McKinsey uses for the US firms in their 2018 and 2020 studies.

These methods enable us to judge the race/ethnicity of 6,931 of the 7,246 S&P 500® executives we identified as being in place in mid-2020. The top portion of Table 4 classifies executives by the eight racial/ethnic categories in McKinsey (2015), while the bottom portion classifies executives by the five IPEDS race/ethnicity categories. Of executives, 0.01 percent are American Indian or Alaska Natives, 2.8 percent are East Asian and 4.4 percent are South Asian (total Asian/Pacific Islander is 7.2 percent), 3.5 percent are African ancestry/Black non-Hispanic, 2.1 percent are Latino/Hispanic, and 1.4 percent are Near Eastern and 85.8 percent are European ancestry (total White non-Hispanic is 87.2 percent). While no approach outside of self-reported identification by each executive would achieve perfect accuracy, and no large database is likely to be completely accurate (and we make no representation of such), we undertook several cross-checks to do our best to obtain accurate assessments of executive race/ethnicity.15

15. These included the qualified assessment of an expert who is fluent in Spanish and deeply involved
Table 4. Numbers and densities of executives in S&P 500 firms at 12/31/19 classified into two sets of racial/ethnic (RAETH) categories

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>All Executives</td>
<td>0.00%</td>
<td>0.01%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>President</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>CFO</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>GC or CLO</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>COO</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Corporate Secretary</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>CHRO</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>CSO</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>CMO</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>Chief Accounting Officer</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>CDO/CHO/CTO/OEO</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>SEVP or EVP</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
<tr>
<td>EVP</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.38%</td>
<td>4.3%</td>
<td>3.5%</td>
<td>2.1%</td>
<td>1.5%</td>
<td>5.8%</td>
</tr>
</tbody>
</table>

Notes: [1] Expanding on McKinsey (2015), we classified an executive’s RAETH into one of nine categories by visually examining their photo and first and last names. All classifications were done by the same coauthor. The categories are as follows (RAETH tag in parentheses): African ancestry (aa), European ancestry (eur), Near Eastern (ne), East Asian (ea), South Asian (sa), Latino (lat), Native American (na), Pacific Islander (pi) and Alaska Native (an). Following McKinsey (2015, 2), we then combined pi and an into the category Other (o) to arrive at McKinsey’s (2015) eight RAETH categories. [2] We also created five RAETH supracategories to parallel those used in much of the historical data in the National Center for Educational Statistics’ Integrated Postsecondary Education Data System (NCES IPEDS). With our tag for each, these categories (RAETH supracategory tag in parentheses) are American Indian/Alaska Native (aian), Asian/Pacific Islander (api), Black (b), Hispanic (h) and White (w), where aian = ai + an, api = ea + sa + pi, b = aa, h = lat, and w = eur + ne. The five IPEDS’ RAETH supracategories closely match those used for US executives in McKinsey (2018; 2020).
McKinsey’s executive racial/ethnic diversity metrics

McKinsey measures the racial/ethnic diversity of a firm’s executives using an inverse normalized version of the Herfindahl-Hirschman Index (HHI) that they apply to the executives in the US + international sets of firms that they identified as being in place in 2014, 2017, and 2019. The Herfindahl-Hirschman Index is a standard measure of market concentration used to determine market competitiveness (link), such as before vs. after a merger or acquisition. Let \( i = 1 \) to \( N \) be mutually exclusive racial/ethnic groups into which an executive may be classified, and for any firm \( j \) let \( n_{ij} \) be the number of firm \( j \)'s executives that are classified in racial/ethnic group \( i \). Further letting the racial/ethnic density of racial/ethnic group \( i \) in firm \( j \) be given by \( RAED_{ij} = \frac{n_{ij}}{\sum_{i=1}^{N} n_{ij}} \), McKinsey defines \( HHI_j \) as:

\[
HHI_j = \sum_{i=1}^{N} RAED_{ij}^2
\]  

(1)

McKinsey then defines racial/ethnic diversity \( NHHI_j \) for firm \( j \) on a normalized and inverse basis:

\[
NHHI_j = 1 - \frac{HHI_j - N^{-1}}{1 - N^{-1}}
\]  

(2)

We follow McKinsey (2015) by using \( N = 8 \) racial/ethnic groups in our main tests, leaving the less differentiated \( N = 5 \) racial/ethnic groups for robustness tests.\(^{17}\) We denote \( NHHI \) when calculated using \( N = 8 \) as \( DIV_{McK8} \), and using \( N = 5 \) as \( DIV_{McK5} \).

---

16. McKinsey defines \( NHHI \) per equation (2) in their 2018 and 2020 studies (2018, 37; 2020, 49). In their 2015 study, McKinsey defines \( NHHI_j = \frac{HHI_j - N^{-1}}{1 - N^{-1}} \), that is, without applying an inverse by subtracting from one. McKinsey applies an inversion in their 2018 and 2020 studies in order that, per intuition, \( NHHI = 0 \) indicates a firm whose executives are all in the same racial/ethnic group, and \( NHHI = 1 \) indicates that firm’s executives are exactly equally spread out across the \( N \) racial/ethnic groups \( s_{ij} = \frac{N}{N-1} \forall i \). The result of this inversion is that \( NHHI \) in equation (2) is increasing in McKinsey’s definition of the degree of racial/ethnic diversity in a firm’s executives.

17. In their first study, McKinsey (2015, 2) states that \( N = 7 \), but in a footnote on the same page it says that “ethnic and racial categories used were African ancestry, European ancestry, Near Eastern, East Asian, South Asian, Latino, Native American, other” (McKinsey 2015, 2 n.2), which suggests that \( N = 8 \). In their later studies, McKinsey uses \( N = 5 \) for US firms (White/European ancestry, Black/African ancestry, Latino/Hispanic of any race, Asian/Asian ancestry including South Asian, and Other including mixed race; see McKinsey 2018, 37; McKinsey 2020, 49).
Table 5. Relations between McKinsey’s inverse normalized Herfindahl-Hirschman metrics of the racial/ethnic diversity of a firm’s executives DIV_McK8 and DIV_McK5 and the firm’s average annual financial performance FP in McKinsey’s 2015, 2018, and 2020 studies

Panel A. McKinsey’s normalized Herfindahl-Hirschman measures of the racial/ethnic diversity of S&P 500® firms’ executives (DIV_McK8 and DIV_McK5)

<table>
<thead>
<tr>
<th></th>
<th>DIV_McK8</th>
<th>DIV_McK5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50%</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>75%</td>
<td>0.39</td>
<td>0.40</td>
</tr>
<tr>
<td>Max</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td># obs</td>
<td>497</td>
<td>497</td>
</tr>
</tbody>
</table>

Notes: FP is measured by EBIT margin. Above-median financial outperformance (AMFP) is defined as a firm’s mean FP over the benchmark period used in the study less the median FP for the McKinsey-defined industry over the same period. $p(\text{AMFP}) = 1$ if AMFP > 0, else $p(\text{AMFP}) = 0$. Executive race/ethnicity judgments were made by McKinsey researchers during the year following the last year used in calculating annual FP. DIV_McK8 uses the eight different racial/ethnic groups delineated in McKinsey (2015, 2), while DIV_McK5 uses the five racial/ethnic groups delineated in McKinsey (2018, 37; 2020, 49). We assume that each quartile comprises the same number of firms (rounded down to the nearest one firm). Source: McKinsey (2020, 14; 2015, 6).

Panel B. Relations between McKinsey’s inverse normalized Herfindahl-Hirschman metrics of the racial/ethnic diversity of a firm’s executives DIV_McK8 and DIV_McK5 and the firm’s average annual financial performance FP in McKinsey (2015; 2018; 2020)

In panel A of Table 5 we report descriptive statistics on DIV_McK8 and DIV_McK5 for firms in the S&P 500® at 12/31/19. The two distributions are similar, with an average of 27 percent of firms having executives entirely of one
race/ethnicity, and a standard deviation for non-zero observations of 0.15. No firm has a McKinsey diversity score equal to 1.0, indicating that no firm in the S&P 500® had an equal number of the eight narrow or five broader races/ethnicities required under DIV_McK8 and DIV_McK5 for maximum diversity as defined by McKinsey.

**McKinsey’s approach to measuring, analyzing, and evaluating the relations between executive racial/ethnic diversity and firm financial performance**

McKinsey measures, analyzes, and evaluates the relations between the racial/ethnic diversity of firms’ executive teams and firms’ financial performance according to this sequence:

1. Rank the pertinent firms (e.g., all firms, or only US firms) by their $NHHI_j$ as defined in equation (2), with firms in the bottom executive race/ethnicity diversity quartile Q1 being those with the lowest $NHHI_j$ and firms in the top executive race/ethnicity diversity quartile Q4 being those with the highest $NHHI_j$.

2. In each of Q1 and Q4, calculate the “likelihood of financial outperformance” defined by McKinsey as the proportion of firms in a given quartile that have an EBIT margin above their national industry median EBIT margin.

3. Report the 2-tailed $p$-value on the $z$-statistic testing the null hypothesis that the difference in the likelihood of financial outperformance in Q4 versus Q1 is zero.\(^{19}\)

4. Present the percentage by which the likelihood of financial outperformance in Q4 exceeds the likelihood of financial outperformance in Q1. As an example, in Exhibit 1 of McKinsey’s 2018 study the likelihoods of financial outperformance are shown as 44 percent for Q1 and 59 percent for Q4. While McKinsey reports a $p$-value < 0.05 on the difference of $15\% = 59\% - 44\%$, what they emphasize is not the $15\%$, but the $33\%$ that is $15\% + 44\%$ (McKinsey 2018, 8).

---

\(^{19}\) McKinsey does not report what type of statistic underlies their inferences. We assume that McKinsey is calculating a $z$-statistic testing a difference in proportions, namely in this situation the difference in the likelihood of financial outperformance in Q4 versus Q1. McKinsey does not report whether their $p$-values are one-tailed or two-tailed. Based on our calculations using their Exhibit 7 results (McKinsey 2020, 20), we believe McKinsey’s $p$-values are two-tailed.
Results

McKinsey’s results as reported in their 2015, 2018, and 2020 studies

In each of their 2015, 2018, and 2020 studies, McKinsey reports finding a statistically significant positive relation between the industry-adjusted EBIT margin in their global samples of large public firms and the racial/ethnic diversity of the firms’ executive teams. In panel B of Table 5, we tabulate McKinsey’s results on financial outperformance, which we denote as above-median financial performance (AMFP), and the likelihood of financial outperformance, which we denote p(AMFP), in top NHHI-ranked quartile Q4 firms versus bottom NHHI-ranked quartile Q1 firms, based on Exhibit 6 of McKinsey’s 2015 study and Exhibit 7 of McKinsey’s 2020 study.

Columns 2–4 of panel B of Table 5 report McKinsey’s results for the global samples of large public firms used in their studies. For example, in their 2015 study, McKinsey reports that in their full set of global firms, 58 percent of Q4 firms had above-median financial outperformance compared to 43 percent of Q1 firms. We report the 15 percent difference between Q4 and Q1; the \( z \)-statistic of 2.0 on the difference of 15 percent, assuming there are \( n = 91 \) firms in each of Q1 and Q4; and the 2-tailed \( p \)-value of 0.04 on the \( z \)-statistic of 2.0. In column 5 we also report the results that McKinsey presents for just the US in their 2015 study, since our interest is in US S&P 500® firms.

In all four of columns 2–5, McKinsey’s results consistently indicate that for the firms in their samples of large public companies, there is a statistically significant higher likelihood of financial outperformance in top NHHI-ranked quartile firms than in bottom NHHI-ranked quartile firms. Specifically, the \( z \)-statistics on the differences between the means of p(AMFP) in Q4 versus Q1 are 2.0, 2.6, and 2.6 in McKinsey’s 2015, 2018, and 2020 studies that use all McKinsey’s pooled samples of global public firms (number of observations = 366, 589, and 533, and \( p \)-values = 0.04, 0.01, and 0.01, respectively), and 2.0 in McKinsey’s 2015 study that uses only McKinsey’s sample of US plus Canadian public firms (number of observations = 186, \( p \)-value = 0.05).

20. In Tables 5 and 6, means and \( z \)-statistics are reported to the nearest first decimal place and \( p \)-values to the nearest second decimal place. Because, however, the underlying calculations are based on exact figures without any rounding, the reported means, \( z \)-statistics, and \( p \)-values may differ slightly from their actual calculated values.
Our results for the firms in the S&P 500® Index at 12/31/19

In panel A of Table 6 we present the results of applying McKinsey’s approach to the firms in the S&P 500® Index at 12/31/19, and in panels B and C we report the results of expanding our analysis beyond McKinsey’s in several ways. The key takeaway from Table 6 is that, in contrast to McKinsey’s results, we do not find a statistically significant positive correlation between McKinsey’s measures of the racial/ethnic diversity of the executive teams of firms in the S&P 500® Index at 12/31/19 and either the likelihood of financial outperformance over 2015–2019 or financial outperformance per se over 2015–2019.

Table 6. Comparison of the results reported by McKinsey in its 2015 study and our quasi-replication using S&P 500® firms as of 12/31/19, McKinsey’s inverse normalized Herfindahl-Hirschman measures of the racial/ethnic diversity of the S&P 500® firms’ executives (DIV_McK8 and DIV_McK5), and S&P 500® firms’ annual financial performance FP over 2015–2019, measured in six ways: EBIT margin as % of revenues, revenue growth, gross margin as % of revenues, return on assets (ROA), return on equity (ROE), and total shareholder return (TSR).

Panel A. Comparison of McKinsey’s 2015 study results (left-hand side, green) with our results for S&P 500® firms based on the use of eight racial/ethnic categories in calculating McKinsey’s inverse normalized Herfindahl-Hirschman diversity metric DIV_McK8 (middle, dark blue) and the use of five racial/ethnic categories in calculating McKinsey’s inverse normalized Herfindahl-Hirschman diversity metric DIV_McK5 (right-hand side, pink)
Table 6 (cont’d). Panel B. Full results for S&P 500® firms covering multiple measures of firm financial performance, based on the use of eight racial/ethnic categories in calculating McKinsey’s inverse normalized Herfindahl-Hirschman diversity metric DIV_McK8

<table>
<thead>
<tr>
<th>Statistic used to assess average FP in a given decade</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
<th>Column 5</th>
<th>Column 6</th>
<th>Column 7</th>
<th>Column 8</th>
<th>Average across all 6 FP measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean p(MAMP)</td>
<td>Mean AMP</td>
<td>Mean p(MAMP)</td>
<td>Mean AMP</td>
<td>Mean p(MAMP)</td>
<td>Mean AMP</td>
<td>Mean p(MAMP)</td>
<td>Mean AMP</td>
<td>Mean p(MAMP)</td>
</tr>
<tr>
<td>Q1 = lowest exec diversity quartile</td>
<td>51.2%</td>
<td>0.8%</td>
<td>52.0%</td>
<td>1.3%</td>
<td>44.1%</td>
<td>1.1%</td>
<td>52.8%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Q2</td>
<td>45.5%</td>
<td>1.1%</td>
<td>55.3%</td>
<td>2.1%</td>
<td>52.8%</td>
<td>2.7%</td>
<td>48.8%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Q3</td>
<td>49.6%</td>
<td>0.4%</td>
<td>56.9%</td>
<td>1.0%</td>
<td>43.1%</td>
<td>1.1%</td>
<td>43.1%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Q4 = highest exec diversity quartile</td>
<td>54.0%</td>
<td>1.9%</td>
<td>43.5%</td>
<td>0.1%</td>
<td>54.0%</td>
<td>2.1%</td>
<td>52.8%</td>
<td>1.2%</td>
</tr>
<tr>
<td>Q4 - Q1 (exact is rounded to 1 dp)</td>
<td>2.9%</td>
<td>1.1%</td>
<td>1.4%</td>
<td>1.2%</td>
<td>9.9%</td>
<td>1.1%</td>
<td>0.3%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Pearson_correlation(div_McK8, FP)</td>
<td>0.02</td>
<td>0.02</td>
<td>-0.07</td>
<td>-0.07</td>
<td>0.07</td>
<td>0.00</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>t-stat( Pearson_correlation )</td>
<td>0.5</td>
<td>0.5</td>
<td>1.5</td>
<td>1.5</td>
<td>1.6</td>
<td>0.0</td>
<td>-0.5</td>
<td>-0.4</td>
</tr>
</tbody>
</table>

Panel C. Full results for S&P 500® firms covering multiple measures of firm financial performance, based on the use of five racial/ethnic categories in calculating McKinsey’s inverse normalized Herfindahl-Hirschman diversity metric DIV_McK5

<table>
<thead>
<tr>
<th>Statistic used to assess average FP in a given decade</th>
<th>Column 2</th>
<th>Column 3</th>
<th>Column 4</th>
<th>Column 5</th>
<th>Column 6</th>
<th>Column 7</th>
<th>Column 8</th>
<th>Average across all 6 FP measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean p(MAMP)</td>
<td>Mean AMP</td>
<td>Mean p(MAMP)</td>
<td>Mean AMP</td>
<td>Mean p(MAMP)</td>
<td>Mean AMP</td>
<td>Mean p(MAMP)</td>
<td>Mean AMP</td>
<td>Mean p(MAMP)</td>
</tr>
<tr>
<td>Q1 = lowest exec diversity quartile</td>
<td>52.5%</td>
<td>1.4%</td>
<td>52.5%</td>
<td>1.5%</td>
<td>47.5%</td>
<td>2.3%</td>
<td>51.8%</td>
<td>1.4%</td>
</tr>
<tr>
<td>Q2</td>
<td>48.9%</td>
<td>1.2%</td>
<td>55.9%</td>
<td>1.8%</td>
<td>52.5%</td>
<td>2.4%</td>
<td>50.8%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Q3</td>
<td>47.9%</td>
<td>0.1%</td>
<td>59.7%</td>
<td>1.5%</td>
<td>45.4%</td>
<td>0.1%</td>
<td>46.2%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Q4 - Q1 (exact is rounded to 1 dp)</td>
<td>-1.2%</td>
<td>0.1%</td>
<td>-1.0%</td>
<td>-1.8%</td>
<td>1.2%</td>
<td>2.3%</td>
<td>-3.9%</td>
<td>-0.7%</td>
</tr>
<tr>
<td>Pearson_correlation(div_McK5, FP)</td>
<td>-0.2</td>
<td>0.1</td>
<td>-2.1</td>
<td>-2.0</td>
<td>0.2</td>
<td>-3.0</td>
<td>-0.6</td>
<td>-1.0</td>
</tr>
<tr>
<td>t-stat( Pearson_correlation )</td>
<td>-0.2</td>
<td>-0.2</td>
<td>-1.8</td>
<td>-1.6</td>
<td>0.5</td>
<td>-0.9</td>
<td>0.7</td>
<td>-0.8</td>
</tr>
</tbody>
</table>

Notes: Above-median financial outperformance (AMFP) is defined as a firm’s mean FP over the period 2015–2019 less the median Fama-French 12-industry FP over 2015–2019. The likelihood of above-median financial performance = p(AMFP) is set to 1 if AMFP > 0, else p(AMFP) = 0. Executive race/ethnicity judgments were made by the same coauthor during May–August 2020 as described in the section “Executive judged race/ethnicity.” DIV_McK8 uses the eight different racial/ethnic groups delineated in McKinsey (2015, 2), while DIV_McK5 uses the five racial/ethnic groups delineated in McKinsey (2018, 37, 2020, 49).
Our results in Table 6 that are most directly comparable to McKinsey’s are shown in panel A. On the left-hand side of panel A we show McKinsey’s 2015 results per the far right-hand column of Table 5, panel B. In the middle and right-hand side of Table 6, panel A, we present for comparison our results based on the use of McKinsey’s executive racial/ethnic diversity metrics DIV_McK8 and DIV_McK5, respectively. We propose that these results are comparable because each set of firms consists of large US public companies (n = 186 McKinsey US firms vs. n = 497 S&P 500® firms) for which financial performance is measured reasonably after the Great Recession (2010–2013 for McKinsey’s US firms vs. 2015–2019 for our S&P 500® firms). However, in sharp contrast to McKinsey’s results in the left-hand side of Table 6, panel A, in the middle of this panel we show that only 54.0 percent of the S&P 500® firms that are in the top quartile Q4 of McKinsey’s 2015 executive racial/ethnic diversity metric DIV_McK8 have a positive industry-adjusted EBIT margin, vs. 51.2 percent in the bottom DIV_McK8 quartile Q1.21 The z-statistic on the difference of 2.8 percent is not statistically significant (p-value = 0.65).

Also in sharp contrast to McKinsey’s results in the left-hand side of panel A of Table 6, in the right-hand side of the panel we find that only 51.3 percent of the S&P 500® firms that are in the top quartile Q4 of McKinsey’s 2015 executive racial/ethnic diversity metric DIV_McK5 have a positive industry-adjusted EBIT margin, vs. 52.5 percent in the bottom DIV_McK5 quartile Q1. The z-statistic on the difference of −1.2 percent is a not statistically significant −0.2.

Because the results of our quasi-replication contradict those reported by McKinsey in their 2015, 2018, and 2020 studies, we expand our tests in five ways to assess the robustness of our inference. First, in Table 6, panel B, column 2, we report the mean p(AMFP) for DIV_McK8 quartiles Q2 and Q3. If greater executive racial/ethnic diversity is positively associated with industry-adjusted EBIT margin, it should be that the mean p(AMFP) in Q1 < mean p(AMFP) Q2 < mean p(AMFP) Q3 < mean p(AMFP) Q4. However, the mean p(AMFP) figures for Q2 and Q3 do not support this prediction. Thus, the z-statistic on the difference of −5.7 percent between the mean p(AMFP) in Q2 and Q1 is −0.9 (p-value = 0.37). Second, we extend beyond the probability of financial outperformance per se and compare the mean levels of industry-adjusted EBIT margin in Q4 vs. Q1. Here too, however, we find a not statistically significant difference: the mean industry-adjusted EBIT margin for Q4 is 1.9 percent vs. 0.8 percent in Q1, and the t-statistic on the 1.1 percent difference in means is 0.9 (p-value = 0.37). Third, using the data in all four quartiles but in a continuous and unranked manner,

---

21. All n = 127 S&P 500® firms with zero executive racial/ethnic diversity are in the bottom quartile, and the n = 124 firms with the highest executive racial/ethnic diversity are in the top quartile.
we find that executive racial/ethnic diversity is uncorrelated with the likelihood that a firm’s industry-adjusted EBIT margin is positive (Pearson correlation coefficient = 0.02, \( t \)-statistic = 0.5), and with the firm’s industry-adjusted EBIT margin (Pearson correlation = 0.02, \( t \)-statistic = 0.5).

Fourth, since in their 2018 and 2020 studies McKinsey use five rather than eight racial/ethnic groups to measure US executive racial/ethnic diversity, in column 2 of Table 6, panel C, we repeat our tests from column 2 of Table 6, panel B, using \( DIV\_ McK5 \) instead of \( DIV\_ McK8 \). We find \( z \)-statistics and correlations using \( DIV\_ McK5 \) that are even closer to zero than those found using eight racial/ethnic groups. Fifth, in panels A and B of Figure 3 we present the scatterplots and univariate regression lines for \( DIV\_ McK8 \) (y-axis) vs. \( AMPF \) (x-axis), and \( DIV\_ McK5 \) vs. \( AMPF \), respectively. The scatterplots do not reveal any outliers.

**Figure 3.** Scatterplots of the relations between McKinsey’s inverse normalized Herfindahl-Hirschman measures of the racial/ethnic diversity of S&P 500® firms’ executives \( DIV\_ McK8 \) (panel A) and \( DIV\_ McK5 \) (panel B) and above-median financial outperformance (\( AMPF \)) defined as average EBIT margin over 2015–2019 less the median Fama-French 12-industry firm performance over 2015–2019.

**Panel A.** Above-median financial outperformance (\( AMPF \), y-axis) vs. Executive diversity using eight races/ethnicities (\( DIV\_ McK8 \), x-axis), with univariate OLS regression line.
and the univariate regression lines appear visually sound and robust. Sixth, in columns 3–7 of Table 6, panels B and C, we examine five additional measures of firm financial performance: sales growth, gross margin, ROA, ROE, and TSR, all on an industry-adjusted basis. For each, we repeat the tests done for industry-adjusted EBIT margin. This yields 40 non-independent \( z \)-statistics or \( t \)-statistics testing the null hypothesis that there is no relation between firm financial performance and McKinsey’s metric for executive racial/ethnic diversity in US S&P 500\(^\text{®} \) firms. We find that 37 of the 40 test statistics are insignificant, one is reliably positive, and two reliably negative.

Lastly, in column 8 of panels B and C we calculate the simple average of the corresponding cells in columns 2–7. Once again, we find no evidence of any statistically significant positive relations between the financial performance of S&P 500\(^\text{®} \) firms and McKinsey’s measures of the racial/ethnic diversity of their executives.

The totality of the results we report in Table 6 suggest that despite the imprimatur given to McKinsey’s (2015; 2018; 2020) studies, substantial caution is warranted in relying on their findings to support the view that US publicly traded firms can deliver improved financial performance if they increase the racial/ethnic diversity of their executives.
diversity of their executives. Hewing closely to McKinsey’s approach using a sample of large US public firms, in our quasi-replication we do not find evidence that is consistent with McKinsey’s results for firms that were in the US S&P 500® at 12/31/19, using average annual financial performance over 2015–2019 and executive race/ethnicity measured in mid-2020.

**Discussion**

In this section we provide additional critiques of McKinsey’s studies. We first appraise McKinsey’s inverse normalized Herfindahl-Hirschman measure of racial/ethnic diversity, highlighting its strengths and weaknesses. We then discuss McKinsey’s views of what the positive correlations it reports between executive racial/ethnic diversity and firm financial performance say about causality. Lastly, we report on work that has sought to carefully identify and measure the presence, sign, and magnitude of causal relations between executive racial/ethnic diversity and firm financial performance.

**Strengths and weaknesses of McKinsey’s measure of executive team diversity**

Despite its careful, albeit varied, delineation in academic research (e.g., Williams and O’Reilly 1998; Harrison and Klein 2007; Lu, Naik, and Teo 2024), the word *diversity* is rarely defined in a careful manner in either the business or common vernacular. In contrast, a strength of McKinsey’s studies is that McKinsey clearly and algebraically defines their diversity measure in all three of their 2015, 2018, and 2020 reports. This feature notwithstanding, McKinsey’s *HNNI* inverse normalized Herfindahl-Hirschman definition of executive racial/ethnic diversity has three weaknesses.

First, McKinsey’s *HNNI* diversity metric is maximized when there are equal numbers or densities of executives from all *N* races/ethnicities in a given firm. This is problematic because neither the US population nor the US labor force contains equal numbers of each race/ethnicity. From a real-world point of view, McKinsey’s measure of executive racial/ethnic diversity can therefore likely only be at its maximum in a subset of US firms, not in all US firms.

Second, McKinsey’s measure of executive racial/ethnic diversity yields the result that any set of executive racial/ethnic densities (RAEDs) that differs from equal densities is less diverse than one with equal densities. We suggest that this runs counter to the intuition that a firm whose executive RAEDs are equal to those of the US population is more racially/ethnically diverse than a firm whose
executive RAEDs are equal across all races/ethnicities.\textsuperscript{22}

Third, McKinsey’s diversity metric yields what we propose is the counterintuitive outcome that firm ABC that has the same executive racial/ethnic densities of the US population is as equally diverse as firm XYZ that has the same race/ethnicity densities as the US population except that the race/ethnicity densities are spread out “oppositely” or in some other way different from those of the US population. For example, the 2019 US population RAEDs are American Indian/Alaska Native = 1.0%, Asian/Pacific Islander = 6.4%, Black = 13.0%, Hispanic = 18.5%, and White = 61.2% (Green and Hand 2021, appendix C). One can readily calculate that $HNNI_{\text{(aian, api, b, h, w)}} = HNNI_{(1.0\%, 6.4\%, 13.0\%, 18.5\%, 61.2\%)} = 0.77 = HNNI_{(61.2\%, 18.5\%, 13.0\%, 6.4\%, 1.0\%)} = HNNI_{(6.4\%, 18.5\%, 61.2\%, 13.0\%, 1.0\%)}. However, we propose that it is unlikely that most business leaders and employees will view a firm whose executive team is 61.2% American Indian/Alaska Native, 18.5% Asian/Pacific Islander, 13.0% Black, 6.4% Hispanic and 1.0% White (exactly the inverse of the 2019 US population RAED percentages) to be as equally racially/ethnically diverse as a firm whose executive team is 1.0% American Indian/Alaska Native, 6.4% Asian/Pacific Islander, 13.0% Black, 18.5% Hispanic and 61.2% White (the 2019 US population RAEDs).

\textbf{Causality: Testing for the presence, sign, and magnitude of causal relations between executive racial/ethnic diversity and firm financial performance}

Given the business and societal importance of determining whether greater racial/ethnic diversity in corporate executives does or does not cause or drive higher firm financial performance in a statistically reliable way, we believe it is crucial to highlight that even if our results had agreed with McKinsey’s, McKinsey’s interpretation of their results, namely that US publicly traded firms can deliver improved financial performance if they increase the racial/ethnic diversity of their executives, is flawed because their tests are structured so as to evaluate the exact opposite direction of causality, namely that higher firm financial performance leads to greater executive racial/ethnic diversity. McKinsey measures firm financial performance over the four or five years leading up to the year in which they measure the race/ethnicity of the firm’s executives, making the default direction.

\textsuperscript{22} While executive RAEDs that are equal to the US population are more representative of the US population, they might still be seen by some as less diverse than RAEDs that are equal across all N races/ethnicities. Such ambiguities and differences are unfortunately inevitable when there is not a uniquely accepted definition of racial/ethnic diversity.
of causality captured in their correlations that of better firm financial performance causing companies to diversify the racial/ethnic composition of their executives, not the reverse. In this regard, we make three points.

First, McKinsey notes the reverse causal nature of their tests. In all three of their studies, McKinsey states that their positive relation between executive racial/ethnic diversity and EBIT margin is a correlation, and not a causal link that shows that higher racial/ethnic diversity of executives causes higher firm financial performance. McKinsey also notes that better firm financial performance may lead companies to diversify as defined by their measures of diversity.

Second, despite this, McKinsey has not emphasized the nature of their tests in their public statements concerning their 2015, 2018, 2020, and 2023 studies. They have also not sought to estimate the directionally correct causal relations between their measures of diversity in executive race/ethnicity and firm financial performance, instead arguing, “As with many levers of business performance, particularly at such a high level, this [a causal link] would be challenging to demonstrate, likely requiring detailed longitudinal studies” (2020, 51).

Third, the longitudinal and causally oriented analyses that McKinsey has not done have nevertheless been done, in our work with Sekou Bermiss, for S&P 500® firms (Bermiss, Green, and Hand 2024). That study gathered data on the race/ethnicity of the individuals defined to be executives of S&P 500® firms on the leadership pages of these firms’ websites as of mid-2011, 2014, 2017, 2020, and 2021. The study then empirically assessed whether any of nine different measures of executives’ racial/ethnic diversity reliably predict cross-sectional variation in any of six measures of firm financial performance over the next fiscal year and find that they do not, neither over the full 11-year span of the data, nor in the period of America’s ‘awakening’ to systemic racism after the George Floyd murder in 2020.

23. “The relationship between diversity and performance highlighted in the research is a correlation, not a causal link” (2015, 1); “correlation does not prove that the relationship is causal” (2015, 3); “correlation does not demonstrate causation” (2018, 2); “the same caveats apply to the correlation analyses reported here as did in Why Diversity Matters: correlation is not causation” (2018, 5). Very similar statements to those from McKinsey’s 2018 study are included in their 2020 study. At the same time, however, the titles of McKinsey’s studies imply in their wordings that they are taking a causal view of the evidence they present: thus, “Diversity Matters” (2015), “Delivering Through Diversity” (2018) and “Diversity Wins: How Inclusion Matters.”

24. “It is theoretically possible that the better financial outperformance enables companies to achieve greater levels of diversity. Companies that perform well financially may choose to deploy more of their resources toward more advanced talent strategies, thus allowing them to attract more diverse talent, for example” (2018, 39).

25. We approached McKinsey and asked if they would share their data with us so that we could undertake a longitudinal analysis of it. They declined, citing internal policies pertaining to not releasing data that would relate to clients. The severity of this stricture meant that McKinsey would not release to us even the names of the firms in their datasets.
Of the total of 270 estimated coefficients on the nine measures of executive racial/ethnic diversity across the six measures of 1-year-ahead firm financial performance over the years 2012, 2015, 2018, 2021, and 2022, the study finds that just under 4 percent are significantly non-zero at a 2-tailed level of a \( p \)-value = 0.05. As such, the results of Bermiss, Green, and Hand (2024) suggest that in contrast to the titles of McKinsey’s 2015 (“Diversity Matters”), 2018 (“Delivering through Diversity”), 2020 (“Diversity Wins: How Inclusion Matters”) and 2023 (“Diversity Matters Even More: The Case for Holistic Impact”) studies, greater racial/ethnic diversity in the executives of large US public companies does not “matter, deliver, or win,” in the specific sense that greater racial/ethnic diversity in the executives of large US public companies does not on average correlate in a statistically reliable way with higher one-year-ahead firm financial performance.

**Caveats**

As with any study, our research comes with several caveats. First, our sole focus is on US firms. We therefore make no comments regarding McKinsey’s findings on non-US firms. Second, S&P 500\(^\circ\) firms are not a random sample of US publicly traded firms. Our results should therefore not be assumed to automatically generalize to the population of US publicly traded firms. Third, because we do not undertake in-depth biographical analysis, our method of identifying executive race/ethnicity is likely to undercount non-Whites and overcount Whites, primarily because non-White individuals’ faces and/or names can sometimes appear similar to European faces and/or names, and vice-versa. While we do not believe this is likely to lead to biases in the inferences we make in our study, executive-specific information from List Service Direct Inc. (LSDI) could be used to augment our current face-plus-names approach to judging race/ethnicity. LSDI uses a person’s name(s) to estimate their race/ethnicity. The strength of LSDI’s approach is that it provides likely less-biased identification of Hispanics. However, the weakness of LSDI’s approach is that it focuses solely on name-based information and thus sets entirely aside the value of face-based information.\(^{26}\)

---

\(^{26}\) We do not adjust any other visually identified races/ethnicities using LSDI data for two reasons. First, in Green and Hand (2021) we used CEO and CFO data from Crist Kolder Associates that allowed us to cross-check the accuracy of the visual identification method. We found that the visual identification method, which we also use in this paper, identified API and Black executives in a fairly accurate manner. Second, because many Black and White names are not distinguishable, LSDI under-identifies (over-identifies) the number of Black (White) individuals (Brochet et al. 2019; Flam et al. 2023).
Conclusions

In a series of studies that are highly influential in the business world, McKinsey (2015; 2018; 2020; 2023) report finding statistically significant positive relations between the industry-adjusted EBIT margin of global samples of large public firms and the racial/ethnic diversity of their executives. However, when we conduct a quasi-replication of McKinsey’s tests using data for US S&P 500® firms as of 12/31/19, we find a not statistically significant relations between McKinsey’s measures of executive racial/ethnic diversity and not only industry-adjusted EBIT margin, but also industry-adjusted sales growth, gross margin, return on assets, return on equity, and total shareholder return.

Our findings lead us to two main conclusions and an emphasis. First, we conclude that caution is warranted in relying on McKinsey’s findings to support the view that US publicly traded firms can deliver improved financial performance if they increase the racial/ethnic diversity of their executives—not only because we are unable to replicate the same statistically reliable association between firm financial performance and executive race/ethnic diversity as they report, but also because the structure of McKinsey’s tests are such that by measuring firm financial performance over the four or five years leading up to the year in which they judge the race/ethnicity of firms’ executives, the default direction of causality that McKinsey capture in the positive correlation they report is that better firm financial performance causes firms to diversify the racial/ethnic composition of their executives, not the reverse.

Second, we conclude that in light of the prominence of the connections between firm financial performance and the racial/ethnic composition of their employees, not just in the US but around the world, there is great value in future research that would seek to empirically test for the presence, sign, magnitude, and direction of any causal relations that exist. Such longitudinal and causality-oriented studies may also help bring into sharper focus the identities and sizes of the costs and benefits, as well as the risks and returns, that are associated with higher or lower racial/ethnic diversity, not only in firms’ executives, but in their Boards of Directors and rank-and-file employees. In this regard, we believe that our own work, published in a separate paper to this (Bermis, Green, and Hand 2024), represents a useful beginning.

Lastly, we emphasize that in light of the challenging nature of matters to do with race/ethnicity in the US, our findings, like those of McKinsey, are limited. While our results do speak to the lack of robustness of McKinsey’s studies vis-à-vis large public US firms, they do not speak to the connections between racial/ethnic diversity in employees and/or boards and either firm financial performance or non-financial firm goals, nor to intrafirm activities. Nor do they speak to any
social or moral contributions that racial/ethnic diversity in US executives provides. Such research is worthwhile and important, and we hope that it will be undertaken and well so by business scholars.

## Data and code

While EJW has a policy of making data and code immediately available (as stated here), EJW has agreed to letting us depart from that policy here; we are not posting our data and code just yet because the data constitute an excerpt of research that is currently in progress. However, at this time, data and code may be available from the corresponding author (John Hand) upon reasonable request. Furthermore, we commit to posting the data and code by 31 December 2026.

## Appendix

This appendix presents screenshots of the raw firm and executive data items for two example firms in the S&P 500® dataset, along with an explanation of what each data item means, how it was collected, and how it was coded. Not all the data items shown in Figure A1 and Figure A2 are used in this study.

### Figure A1. Items 1–19

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Firm ID.</td>
</tr>
<tr>
<td>2.</td>
<td>Firm name per Compustat.</td>
</tr>
<tr>
<td>3.</td>
<td>Webpg 1 = First level in firm’s website address identifying the page with the executive on it.</td>
</tr>
<tr>
<td>4.</td>
<td>Webpg 2 = Second level in firm’s website address identifying the page with the executive on it.</td>
</tr>
<tr>
<td>5.</td>
<td>Webpg 3 = Third level in firm’s website address identifying the page with the executive on it.</td>
</tr>
</tbody>
</table>

**Item 1:** Firm ID.

**Item 2:** Firm name per Compustat.

**Item 3:** Webpg 1 = First level in firm’s website address identifying the page with the executive on it.

**Item 4:** Webpg 2 = Second level in firm’s website address identifying the page with the executive on it.

**Item 5:** Webpg 3 = Third level in firm’s website address identifying the page with the executive on it.
executive on it.

Item 6: Webpg 4 = Fourth level in firm’s website address identifying the page with the executive on it.

Item 7: Executive #, coded in the order shown on firm’s website (if in a row, order taken is left to right).

Item 8: Last name(s) of executive.

Item 9: First name(s) of executive.

Item 10: Middle initial(s) of executive.

Item 11: Chief or Officer 1 = First of a maximum of two Chief or Officer positions ascribed to the executive.

Item 12: Chief or Officer 2 = Second of a maximum of two Chief or Officer positions ascribed to the executive.

Item 13: Chief or Officer Domain = category covering one or more Chief or Officer 1 or 2 positions.

Item 14: Rank or Title = rank or title of executive, outside of Chief and Officer 1 or 2.

Item 15: Rank or Title Domain = category covering one or more Ranks or Titles.

Item 16: Area = area of business responsibility covered by the executive, as judged by the authors based on the text provided about the executive on firm’s website.

Item 17: Photo = y if a photo of the executive was found on the firm’s website, else the executive’s LinkedIn page (LIN), else the firm’s Bloomberg profile (BB), else business media (OTH).

Item 18: Photo source. If photo = y, photo source = firm’s website, LIN, BB or OTH.

Item 19: Gender: Male or female, based on the executive’s photo and/or bio, where available.

**Figure A2.** Items 20–32
Item 20: McK 2015 race/ethnicity: We classified an executive’s race/ethnicity by visually examining their photo and first and last names. All classifications were done by the same coauthor. The most granular racial/ethnic categories we employ are those of McKinsey (Hunt, Layton, and Prince, 2015). With our lowercase descriptor tag of each race/ethnicity category shown in parentheses, these are African ancestry (aa), European ancestry (eur), Near Eastern (ne), East Asian (ea), South Asian (sa), Latino (lat), Native American (na), and Other (o). We specify Other as either Pacific Islander (pi) or Alaska Native (an). We use the nomenclature American Indian (ai) rather than Native American because American Indian is the nomenclature used in much of the historical data that we extract from the National Center for Educational Statistics’ Integrated Postsecondary Education Data System (NCES IPEDS).

Item 21: NCES IPEDS race/ethnicity. NCES IPEDS specifies the following five race/ethnicity categories outside of Nonresident aliens (lowercase descriptor tag of each race/ethnicity category in parentheses): American Indian/Alaska Native (aian), Asian/Pacific Islander (api), Black (b), Hispanic (h), White (w). We connect McK 2015 race/ethnicity categories with the NCES IPEDS race/ethnicity categories by defining $b = aa$, $w = eur + ne$, $api = ea + sa + pi$, $h = lat$, $aian = na + an$ (see item 23 for McK category descriptor tags). NCES IPEDS’ race/ethnicity categories match closely with those used for US executives in McKinsey’s 2018 and 2020 studies.

Item 22: Visual est. age: Age of the executive as judged by the same coauthor from the executive’s photo, assigned into one of the following point estimates: 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90.

Item 23: Formal attire? = y if executive was wearing formal attire as judged from the executive’s photo by the same coauthor (sometimes not possible if photo was only of the executive’s face).

Item 24: Jacket? = y if executive was wearing a jacket as judged from their photo by the same coauthor (sometimes not possible if photo was only of the executive’s face).

Item 25: Tie? = y if executive was wearing a tie as judged from the executive’s photo by the same coauthor (sometimes not possible if photo was only of the executive’s face).

Item 26: Smile (1–10). Degree of genuine smile on the executive’s face as judged from the executive’s photo by the same coauthor, where 1 = not at all smiling/“very grumpy” and 10 = very wide, “joyous” smile.

Item 27: Pay ($M) Yahoo! Finance. If executive is one of the maximum of five individuals listed on the firm’s Yahoo! Finance Profile page, Pay is the amount of “salary, bonuses etc.” for the last fiscal year ending December 31, 2019.

Item 28: Year Born Yahoo! Finance. If executive is one of the maximum of five
individuals listed on the firm’s Yahoo! Finance Profile page, Year Born is the executive’s YYYY year of birth.

Item 29: True Age @ Feb-20. If Year Born is available, True Age @ Feb-20 is the age of the executive to the nearest one year as of February 2020.

References


Kim, Michelle Mijung. 2018. If Your Boss Is Still Asking About the “Business Case” for Diversity, Your Company’s in Trouble. Awaken Blog (Awaken, Oakland, Cal.),
March 26. Link


About the Authors

Jeremiah Green is an Associate Professor of Accounting and holds the Ernst & Young Professorship of Accounting at the Mays School of Business at Texas A&M University. He does research on executive race/ethnicity in US public companies, and capital markets research that focuses on the use of accounting information. Dr. Green also studies hedge funds, equity and debt analysts, auditors, managers, the business press, and equity trading strategies. His teaching centers on data analytics and analytics for financial reporting. His email address is jgreen@mays.tamu.edu.
John R. M. Hand is the Robert March & Mildred Borden Hanes Distinguished Professor of Accounting at the Kenan-Flagler Business School at UNC–Chapel Hill and a Visiting Professor of Accounting at The University of Chicago Booth School of Business. His research interests span many areas, most recently focusing on matters relating to executive race/ethnicity in US public companies. Dr. Hand has twice won the American Accounting Association’s competitive manuscript competition for his scholarship and is coauthor with Baruch Lev of New York University of Intangible Assets: Values, Measures and Risks (Oxford University Press, 2003). He teaches MBAs, MAC students, undergraduates, and executives about financial statement analysis & valuation. His email address is John_Hand@kenan-flagler.unc.edu.
Global Non-Linear Effect of Temperature on Economic Production: Comment on Burke, Hsiang, and Miguel

David Barker

The journal Nature published an influential article in 2015 by Marshall Burke, Solomon M. Hsiang, and Edward Miguel (hereafter BHM) purporting to show that higher temperatures will lower economic growth in warm countries. The headline result is that unrestrained global warming will reduce world GDP per capita by 23 percent in the year 2100, approximately nine times larger than the estimate of William Nordhaus (2018).

The Web of Science reports that the paper is in the top six one hundredths of one percent of economics and business publications by citations, and Google Scholar shows 2,269 citations. BHM (2015) also received significant attention in the popular press. Hsiang further developed this work and cowrote a chapter of the National Climate Assessment (Hsiang et al. 2023) claiming that higher temperatures would reduce the rate of economic growth.

BHM’s analysis is shallow and misleading. The authors use data with characteristics that are known to create spurious regression results without making proper adjustments or even acknowledging these characteristics. They estimate...
parameters of a quadratic curve relating temperature to growth, and then cherry-pick countries to include in a chart that appears to confirm the shape of this curve. The curve is then used to project growth rates into the distant future using temperature scenarios that a more recent comment in *Nature* described as either “extremely unlikely” or “unlikely” (Hausfather 2020). BHM say that results from the subsamples 1960–1989 and 1990–2010 are similar, indicating that the effects of higher temperatures are not being mitigated, but do not report that the results from subsamples 1960–1990 and 1991–2010 are different. The headline result, that warming will reduce global GDP per capita by 23 percent, is more than double the mean estimate of BHM’s bootstrap estimation, which they do not report. BHM claim that their result is “globally representative” (2015, 237), but it does not hold without Greenland and the regions of the Sahara and Central Africa, and it does not hold in large regions of the world.

Simulations support the hypothesis that spatial autocorrelation may be the cause of BHM’s results, and robustness checks also suggest that their results may be spurious. BHM has been the subject of methodological criticism (Newell et al. 2021; Tol 2019; Rosen 2019), but my paper is the first to precisely document its deceptive practices.

I begin with a review of previous criticism of BHM, then I describe BHM, and then I discuss problems with their analysis.

## Previous criticism of BHM

BHM has received thousands of positive citations and glowing coverage in the popular press. The paper has, however, faced some criticism. In a short letter in *PNAS*, Richard Rosen (2019) argues that using year dummies in a regression of economic growth is invalid because it ignores the many economic factors that affect growth. Rosen (2019) also argues that using a single equation to estimate optimal temperatures around the world is invalid because the economic conditions of countries are very different. Finally, he argues that equally weighting all countries distorts the results, because very small countries are given the same importance as very large countries. Rosen (2019) describes these issues in his one-page letter, but does not perform independent econometric analysis.

Richard Tol (2019) has written a footnote stating the following:

> The econometrics of Burke et al. (2015) do not stand up to scrutiny. They regress the difference of the log of per capita income, a stationary variable, on

---

4. BHM’s data on GDP per capita begin in 1960, so the first observation of GDP growth is in 1961.
temperature, a non-stationary variable, and year dummies. As a non-stationary variable cannot explain a stationary one, Burke’s year dummies must have de facto detrended temperature. This is indeed the case, as confirmed by their replication package. To the best of my knowledge, the statistical properties of regressing a stationary variable on a cointegrating vector are not known. However, that cointegration vector is measured with error. Errors-in-variables induce bias, of unknown sign in non-linear models (Griliches and Ringstad, 1970). Burke’s unusual procedure works fine in-sample, but goes off the rails out-of-sample (Newell et al.) as the year dummies cannot be predicted. (Tol 2019, 556 n.1)

Tol (2019) also does not perform independent econometric analysis of BHM. Richard Newell et al. (2021), as Tol (2019) describes, finds that BHM’s model does not predict well out of sample. They test the model’s ability to predict out of sample by dividing the data into different portions, estimating the model on one portion of data, and using the model to predict the other portions.

There have also been discussions of BHM on economics blogs. Tol blogged in 2015 that “The graphs are fantastic. The analysis is less impressive” (link). Tol also stated that “Burke and co fail to test for spatial autocorrelation, a feature of both economic growth and temperature. Their standard errors are thus downward biased.” Marshall Burke replied on his own blog that Tol’s criticism about “picking up spurious time trends” was “annoying” (link). Burke went on to say:

If you still think we messed this up, then download our data and show us. The onus is on you at this point, and just making claims about the potential for spurious trends does a disservice to the debate.

In this paper I take up Burke’s challenge. I go beyond the work of Rosen (2019), Tol (2019), and Newell et al. (2021) to look carefully at BHM’s model and a variety of econometric issues that cause BHM (2015), as Tol (2019) puts it, to go off the rails.

**Description of BHM (2015)**

BHM (2015) use annual data representing 166 countries from 1961 to 2010 on temperature and economic growth. All countries are equally weighted, and every country is assigned a single average temperature for each year. Because some data are missing, there is a total of 6,584 country/year observations instead of the 8,300 that could be used if data from all years in all countries were available.

The basic idea of BHM (2015) is to estimate GDP per capita growth as
a quadratic function of temperature, and then use this function and CMIP5 estimates of future temperature increases to estimate year-2100 GDP for each country. BHM find that 77 percent of all countries would be poorer with temperature increases than without increases, and 5 percent of countries would be poorer in 2100 than they are today because of temperature increases. Using more pessimistic IPCC growth scenarios, but without adjusting temperature changes to account for lower growth, BHM unsurprisingly find even larger effects.

The quadratic curve BHM estimate has an inverted U shape, which means that cool countries will grow more rapidly if they warm, countries with moderate temperatures will see little change, and warm countries will grow more slowly. BHM write:

We find country-level economic production is smooth, non-linear, and concave in temperature (Fig. 2a), with a maximum at 13° C, well below the threshold values recovered in micro-level analyses and consistent with predictions from equation (1). Cold-country productivity increases as annual temperature increases, until the optimum. Productivity declines gradually with further warming, and this decline accelerates at higher temperatures (Extended Data Fig. 1a–g). This result is globally representative and not driven by outliers. (BHM 2015, 236–237)

To find GDP per capita growth as a quadratic function of temperature, BHM regress annual growth of GDP per capita on temperature, temperature squared, precipitation, precipitation squared, and four sets of control variables. The first set of control variables consists of dummy variables for each country. The second consists of dummy variables for each year. The third consists of the dummy variables for each country multiplied by an index value for year, with observations for 1961 taking a value of 1, 1962 a value of 2, and so on. The fourth set of control variables consists of the dummy variables for each country multiplied by the square of the index value for year. Standard errors are clustered by country.

The precipitation variables are not statistically significant at the 5-percent level, and leaving them out of the regression makes very little difference to the

---

5. The Coupled Model Intercomparison Project is a project of the World Climate Research Programme, which is made up of organizations that are either part of or funded by the United Nations.
6. The Intergovernmental Panel on Climate Change (IPCC) describes itself as “the United Nations body for assessing the science related to climate change.”
7. The scenarios, known as Shared Socioeconomic Pathways (SSPs), are developed by the International Institute for Applied Systems Analysis (IIASA), a research institute that is funded by scientific organizations in member countries. The IIASA is a member of the IPCC. The economic growth projections associated with SSPs used in BHM (2015) are developed by the OECD, an intergovernmental organization.
estimated coefficients that are used in subsequent analysis.

The country dummy variables control for the mean of each country’s growth over time, and the year dummy variables control for each year’s average growth across countries. The rationale for these control variables is to ensure that the estimate of the effect of temperature on growth is not affected by the fact that some countries have persistent differences in growth rates, nor the fact that growth rates in all countries have common trends.

The final two sets of control variables are included to allow the growth rate of each country to have a quadratic trend independent of temperature. For example, if the growth in a country tends to increase or decrease with a concave or convex pattern, or if it follows a U or an inverted U shape, then any measured effect of temperature on growth for a particular country will be relative to this pattern.

Another way of describing the effects of these control variables is that the coefficients on temperature and temperature squared will show the effect of temperature independent of means of growth by country and year, and independent of any country-specific quadratic trend of growth over time. China, for example, had GDP per capita growth of −33 percent in 1962, the second year of the sample. Growth was then weak and variable until the late 1970s, when growth became consistently positive, and high. Because of this trend, the coefficients on the control variables indicating a quadratic trend in growth for China are highly statistically significant. This quadratic trend is assumed to be independent of temperature, and any effect of temperature on growth that is found is relative to this baseline pattern of growth.

This estimated quadratic curve is the basis of all forecasts in the paper. Expected temperature increases under RCP8.5 calculated by the CMIP5 for global grid points from the 1986–2005 average to the 2080–2100 expected average are weighted by population and aggregated by country. Temperature increases for each year are linear interpolations of this overall expected temperature increase. These temperature increases for each country are plugged into the estimated quadratic curve to form a predicted path of GDP per capita for each year until 2100. Some countries show higher income in 2100 than they would have with no temperature increase, while others show lower income. Overall, under the IPCC’s RCP8.5 scenario, BHM find that 77 percent of countries will be poorer in 2100 because of temperature increases than if there is no climate change.

Using Burkina Faso as an example, BHM use 28.07 degrees Celsius as the starting temperature for their forecasts that begin in 2010 based on the 1986–2005 average. The starting growth rate is 3.3 percent, based on the SSP3 and SSP5 scenarios with growth projected for each country by the OECD. The temperature in Burkina Faso is expected to increase by 4.5 degrees by the year 2100 based on the RCP8.5 emissions scenario, and so they are expected to increase by 4.5/90,
0.05 each year between 2010 and 2100. For the first year, the temperatures 28.07 and 28.02 are plugged into equation 1.

\[ g = 0.0127184 t - 0.0004871 t^2 \]  

(1)

Equation 1 predicts growth of −2.7 percent without warming and −2.8 percent with warming of 0.05 degrees, for a growth loss of 0.1 percent. This growth is subtracted from the beginning growth rate of 3.3 percent and multiplied by the beginning value of GDP per capita. This process is repeated each year. By the year 2100, the equation predicts growth of −10.3 percent at a temperature of 32.6. BHM cap temperatures at 30 so as not to predict out of sample, so for the year 2100 growth is estimated at a temperature of 30, for a growth reduction of 3 percent. SSP3 and SSP5 predict declining growth rates for Burkina Faso without climate change, down to 1.7 percent per year by the year 2100, so the BHM model predicts growth of −1.3 percent.

Burkina Faso is well above the estimated optimal temperature of 13, so temperature increases reduce growth. For a cold country such as Iceland, the quadratic curve predicts that temperature increases will increase growth. For a country near 13 degrees Celsius, like France, a small change in temperature would have little effect on economic growth.

Using the Stata and R code provided on BHM’s website, I was able to exactly replicate the results of the paper. Some data is common to that of Melissa Dell et al. (2012), and some errors, such as the inadvertent omission of Burma, are also common to both papers.

BHM present the quadratic curve differently than simply plotting equation 1. Figure 1 shows BHM’s Figure 2a, which is the estimate of the total effect of temperature on growth of GDP per capita over the period 1960–2010.8 The blue area is a 90-percent confidence interval.

Growth of GDP per capita peaks at approximately 13 degrees Celsius. Projected GDP per capita is close to 15 percent lower for cold and hot countries than for countries near the temperate optimal. The figure may remind one of Aristotle’s view of virtue as moderation between extremes.

To produce the curve in Figure 1, BHM calculated, for each observation and using values and coefficients for all independent variables in the regression, predicted growth when temperatures were allowed to range from −5 to 35 degrees Celsius. For each temperature in that range, they took the average predicted growth across observations. More specifically, they set the temperature at −5, −4, etc., and calculated the predicted growth for Argentina in 1962, Canada in 1970, etc., and

---

8. BHM’s data on GDP per capita begin in 1960, so the first observation of GDP growth is in 1961.
averaged them all. The result is overall predicted growth for each temperature with each country weighted equally. The pattern of growth with temperature calculated in this way is fashioned using the estimated coefficients of growth on temperature and temperature squared, and so it must, by construction, follow a smooth quadratic path. The estimated regression coefficients result in this curve having an inverted U shape, with growth peaking at approximately 13 degrees Celsius. Any data, no matter how noisy, will generate a smooth quadratic curve if one variable is regressed on another and its square and the predicted values of the dependent variable are plotted against possible values of the independent variable.

**Figure 1.** Quadratic relationship between growth and temperature from BHM (2015, 236 Figure 2a)

The headline result of a 23 percent reduction in GDP comes from taking each country’s projected GDP per capita with and without climate change, then taking the weighted average by population, and then taking the percentage difference between the weighted sum with and without climate change.

### Problems with BHM’s analysis

**An overly ambitious graph**

As Tol (2019) commented, the graphs in BHM are impressive. One of them, BHM’s Extended Data Figure 1, panels a–f, is reproduced here as Figure 2. The graph appears to confirm the quadratic relationship between growth and temperature described in the previous section in Equation 1 and Figure 1. BHM
estimate the quadratic relationship using the full panel of 6,584 country/year observations, but Figure 2 shows the average relationship between growth and temperature for 166 countries to illustrate the overall pattern. It also highlights five countries that clarify the pattern. The overall pattern, however, is not statistically significant, and the five countries are cherry-picked to make the relationship appear to be significant. While the figure is not a crucial part of BHM’s analysis, it is indicative of the misleading approach of the paper, and suggests alternative methods of measuring the relationship between growth and temperature.

Panel (a) at the top of Figure 2 shows the quadratic curve discussed earlier and temperature ranges of five countries on different continents. The panels b–f show individual country data on what BHM call the marginal effect of temperature fluctuations on growth. To calculate this marginal effect for each country, separate regressions for each country are run that are similar to the regression on all countries that generated the quadratic curve shown in Figure 1, but of course the control variables that are based on country dummy variables cannot be included on regressions run separately for each country. The regressions include precipitation, year, and year squared. For reasons that are not explained, precipitation squared is not included. The plots are produced as follows:

1. Regress growth on precipitation, year, and year squared, but not temperature.
2. Regress temperature on precipitation, year, and year squared.
3. Draw scatter plots of the residuals from #1 and the residuals from #2.

**Figure 2.** Reproduction of Extended Data Figure 1 panels a–f from BHM (2015) showing relationship between growth and temperature for countries
The resulting scatterplot points are detrended growth and temperature, controlled for precipitation. For Iceland, the coldest country shown, the slope of points is upward, corresponding to the point on which Iceland lands on the quadratic curve, meaning that higher temperatures were associated with higher growth. For France and the United States, near the maximum of the quadratic curve, the slopes are much flatter. For the warmer countries, Vietnam and Mali, the slopes are downward.

Panel (h) of BHM’s Extended Data Figure 1 shows the slopes for all 166 countries, and adds an OLS line showing a downward slope. I recreated the scatterplot in panel (h) as Figure 3, and highlighted the five countries selected by BHM, as BHM also do in panel (h). A negative slope of the OLS line shows that the slope is positive and high for cold countries, near zero for intermediate countries, and negative for warm countries. Figure 3 shows these slopes plotted against temperature for all countries. The examples shown in Figure 2, Iceland, France, the United States, Vietnam and Mali, are shown in red, while other countries are shown in blue.

**Figure 3.** Slopes of growth-temperature regressions by country

Panels a–f are misleading because they cherry-pick five countries that make it appear that slope is strongly negatively related to temperature, when the actual relationship is quite noisy. The correlation of slope and temperature for all countries is $-0.13$, while the correlation for the five selected countries is $-0.94$. The country-level regression coefficient on temperature for the five selected countries
is 74.5 percent higher than the coefficient obtained by running the regression on all countries. The regression slope for all countries is not statistically significant at the 10-percent level. A regression test of a quadratic relationship shows no significance of either temperature or temperature squared.

A closer examination of the data shown in Figure 3 reveals that the very weak relationship between slope and temperature is not consistent over the range of temperature. The median temperature of the 166 countries is 20.6 degrees, while the range of Figure 3 is from −10 to 30 degrees, so most countries are on the far-right side of the graph. For countries above the median, an OLS line would have a positive slope, although it is statistically insignificant. In other words, for the warmest half of all countries, the relationship that BHM highlight in their Extended Data Figure 1 does not exist. This is particularly important, since it is only in the warmest countries that BHM claim that warming temperatures will reduce growth.

**Figure 4.** Local polynomial regression of slope and temperature

Figure 4 shows a local polynomial regression of the same data from Figure 3. The gray band represents a 95-percent confidence interval. With the exception of the two coldest countries, Greenland and Mongolia, the slope is not statistically different from zero at any temperature. In other words, for nearly all countries, these data do not show a statistically significant relationship between growth and temperature. For the warmest half of all countries, those with temperatures above 20.6 degrees, the slope increases with temperature. Three of the five countries selected by BHM to illustrate the relationship are outside of the 95-percent confidence interval, and the other two are at the visual midpoint of the graph,
creating the impression of a downward relationship across all temperatures.

For the purposes of their Extended Data Figure 1, BHM examine the relationship between detrended growth and detrended temperature. In their main regression analysis, however, BHM use unadjusted growth and temperature with fixed effect controls. I will show later that BHM’s results disappear if detrended growth and temperature are used. BHM do not report this result.

It is worth repeating that this figure is not a crucial part of BHM. I discuss it here only as an illustration of the character of the presentation of data in BHM. In the following sections I examine BHM’s data and results in greater detail.

Data characteristics

The full panel of data that BHM use to estimate the quadratic curve shown in Figure 1 present a number of difficulties for estimation. The measures of GDP per capita growth contain extreme outliers, with annual growth ranging from a 70 percent drop in GDP per capita in Liberia in 1990 to an 88 percent increase in Equatorial Guinea in 1997. All countries are equally weighted, so unusual growth in a small country can have a large effect on the results. The panel data is unbalanced because data are missing in many years in some countries. Because of the changing mix of countries over time, the average yearly temperature across countries falls from 1960 to 1996 and then increases. Average growth generally falls until 1992 and then generally rises.

Residuals from the basic regression that produces the quadratic curve show a skewness coefficient of −0.67 and kurtosis of 28.4. That skewness level is high, and that kurtosis level is extreme. With a very high level of confidence, the residuals fail a variety of normal normality tests. Failures of these tests do not invalidate BHM’s results, but they should suggest additional caution in evaluating them. BHM do not mention anything about the distribution of their residuals.

The Variance Inflation Factors (VIF) in the basic regression are 1,310 for temperature and 1,484 for temperature squared. A rule of thumb for VIF is that a value larger than ten is cause for concern over multicollinearity (Forthofer 2007). The VIFs here, then, are more than 130 times higher than the threshold for cause for concern.

A Wooldridge test for autocorrelation (Wooldridge 2002) of the residuals from BHM’s main regression, a test that takes account of the panel structure of the data, rejects the null hypothesis of no autocorrelation with a very high level of confidence. The F statistic is 28.6, while the critical value at a 99-percent level of confidence is 6.1. The same test for temperature produces a test statistic of 276.5. In the presence of a unit root, regressing an autocorrelated variable on another autocorrelated variable can produce spurious findings in the sense that the results
appear to be statistically significant even when there is no underlying relationship between the two variables (Granger 1974).

Spurious regression results can also be caused by spatial autocorrelation (Muller 2023). Spatial autocorrelation also exists in BHM’s data. For each country in the sample, I determined the nearest neighboring countries and regressed annual GDP per capita growth on growth in the neighboring countries, as shown in Table 1. The first five spatial lags were all statistically significant at a 95-percent or higher level of confidence. Regressing temperature on its spatial lags produced even stronger results.

<table>
<thead>
<tr>
<th>Spatial lag</th>
<th>Coeff.</th>
<th>T-stat</th>
<th>P-value</th>
<th>Coeff.</th>
<th>T-stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.152</td>
<td>5.70</td>
<td>0.000</td>
<td>0.295</td>
<td>12.44</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.044</td>
<td>2.21</td>
<td>0.027</td>
<td>0.372</td>
<td>11.14</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.049</td>
<td>2.03</td>
<td>0.043</td>
<td>0.072</td>
<td>2.98</td>
<td>0.003</td>
</tr>
<tr>
<td>4</td>
<td>0.108</td>
<td>4.54</td>
<td>0.000</td>
<td>0.207</td>
<td>9.56</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>0.085</td>
<td>3.05</td>
<td>0.002</td>
<td>0.231</td>
<td>6.57</td>
<td>0.000</td>
</tr>
<tr>
<td>Constant</td>
<td>0.012</td>
<td>8.11</td>
<td>0.000</td>
<td>−1.725</td>
<td>−5.41</td>
<td>0.000</td>
</tr>
</tbody>
</table>

The data are also heteroskedastic, meaning that the variance of economic growth is very different for different countries. France, for example, has a standard deviation of growth of 1.9 percent, while Liberia has a standard deviation of 19.3 percent. Other than heteroskedasticity and temporal autocorrelation, none of these data characteristics are discussed in BHM, other than to say that “this result is globally representative and not driven by outliers” (2015, 236–237). The only corrections applied to the data are the fixed-effect control variables discussed earlier and the adjusting of standard errors for clustering by country to account for heteroskedasticity. These corrections make little difference to the results. A simple regression of growth on temperature and temperature squared with no controls and no corrections of standard errors shows the same quadratic relationship that is statistically significant, with an optimal temperature of 10.3 instead of 13.0 degrees.

BHM mention the spurious regression problem in their supplementary materials (link, p. 39), but only with regard to temporal serial correlation. There is no mention of potential spatial or other cross-sectional autocorrelation that might result in unit roots that could produce spurious results.

BHM do perform robustness checks for temporal autocorrelation by including lags of the dependent variable, GDP per capita growth. A graph in their Extended Data Figure 2 shows that including lags dramatically widens the confidence interval around the effect of temperature on growth, but the finding is
not discussed in the text of the paper. In their Extended Data Table 1, columns 9 and 10 also show that the inclusion of lags of the dependent variable substantially reduces the estimated magnitude of the effect of temperature on growth and its statistical significance, but, when including three lags, the result is still practically and statistically significant. This robustness check, however, says nothing about the possibility of spatial or other cross-sectional autocorrelation. Spatial autocorrelation is more likely to be a problem than temporal autocorrelation, since there are 166 countries in the sample and only 50 years. The sample is unbalanced, so some countries have as few as eight years of observations.

The presence of a unit root in economic growth or in temperature could cause spurious regression results. The measure of growth used by BHM is the first difference of GDP per capita, so in a single-country regression growth would not be expected to contain a unit root. In panel data with multiple countries, however, a unit root in economic growth is possible. Using the Levin-Lin-Chu (LLC) test for a unit root in panel data (Levin et al. 2002) that takes account of cross-sectional correlation, a unit root in economic growth cannot be rejected, and a unit root in temperature also cannot be rejected. The failure to reject occurred using a variety of lag structures and testing methods. The null hypothesis of the LLC test is that all panels contain unit roots. Even if this null hypothesis is rejected, it is possible that enough panels contain unit roots to cause spurious regression results. The Hadri test (Hadri 2000) has as a null hypothesis that all panels are stationary. This hypothesis is rejected, again using a variety of lag structures and testing methods for both growth and temperature. Both tests require balanced panels, so I only used countries with 50 years of data, which eliminated 80 out of 166 countries.

Unit roots in growth could be the result of the effects of location on growth that are independent of temperature. When latitude and longitude are added to BHM’s primary regression with growth of GDP per capita as the dependent variable, they are highly statistically significant, even though temperature and rainfall are controlled for. These results are shown in Table 2. When temperature is regressed on latitude and longitude, rainfall, and all of the fixed effect control variables in BHM’s primary regression, latitude and longitude are extraordinarily statistically significant. Rainfall and the control variables are also highly significant. Table 3 shows that the effects are nonlinear, with latitude squared significant in the growth regression, and both latitude and longitude squared and the product of latitude and longitude significant in the temperature regression. BHM’s analysis clearly has an endogeneity problem, and there is a possibility of spurious results coming from regressing dependent and independent variables that may contain unit roots.
TABLE 2. Growth and temperature regressed on latitude and longitude

<table>
<thead>
<tr>
<th></th>
<th>BHM Growth</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>T-stat</td>
</tr>
<tr>
<td>Temp</td>
<td>0.013</td>
<td>3.358</td>
</tr>
<tr>
<td>Temp$^2$</td>
<td>0.000</td>
<td>−4.114</td>
</tr>
<tr>
<td>Lat</td>
<td>0.013</td>
<td>2.596</td>
</tr>
<tr>
<td>Long</td>
<td>0.000</td>
<td>2.700</td>
</tr>
<tr>
<td>Rain</td>
<td>0.000</td>
<td>1.440</td>
</tr>
<tr>
<td>Rain$^2$</td>
<td>0.000</td>
<td>−1.861</td>
</tr>
<tr>
<td>R$^2$</td>
<td>0.286</td>
<td>0.288</td>
</tr>
<tr>
<td>Obs</td>
<td>6584</td>
<td>6519</td>
</tr>
</tbody>
</table>

TABLE 3. Growth and temperature regressed on latitude and longitude with nonlinearity

<table>
<thead>
<tr>
<th></th>
<th>BHM Growth</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>T-stat</td>
</tr>
<tr>
<td>Temp</td>
<td>0.013</td>
<td>3.358</td>
</tr>
<tr>
<td>Temp$^2$</td>
<td>0.000</td>
<td>−4.114</td>
</tr>
<tr>
<td>Lat</td>
<td>0.013</td>
<td>2.596</td>
</tr>
<tr>
<td>Lat$^2$</td>
<td>0.000</td>
<td>2.700</td>
</tr>
<tr>
<td>Long</td>
<td>0.005</td>
<td>1.527</td>
</tr>
<tr>
<td>Long$^2$</td>
<td>0.000</td>
<td>1.423</td>
</tr>
<tr>
<td>Lat×Long</td>
<td>0.000</td>
<td>0.880</td>
</tr>
<tr>
<td>Rain</td>
<td>0.000</td>
<td>1.440</td>
</tr>
<tr>
<td>Rain$^2$</td>
<td>0.000</td>
<td>−1.861</td>
</tr>
<tr>
<td>R$^2$</td>
<td>0.286</td>
<td>0.288</td>
</tr>
<tr>
<td>Obs</td>
<td>6584</td>
<td>6519</td>
</tr>
</tbody>
</table>

The quadratic relationship between growth and temperature

Basic robustness checks

All of BHM’s results depend on the quadratic function of growth that they estimate with respect to temperature using these data. The growth~temperature relationship is specified in Equation 1 and illustrated in Figure 1. BHM estimate Equation 1 by regressing growth on temperature and control variables using panel data consisting of annual observations from 166 countries. Average temperatures vary considerably between countries, and temperatures within in each country vary by year, but not as much as country averages differ from each other. The data characteristics described in the previous section recommend some basic robustness checks on the estimated quadratic relationship between temperature and growth.
Weighting

Table 2 shows BHM’s results, along with results from weighting residuals to be minimized by population to see whether small countries have a disproportionate effect on BHM’s estimates. In order to calculate country average annual temperatures from gridded temperature data, BHM weight the gridded temperatures by gridded population, but countries themselves are not weighted by population in their analysis. Weighting at the sub-country level but not at the country level seems arbitrary, and seeing whether population weighting of residuals affects BHM’s results would be a reasonable robustness check. While weighting by population creates its own problems by allowing large countries to dominate the results, substantial differences between population-weighted and non-weighted regression are indicators of potential problems. Results that are driven by small countries may be irrelevant for most of the world.

In addition, to see if heteroskedasticity affects the results, it seems reasonable to weight residuals by the inverse of the variance of growth by country. The second and third columns of numbers in Table 4 show that these weightings substantially reduce the statistical significance of BHM’s results. Weighting by population and variance increases the p-value of temperature and temperature squared from less than one tenth of one percent to over 10 percent.

<table>
<thead>
<tr>
<th>TABLE 4. Robustness checks for BHM results</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHM</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>Temp</td>
</tr>
<tr>
<td>Se</td>
</tr>
<tr>
<td>T-stat</td>
</tr>
<tr>
<td>P-value</td>
</tr>
<tr>
<td>Temp²</td>
</tr>
<tr>
<td>Se</td>
</tr>
<tr>
<td>T-stat</td>
</tr>
<tr>
<td>P-value</td>
</tr>
<tr>
<td>Opt. temp</td>
</tr>
<tr>
<td>Temp diff</td>
</tr>
<tr>
<td>R²</td>
</tr>
<tr>
<td>Obs</td>
</tr>
</tbody>
</table>

Lagged dependent variables

The fourth column of numbers shows the regression results with the
inclusion of first and second lagged values of growth as independent variables. BHM allow growth to have quadratic trends, but in their main specification they do not allow for an autocorrelated growth process. Their Extended Data Figure 2 shows that confidence intervals for the effect of temperature are wider with the addition of lagged growth. Moreover, in their Extended Data Table 1, columns 9 and 10 show a 50-percent reduction in the coefficient on temperature and a reduction in statistical significance from the 1-percent to the 10-percent level, and a 40-percent reduction in the coefficient on temperature squared. But in the text of the article BHM only say that their results are “robust to estimation procedures that...account for multiple lags of growth” (2015, 237). In their Supplementary Materials (link) they say that “our main result is robust across models that use alternative set[s] of controls” (p. 43).

The results in Table 4 show that in the weighted regression, the addition of two lags of growth eliminate the statistical significance of temperature and reduce the effect of changing temperature from the optimal level to 35 degrees by 76 percent. Dropping just two out of 6,252 observations—namely, India 1979 and Indonesia 1998—changes the sign of the coefficient on temperature squared, implying a positive effect of rising temperature, although with no statistical significance. In 1979 GDP per capita in India dropped by 7.7 percent, the worst year for India in BHM’s sample. The year 1979 was the 7th warmest year in BHM’s sample of 50 years. The decline in India’s GDP can be attributed to a balance of payments crisis triggered by oil price increases. Indonesia’s GDP per capita dropped by 15.5 percent in 1998, by far the worst in the sample. The year 1998 also happened to be the warmest year in the sample of 50 years. The 1998 drop in GDP can be attributed to the Asian financial crisis.

Weighting by country size and volatility, adding temporal and spatial lags of growth, and dropping two outlier observations out of 6,252 are reasonable model modifications that completely eliminate BHM’s results.

**Country effects vs. temporal effects**

A natural question to ask is how much of the measured effect of temperature on growth in BHM comes from differences in country averages versus annual fluctuations within countries. Dell et al. (2012) and Michael Kiley (2021) explain that observers have for centuries noted an association between the long-term average temperature of countries and economic development, which may be due to “spurious associations of temperature with national characteristics such as institutional quality” (Dell et al. 2012, 66). Country fixed-effect variables are included in their models of the effect of temperature on growth in order to ensure that their results are due to short-term temperature fluctuations, not “spurious
associations” with long term averages. Kiley (2021, 4) says about the inclusion of fixed effect variables that “this specification eliminates the ‘permanent’ component of weather, and hence may control for concerns regarding the link between the average temperature and the level of income across countries.” BHM explain their use of country fixed effects as follows:

We deconvolve economic growth to account for: (1) all constant differences between countries, for example, culture or history; (2) all common contemporaneous shocks, for example, global price changes or technological innovations; (3) country-specific quadratic trends in growth rates, which may arise, for example, from changing political institutions or economic policies; and (4) the possibly non-linear effects of annual average temperature and rainfall. (BHM 2015, 236)

If random temperature fluctuations affect economic growth independently of an association of average country temperature with average growth, then demeaned growth should show a relationship with demeaned temperature. In simulated data, if random temperature fluctuations affect growth while at the same time there is a correlation between average growth and average temperature, then fixed-effect controls remove the effect of average temperature on average growth, and the effect of temperature on growth is statistically significant. In addition, a regression of demeaned growth on demeaned temperature will show a statistically significant effect. If in simulated data there are quadratic trends in growth and/or temperature, then BHM’s fixed-effect controls will remove the effects of these trends and show any true effect of random temperature fluctuations affecting economic growth. I find, however, that after demeaning and detrending growth and temperature, there is no effect of temperature on growth.

In Table 5, the first column of numbers shows BHM’s results. The second column shows the results using as the dependent variable the demeaned growth, that is, the temperature for a country in a particular year minus the mean temperature of the country over the entire sample period. The third column shows the results from removing the mean and country-specific trend from growth. In other words, the dependent variable is the residual of a regression of growth on country fixed-effect variables, year fixed-effect variables, and interacted country fixed-effect variables (one set of them multiplied by time and another set multiplied by time squared). Time is an index, with 1961 equal to one, 1962 equal to two, etc.

The ninth row of numbers shows the optimal temperature calculated from the regression coefficients, and the tenth row shows the difference in the growth rate from the optimal temperature to 35 degrees Celsius, which is approximately the maximum projected 2100 temperature for any country.
The calculated optimal temperature is significantly higher once country means are removed, and the effect of temperature on growth is reduced by 84.6 percent. In BHM’s results, controlling for rainfall, fixed effects and trends in growth, annual growth is 23.5 percentage points lower in very hot countries compared to countries at the optimal temperature. Demeaning growth lowers this estimate to 3.6 percentage points. Removing growth trends from the dependent variable lowers the estimate to 0.22 percentage points. Removing growth trends also reduces \( R^2 \) from 28.6 percent to 0.018 percent.

Weighting residuals by the inverse of variance reduces the estimate to 0.1 percentage points. In the weighted regression, this small effect is statistically significant, but removing one out of the 6,584 observations, Greenland 1990, eliminates this significance. Greenland is the coldest country in the sample, and 1990 was the worst year for growth in that country. In 1990 the Black Angel mine in Greenland closed after financial losses that began in 1985 and exhaustion of extractable reserves, not because of the temperature in 1990. Annual production had been rising since the 1970s, and suddenly stopped when the mine was closed (Thomassen 2003). Gross sales from the mine were $112.6 million in 1989, 12.1 percent of Greenland’s GDP at the time. Greenland’s GDP per capita fell by 13.0 percent in 1990. The ocean fishing harvest was also significantly lower in 1990 than in 1989 (Booth and Knip 2014). The ocean fishing harvest seems unlikely to have been affected by air temperatures on land.

The measured effect of temperature on growth appears to be primarily the result of differences in country averages, not annual temperature fluctuations within countries. When country-specific quadratic trends are also removed, the effect of temperature on growth is reduced to statistical insignificance.

### TABLE 5. Regression results for quadratic relationship between growth and temperature

<table>
<thead>
<tr>
<th></th>
<th>BHM</th>
<th>Demeaned</th>
<th>Detrended</th>
<th>Clustered</th>
<th>Weighted</th>
<th>W/o 1 obs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>.0127</td>
<td>.00368</td>
<td>.000202</td>
<td>.000202</td>
<td>.000121</td>
<td>.0000614</td>
</tr>
<tr>
<td>Se</td>
<td>.00379</td>
<td>.00159</td>
<td>.00039</td>
<td>.0000625</td>
<td>.000044</td>
<td>.0000657</td>
</tr>
<tr>
<td>T-stat</td>
<td>3.36</td>
<td>2.32</td>
<td>.519</td>
<td>3.24</td>
<td>2.74</td>
<td>.935</td>
</tr>
<tr>
<td>P-value</td>
<td>.000975</td>
<td>.0216</td>
<td>.604</td>
<td>.00145</td>
<td>.00681</td>
<td>.351</td>
</tr>
<tr>
<td>Temp(^2)</td>
<td>-.000487</td>
<td>-.000109</td>
<td>-.626e-06</td>
<td>-.626e-06</td>
<td>-3.46e-06</td>
<td>-1.74e-06</td>
</tr>
<tr>
<td>Se</td>
<td>.000118</td>
<td>.0000538</td>
<td>.0000116</td>
<td>1.88e-06</td>
<td>1.50e-06</td>
<td>2.05e-06</td>
</tr>
<tr>
<td>T-stat</td>
<td>-4.11</td>
<td>-2.03</td>
<td>-.54</td>
<td>-3.33</td>
<td>-2.3</td>
<td>-8.49</td>
</tr>
<tr>
<td>P-value</td>
<td>.0000612</td>
<td>.0437</td>
<td>.589</td>
<td>.00107</td>
<td>.0228</td>
<td>.397</td>
</tr>
<tr>
<td>Opt. temp</td>
<td>13.1</td>
<td>16.8</td>
<td>16.2</td>
<td>16.2</td>
<td>17.4</td>
<td>17.6</td>
</tr>
<tr>
<td>Temp diff</td>
<td>.235</td>
<td>.0361</td>
<td>.00222</td>
<td>.00222</td>
<td>.00107</td>
<td>.000527</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>.286</td>
<td>.147</td>
<td>.000179</td>
<td>.000179</td>
<td>.000583</td>
<td>.0000329</td>
</tr>
<tr>
<td>Obs</td>
<td>6584</td>
<td>6584</td>
<td>6584</td>
<td>6584</td>
<td>6584</td>
<td>6583</td>
</tr>
</tbody>
</table>
Interestingly, when standard errors are adjusted for clustering, these small effects are still statistically significant, but with standard OLS regression they are not. This difference can be explained by heteroskedasticity of growth by country. Adjusting standard errors for clustering normally increases estimated standard errors, because variation within panels is normally smaller than variation across panels. Once the mean of growth is removed, there is more variation within some countries than between countries, leading the adjustment for clustering to lower standard errors and inflate the statistical significance of temperature. Weighting residuals by the inverse of the variance of growth of each country helps to control for heteroskedasticity, which reduces the cluster-adjusted statistical significance of temperature.

We can also demean and detrend temperature. Table 6 shows the results. Statistical significance is eliminated, although the coefficient estimates suggest extremely large effects. Weighting by the inverse of variance flips the sign of the effect. Similarly to the results in Table 5, when growth and temperature are detrended, $R^2$ is significantly reduced.

A regression of growth on temperature for 166 country averages produces a weak quadratic relationship between temperature and growth. Using instead, as BHM do, up to 50 annual observations for each country, country averages are replicated over and over, inflating statistical significance without adding additional information.

| TABLE 6. Regression results for quadratic relationship between growth and temperature |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                 | BHM             | Demeaned        | Detrended       | Weighted        |
| Temp                            | 0.0127          | −0.00141        | −0.00086        | 0.000748        |
| Se                              | 0.00379         | 0.0022          | 0.00161         | 0.000928        |
| T-stat                          | 3.36            | −0.642          | −0.532          | 0.806           |
| P-value                         | 0.00975         | 0.322           | 0.595           | 0.42            |
| Temp$^2$                        | −0.00049        | −0.00522        | −0.00155        | 0.00189         |
| Se                              | 0.000118        | 0.00275         | 0.00196         | 0.00105         |
| T-stat                          | −4.11           | −1.9            | −0.791          | 1.81            |
| P-value                         | $6.12 \times 10^{-5}$ | 0.0589          | 0.429           | 0.0699          |
| Opt. temp                       | 13.1            | −0.135          | −0.276          | −0.197          |
| Temp diff                       | 0.235           | 6.45            | 1.93            | −2.35           |
| R$^2$                           | 0.286           | 0.147           | 0.000262        | 0.000559        |
| Obs                             | 6584            | 6584            | 6584            | 6584            |

To ensure that the ‘generated regressor problem’ is not causing the reduction in the effect of temperature on growth, I generated bootstrapped standard errors using bootstrap samples for both stages of the regression. The results are similar.
**Country averages**

The previous section showed that BHM’s results come largely from country-by-country averages, not annual temperature fluctuations. But is there truly a relationship between average country growth and average temperature? There is some sign of such a relationship, but the evidence is statistically weak, as shown in Table 7.

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Var wtg</th>
<th>Pop wtg</th>
<th>Both</th>
<th>Pop w/o 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>0.000879</td>
<td>0.000249</td>
<td>0.00662</td>
<td>0.000669</td>
<td>0.000896</td>
</tr>
<tr>
<td>Se</td>
<td>0.000884</td>
<td>0.000696</td>
<td>0.00129</td>
<td>0.00107</td>
<td>0.000862</td>
</tr>
<tr>
<td>T-stat</td>
<td>0.995</td>
<td>0.357</td>
<td>5.12</td>
<td>0.627</td>
<td>1.04</td>
</tr>
<tr>
<td>P-value</td>
<td>0.321</td>
<td>0.721</td>
<td>8.38e-07</td>
<td>0.532</td>
<td>0.3</td>
</tr>
<tr>
<td>Temp^2</td>
<td>-0.0000427</td>
<td>-0.0000179</td>
<td>-0.000202</td>
<td>-0.0000129</td>
<td>-0.0000121</td>
</tr>
<tr>
<td>Se</td>
<td>0.0000264</td>
<td>0.0000211</td>
<td>0.0000363</td>
<td>0.0000301</td>
<td>0.000025</td>
</tr>
<tr>
<td>T-stat</td>
<td>-1.61</td>
<td>-0.849</td>
<td>-5.56</td>
<td>-0.427</td>
<td>-0.485</td>
</tr>
<tr>
<td>P-value</td>
<td>0.109</td>
<td>0.397</td>
<td>1.07e-07</td>
<td>0.67</td>
<td>0.628</td>
</tr>
<tr>
<td>Opt. temp</td>
<td>10.3</td>
<td>6.93</td>
<td>16.4</td>
<td>26</td>
<td>37</td>
</tr>
<tr>
<td>Temp diff</td>
<td>0.026</td>
<td>0.0141</td>
<td>0.0698</td>
<td>0.00104</td>
<td>0.0000484</td>
</tr>
<tr>
<td>R^2</td>
<td>0.0575</td>
<td>0.0402</td>
<td>0.179</td>
<td>0.0144</td>
<td>0.0702</td>
</tr>
<tr>
<td>Obs</td>
<td>166</td>
<td>166</td>
<td>166</td>
<td>166</td>
<td>163</td>
</tr>
</tbody>
</table>

The first column of numbers shows the results from unweighted regression of country average temperature and temperature squared on average growth. Temperature squared is just short of being statistically significant at the 10-percent level. The coefficient estimates are consistent with a quadratic relationship with an optimal temperature of 10.3 degrees, reasonably close to BHM’s estimate. However, if residuals by country are weighted by the inverse of variance of growth, the significance is eliminated. Weighing by population produces very significant coefficient estimates, and looming large is China, which has a large population, mid-level temperatures, and high growth. Removing China, Japan, and South Korea eliminates the significance of temperature squared, as shown in the last column of numbers. The fourth column of numbers shows the results of weighting by both population and the inverse of variance of growth, where the significance of temperature is also eliminated.

Table 8 shows the results of the same regressions, but without temperature squared, to see if there is a simple linear relationship between country average temperatures and growth. The row labeled “Temp Diff” shows the difference in predicted growth when the average country temperature is −5 versus 35 degrees.
In the unweighted regression, temperature has a negative relationship with growth, but the effect is reduced when residuals by country are weighted by population or the inverse of variance, and the sign reverses when residuals are weighted by both. This is primarily due to high-population countries with moderate, stable growth and low temperatures, such as Germany, France, the United States, and Japan, and other countries with high populations, high and reasonably stable growth, and high temperatures, such as Pakistan, India, Indonesia, and Vietnam.

**Regional effects**

**Figure 5.** Map of large world regions

The influence of China, Japan, and South Korea on the results in Table 7 suggests that regional factors may play a part in the measured relationship between temperature and growth. In this section, I first divide the world into three large regions, which I will call Orange, Green, and Blue, following the colors on the map.
in Figure 5. Orange includes North America except for Mexico, Europe including Russia, and Australia and New Zealand. Green includes continental countries in eastern, southeastern, and central Asia, not including southwestern Asian countries on the Indian Ocean. All other countries are included in Blue. I tested different regions to see if BHM’s results are consistent throughout the world.

Orange is relatively cool and grew at a moderate pace from 1960–2010. Green has intermediate temperatures and grew quickly, and Blue is relatively warm and grew slowly.

Data for Somalia and some very small countries are missing, and Burma is missing because of a coding error in BHM that came from Dell et al. (2012). Figure 6 shows the average growth and temperatures of the three regions over the period 1960–2010 with a quadratic curve fitting the three points, and Figure 5 shows a map of these regions. Figure 6 is constructed with actual growth rates and temperatures without adjustment for the fixed effects variables in BHM.

**Figure 6.** Fitted quadratic relationship between growth and temperature for three world regions

Figure 6 shows that a quadratic relationship between growth and temperature can appear to be the case, even when other factors are far more important than temperature. The relative unimportance of temperature is clear from more distant history. China and surrounding countries grew more slowly than Europe, North America, and Australia during the 19th and early 20th centuries (Pomeranz 2001),
and GDP of Africa and South America grew rapidly during the early 20th century. In other words, relative growth of the regions has changed over time, as relative temperatures have changed little.

Table 9 shows the regression results from BHM for the entire world and for the three regions. The effect of temperature on growth that BHM reports for the entire world is highly statistically significant. Within regions, however, the relationship between growth and temperature is weak and the coefficient estimates are unstable. Removing 1998, the year following the Asian financial crisis, from the regression for Green eliminates the statistical significance of temperature. Removing two out of 4,466 observations from Blue eliminates the statistical significance of temperature. The observations removed are the two neighboring countries of Georgia and Armenia in 1992, a year of war and upheaval for both.

Another example of the lack of consistency of the effect of temperature on growth by region can be seen by removing Greenland and central Africa from the sample. Table 10 shows the results of BHM, results with Greenland and central Africa removed, and then also removing the year 1992, a year of economic turmoil in many countries following the collapse of the USSR. Figure 7 shows a map of the countries removed.
TABLE 10. Regression results by region

<table>
<thead>
<tr>
<th></th>
<th>BHM</th>
<th>−GL, AF</th>
<th>−1992</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temp</td>
<td>0.0127</td>
<td>0.00736</td>
<td>0.00549</td>
</tr>
<tr>
<td>Se</td>
<td>0.00379</td>
<td>0.00437</td>
<td>0.00336</td>
</tr>
<tr>
<td>T-stat</td>
<td>3.36</td>
<td>1.68</td>
<td>1.63</td>
</tr>
<tr>
<td>P-value</td>
<td>0.000975</td>
<td>0.095</td>
<td>0.105</td>
</tr>
<tr>
<td>Temp²</td>
<td>−0.000487</td>
<td>−0.000187</td>
<td>−0.000111</td>
</tr>
<tr>
<td>Se</td>
<td>0.000118</td>
<td>0.00013</td>
<td>0.000116</td>
</tr>
<tr>
<td>T-stat</td>
<td>−4.11</td>
<td>−1.44</td>
<td>−0.962</td>
</tr>
<tr>
<td>P-value</td>
<td>0.0000612</td>
<td>0.152</td>
<td>0.338</td>
</tr>
<tr>
<td>Opt. temp</td>
<td>13.1</td>
<td>19.6</td>
<td>24.7</td>
</tr>
<tr>
<td>Temp diff</td>
<td>0.235</td>
<td>0.0441</td>
<td>0.0118</td>
</tr>
<tr>
<td>R²</td>
<td>0.286</td>
<td>0.303</td>
<td>0.29</td>
</tr>
<tr>
<td>Obs</td>
<td>6584</td>
<td>5503</td>
<td>5373</td>
</tr>
</tbody>
</table>

Figure 7. Greenland and Saharan and Central Africa

Simulation

The control variables used by BHM are supposed to extract the signal of temperature affecting growth that is independent of any noise from growth trends or country or year averages. But is it possible for a regression that includes BHM's extensive control variables to show an effect of temperature on growth when no such effect exists?

I simulated random temperatures over 50 years for 166 countries, with
growth and temperatures autocorrelated by location and time, with location being one dimensional. In other words, the simulated countries are labeled by number from 1 to 166, and countries with labels numerically close to each other have growth rates and temperatures that are correlated with each other more than do countries with numbers that are further apart. The temperature of the first country is normally distributed. The second country’s temperature is the first country’s temperature plus another normally distributed random number. The third country adds another random variable. This is repeated for each year. In this way, countries that are closer in distance have more correlated temperatures. Simulated growth is constructed in the same way. There is no relationship between growth and temperature, but both are spatially autocorrelated.

When growth is regressed on temperature and temperature squared using 1,000 different random number draws, even with all of the controls in BHM, in 71.3 percent of the 1,000 simulations temperature and temperature squared are both statistically significant at the 5 percent level. They are both significant at the one percent level 62.7 percent of the time, and they are both significant at the 0.1 percent level 55.3 percent of the time. The coefficient on temperature is significantly positive and temperature squared significantly negative at the 5 percent level 17.6 percent of the time. Adding temporal autocorrelation increases the false statistical significance of the results. Adding a small effect of temperature on growth that is reversed in the following year produces positive coefficients on temperature and negative coefficients on temperature squared with p-values less than 5 percent in 76.1 percent of the simulations.

Using the two-step process of first regressing growth on all of the control variables without temperature and obtaining residuals, and second, regressing temperature and temperature squared on those residuals, the coefficient estimates and their statistical significance are greatly diminished, just as was the case with BHM’s results.

Figure 8 shows the distribution of t-statistics in 1,000 simulations using the BHM fixed-effect model (blue line) and regressing temperature and temperature squared on the residuals of a regression of simulated growth on all of the fixed-effect controls but without temperature (red line). Using residuals reduces the statistical significance of temperature squared. (A figure for temperature is similar.) The results are similar to those obtained using BHM’s actual data, where the effect of temperature on growth is strong in a model containing both temperature and fixed effects, but much weaker when residuals from a regression using fixed effects without temperature are regressed on temperature.

The results of these simulations suggest that the large number of fixed-effect control variables in BHM might not be sufficient to correct for effects, such as spatial autocorrelation, that can produce spurious results.
Mitigation of effect over time

Persistence of effects

In section C.2 of the online Supplementary Materials for BHM is a discussion of a test for level vs. growth effects in which lagged temperatures are included as independent variables. The tests are not discussed in the text of BHM, although their Extended Data Figure 2 contains an illustration of the results without any discussion of what the results show or their importance. BHM admit (in the Supplementary Materials) that “in models that account for lagged effects…, projections become more uncertain” (p. 26).

The idea of the test is that if a hot year reduces growth in the current year, but growth is higher than normal the next year, then there will be no persistent effect of temperature. This would be reflected in a negative coefficient on temperature in the current year and a positive coefficient on lagged temperature. If the coefficient on lagged temperature is zero while the coefficient on temperature is negative, then the loss of output in the hot year will not be recouped. Additionally, if temperatures
rise over time, growth rates would fall over time. But if the signs are opposite and of similar magnitude, then there would be no long-term effect of temperature on GDP.

In their Supplementary Materials (p. 15), BHM include Supplementary Table S2, which shows the sum of the marginal effect of current and lagged temperatures for different numbers of included lags at different temperatures. The table shows estimated total effects and their standard deviations, but no p-values, t-statistics, or discussion of their statistical significance. Extended Data Figure 2 in BHM shows blue shading to illustrate the widening confidence intervals as lags are added, but there is no discussion of this in the paper. Calculating t-statistics and p-values from their Table S2 shows that adding one lag of temperature dramatically reduces the statistical significance of the effect. At 5 degrees Celsius, the p-value goes from 1 percent to 25.3 percent, and at 25 degrees it goes from 0.12 percent to 10.8 percent.

For Table S2, BHM include lagged temperature and lagged temperature squared and evaluate the marginal effects at different temperatures. Table 11 shows the results of a simpler approach. BHM model growth as a quadratic function of temperature, with higher temperatures increasing growth up to the optimal temperature and decreasing growth at temperatures above the optimum. Below the optimum, therefore, a linear model should show a positive relationship between growth and temperature, and it should show a negative linear relationship for temperatures above the optimum. A dummy variable called cold equal to zero for temperatures above the optimal temperature of 13.1 and another called hot that is the reverse of cold can be multiplied by the temperature variable and lagged temperature. Table 11 shows that the positive effect of temperature on growth for cold countries is a little more than offset by the effect in the next year, and the negative effect is almost offset in hot countries. A test of equality for each of the pairs of coefficients, however, shows no statistically significant difference.

<table>
<thead>
<tr>
<th>TABLE 11. Lagged temperatures, effect varying by temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>cold×temp</td>
</tr>
<tr>
<td>cold×lag temp</td>
</tr>
<tr>
<td>hot×temp</td>
</tr>
<tr>
<td>hot×lag temp</td>
</tr>
<tr>
<td>Test cold×temp=cold×lag temp</td>
</tr>
<tr>
<td>Test hot×temp=hot×lag temp</td>
</tr>
<tr>
<td>R²</td>
</tr>
<tr>
<td>Obs</td>
</tr>
</tbody>
</table>

If there is an effect of temperature on growth, it is clear from these results that growth bounces back from any reduction resulting from warm years, elimina-
ting any significant net effect. BHM apparently found this result and relegated it to an online appendix and a graph that is not explained or mentioned in the paper.

**Reduction in effect over time**

BHM claim that there is no difference in the effect of temperature on growth over time. In other words, the effect is just as strong in the early years of their sample as in the later years. BHM write:

We do not find that technological advances or the accumulation of wealth and experience since 1960 has fundamentally altered the relationship between productivity and temperature. Results using data from 1960–1989 and 1990–2010 are nearly identical. (BHM 2015, 237)

“Nearly identical” perhaps means that the coefficient estimate for temperature falls by only 8 percent and the coefficient for temperature squared falls by only 23 percent and both remain statistically significant, but at reduced levels of confidence. This change, however, results in an increase in the optimal temperature from 10.9 degrees to 15.2 degrees and a 48 percent decrease in the change in the growth rate from the optimal temperature to 35 degrees. Moving the cutoff date by one year, from 1989 to 1990, eliminates the statistical significance of both estimates at the 5-percent level. Because data are available for more countries later in the sample, a cutoff date of 1991 instead of 1990 does a better job of balancing the number of observations between the two subsamples of early years and late years. Table 12 shows the results of breaking the sample in 1989 and also in 1990.

<table>
<thead>
<tr>
<th>TABLE 12. Regression results for early and late years</th>
</tr>
</thead>
<tbody>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Temp</td>
</tr>
<tr>
<td>Se</td>
</tr>
<tr>
<td>T-stat</td>
</tr>
<tr>
<td>P-value</td>
</tr>
<tr>
<td>Temp²</td>
</tr>
<tr>
<td>Se</td>
</tr>
<tr>
<td>T-stat</td>
</tr>
<tr>
<td>P-value</td>
</tr>
<tr>
<td>Opt. temp</td>
</tr>
<tr>
<td>Temp diff</td>
</tr>
<tr>
<td>R²</td>
</tr>
<tr>
<td>Obs</td>
</tr>
</tbody>
</table>

9. BHM’s data on GDP per capita begin in 1960, so the first observation of GDP growth is in 1961.
The increase in the optimal temperature between the two subsamples suggests further tests of whether the estimated optimal temperature changes over time. If growth is a quadratic function of temperature, and the coefficients on temperature and temperature squared change over time, then the following equation with interactions between time and temperature could be estimated. Growth is signified by $g$, temperature by $h$, and time by $t$.

$$g = ah + bh^2 + cht + dh^2t$$  \hspace{1cm} (2)

Equation 2 can be rewritten as follows:

$$g = h(a + ct) + h^2(b + dt)$$  \hspace{1cm} (3)

The quadratic form of this function means that the optimal temperature will be minus the coefficient on $h$ divided by twice the coefficient on $h^2$.

$$\text{Optimal temperature} = \frac{-(a + ct)}{2(b + dt)}$$  \hspace{1cm} (4)

Table 13 shows the result of this estimation. The interaction terms are statistically insignificant, but the coefficient estimates imply an increase in the optimal temperature over time, as shown in Figure 9.

<table>
<thead>
<tr>
<th>TABLE 13. Interactions of time and temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>Temperature</td>
</tr>
<tr>
<td>Temperature$^2$</td>
</tr>
<tr>
<td>Time × Temperature</td>
</tr>
<tr>
<td>Time × Temperature$^2$</td>
</tr>
</tbody>
</table>

Figure 10 shows the estimated optimal temperature from rolling eight-year windows of observations beginning with 1961–1968 and ending with 2003–2010. A best-fit line showing an upward slope is also pictured. An eight-year window was chosen because shorter windows resulted in regressions with few observations and noisy results because many countries are missing data from the early years of the sample. Different length windows can produce different results, but most window sizes result in an upward sloping line.

These results from dividing the sample in two suggest that any effect that existed in the early years of BHM’s sample may be weaker in the later years. The results from time-and-temperature interaction and rolling window estimates are weak, but suggest that it is possible that either more countries benefit from higher
temperatures over time, or that countries have become better able to cope with higher temperatures. At the very least, BHM’s dismissal of the possibility that the data indicate the possibility of effect mitigation over time seems overly aggressive.

Figure 9. Optimal temperature over time

![Figure 9](image)

Figure 10. Optimal temperature in 8-year rolling windows of observations

![Figure 10](image)
The headline result

The headline result of BHM (2015, 235) is that “unmitigated warming is expected to reshape the global economy by reducing average global incomes roughly 23% by 2100.” The world average of this warming is 3.7 degrees. To calculate the 23 percent figure, BHM use the RCP8.5 SSP5 scenario to estimate average temperature for each country each year until 2100, and then plug that temperature into the quadratic function they estimated to predict growth for each country each year. GDP per capita in 2100 is calculated using these annual growth rates. Growth rates are then weighted by projected population in 2100 to calculate the overall percentage difference in GDP per capita for the entire world.

To determine the precision of the estimate, BHM perform a bootstrap procedure, where a different selection of countries is chosen 1,000 times to re-estimate the quadratic relationship between growth and temperature. The entire procedure is done each time, resulting in 1,000 different estimates of the percentage difference in world GDP per capita between a scenario with no temperature change and the temperature changes predicted by the RCP8.5 SSP5 scenario. Twenty-three percent is the “point estimate,” which means the result obtained by using the coefficient estimates of the quadratic relationship between growth and temperature from the non-bootstrapped sample. In Extended Data Table 3 in BHM, they report the median of the bootstrap estimates (21 percent) and several percentile values, including the 50th percentile value, which is the median (opposed to the mean). BHM do not report the mean estimate of the 1,000 bootstrap estimates, which is 0.11, less than half of the point estimate that they report as their headline result. While bootstrap estimates are typically used to evaluate standard errors of coefficients rather than mean estimates, such a large difference between the point estimate and the mean of the bootstrap estimates is an indication that something is amiss, perhaps the influence of countries with extreme swings in GDP per capita that are due to factors other than temperature, and also perhaps the effect of spatial and temporal autocorrelation.

BHM also do not discuss that the distribution of their estimated effect of higher temperatures is highly skewed and has extreme kurtosis, meaning that the distribution has a fat tail to the right. From Extended Data Table 3, one can calculate that there is a 29 percent chance that unmitigated warming will result in higher world GDP per capita in 2100, and a 5 percent chance that it will be more than 66 percent higher. The highest estimate in their 1,000 runs is that GDP per capita would be four times higher with warming than without. The model also predicts a one in ten chance that warming will reduce world output by more than 50 percent.

The headline result is also based on the RCP8.5 and SSP5 scenarios of future
emissions. The RCP8.5 scenario is estimated to be higher than the 98th percentile of expected concentrations of greenhouse gases by the year 2100 (van Vuuren 2011).

Conclusion

BHM (2015) is a complicated paper that makes strong claims. The authors use thousands of lines of code to run regressions containing over 500 variables to test a nonlinear model of temperature and growth for 166 countries and forecast economic growth out to the year 2100. Careful analysis of their work shows that they bury inconvenient results, use misleading charts to confuse readers, and fail to report obvious robustness checks. Simulations suggest that the statistical significance of their results is inflated.

BHM’s results are unreliable and their data do not support predictions of large economic losses due to rising temperatures. BHM (2015), along with Dell et al. (2012), Colacito et al. (2019), and Kiley (2021; forthcoming 2024) attempted to show that warming will reduce the rate of growth of GDP per capita. I have shown that all of these papers are seriously flawed (Barker 2022; 2023a; 2023b), and there is no credible evidence suggesting that warming will reduce the rate of growth of GDP per capita.

Continued economic growth at levels similar to what the world has experienced in recent years would increase the level of future economic activity by far more than Nordhaus’ (2018) estimate of the effect of warming on future world GDP. If warming does not affect the rate of economic growth, then the world is likely to be much richer in the future, with or without warming temperatures.

Code

Code used in this research is available from the journal website (link).

References

Barker, David. 2023b. Temperature Shocks and Economic Growth: Comment on Dell,
EFFECT OF TEMPERATURE ON ECONOMIC PRODUCTION


David Barker taught economics and finance at the University of Chicago and the University of Iowa. His Ph.D. is from the University of Chicago and he worked as an Economist at the Federal Reserve Bank of New York. He currently runs a real estate and finance company in Iowa and is a member of the Iowa Board of Regents. His email address is drb@barkerapartments.com.
Rejoinder on Ergodicity Economics

Matthew C. Ford¹ and John A. Kay²

In our critique of ergodicity economics (Ford and Kay 2023), we had three major aims. Firstly, we wanted to set out the reasons ergodicity economics (EE) and expected utility theory (EUT) differ. Secondly, we wanted to set out difficulties with the justification and practical implementation of EE, difficulties that EUT does not face. Finally, much of the literature on EE has criticised EUT, either explicitly or implicitly. We wanted to respond to these criticisms where they were erroneous and in doing so show that the two theories may not be as far apart as they appear in many practical situations, which has implications for experimental work.

In a response (Hulme et al. 2023) to our critique, eleven authors (hereafter ‘the authors’) involved in EE research raise two issues with our paper: our treatment of the mapping between EE and EUT; and our treatment of the link between a converging growth rate and final wealth. Understanding these, they suggest, accounts for many of our conclusions. We are grateful for the generosity and fullness of their reply, but we believe that it is predicated on a misreading of our paper—for which we must take some of the blame, since we made a mistake in one aspect of our analysis which they helpfully correct. We address this below.

Having clarified these issues, we believe that the main points we made about the justification of EE and its practical applicability remain salient and unanswered. The conclusions of our original paper remain: it is still the case that, unlike EUT, there is no clear justification for using EE; there is no clear guidance on which situations it is applicable in; and there is good reason to believe that, compared to EUT, it is often a poor model of the world.

¹. St. John’s College, Oxford OX1 3JP, UK.
². St. John’s College, Oxford OX1 3JP, UK.
Finally, the authors note that they are in at least partial agreement with our concerns about experimental work done on EE and that some of them are conducting a new experiment which addresses these issues (Hulme et al. 2023, 345; Skjold et al. 2023). We discuss the new experiment with two aims: to clarify what we might learn from its results, and to show the practical importance of several of the theoretical issues we raised in our original paper.

## Clarifying concerns

An important section of our original paper addressed an unintuitive result: there are gambles which, in the limit, converge almost surely on a multiplicative growth rate below 1. A multiplicative growth rate below 1 means that wealth will fall over time. Despite this, there exist utility functions which show that agents will take those gambles. But these agents prefer having more wealth to less wealth. Knowing this, it might seem that in the limit these utility functions will not show that agents will take these gambles—that is, perhaps they will take them when the gamble length is short but not when it is long—and therefore EUT and EE converge.

In fact, this intuition is incorrect. In our original paper, we write:

“Samuelson (1971)’s contribution is to show that this is not the case for many utility functions… In the limit case those high-value outcomes are suppressed, explaining why EUT might appear to coincide with growth optimality… risk-neutral EUT is not growth-optimal… Samuelson (1971) can be extended to show that growth-optimal behaviour necessarily violates the axioms of EUT because it is incompatible with any utility function.” (Ford and Kay 2023, 318–319; boldface added here)

The authors, however, appear to think that we argue the opposite. They write: “A mapping between the models exists, but the key condition which needs to be satisfied for this mapping to hold is that the utility function in EUT is chosen to be the ergodicity transformation of EE. It seems that the authors believe that the key condition is merely sufficiently long time scales” (Hulme et al. 2023, 336). We believe that both their points—on the EE-EUT mapping and on the distinction between convergence in growth and convergence in wealth—are based on the misunderstanding contained in that quotation.

Regarding the mapping: we do not dispute the fact that there is a unique mapping between EE and EUT, and the existence of this unique mapping is a result we drew on in our paper to show the difference between EE and EUT. We wrote:
Choosing a gamble to maximise the growth rate of wealth looks equivalent to choosing a specific utility function: if the dynamics are multiplicative we maximise the expected value of \( u(w) = \ln(w) \); if they are additive we maximise the expected value of \( u(w) = w \). Peter Carr and Umberto Cherubini (2020) draw on this insight and vary a stochastic clock to show that a variety of utility functions can be justified in this way, and Peters and Adamou (2021) sets out a general way to find a correspondence between utility functions and dynamics (where this exists). (Ford and Kay 2023, 319)

We believe that the misunderstanding stems from a difference in emphasis. Towards the end of our paper, we pull back from the theoretical analysis and raise the question of what the models imply for “real decisions which real people face” (Ford and Kay 2023, 326). We do believe it is likely that, for many of the choices people face, applying an EUT analysis having estimated a utility function and applying an EE analysis will result in much the same ordering over the choices available: this is what we meant when we wrote that “utility approaches and growth-optimal approaches are likely to give the same answer in many cases” (ibid.). We were implicitly assuming these choices are relatively small in number: in everyday life we do not face a continuum of marginally different stochastic processes. Of course we agree that, faced with the full set of possible stochastic processes, EE and EUT in general do not produce the same ordering.

Furthermore, one of our reasons for stressing this point was that much of the academic literature on EE implicitly misunderstands EUT by treating it as equivalent to myopic EUT. As we will discuss below, this was the case in the experimental analysis run by David Meder et al. (2021) and proposed by Benjamin Skjold et al. (2023). Furthermore, Ole Peters wrote that “by wrongly assuming ergodicity, wealth is often replaced with its expectation value before growth is computed”, and, referring to the experimental work subsequently published as Meder et al. (2021), “the experiment may be flawed in a way we don’t yet understand” (Peters 2019, 1216, 1220). In fact, EUT does not replace wealth with its expectation value before calculating growth, and this was indeed a flaw in Meder et al. (2021). In their reply, the authors claim that “EUT…does not take dynamic information into account” (Hulme et al. 2023, 341). This is simply incorrect, for reasons we explain in our original paper. (We would be happy to expand on this point if the authors could explain why they believe that this is not the case, but no such argument is offered.)

In general, myopic EUT and EUT properly applied will not give the same answers; we discuss the special case where they do below. This is what we meant when we wrote that the fact that the expected value and time average may not equate is “mechanically incorporated” in an EUT analysis (Ford and Kay 2023, 318). We did not dispute that, even with this incorporation, EUT and EE will in
general give different answers.

Regarding convergence: the authors identify a mistake we made when discussing the properties of terminal wealth, and we are grateful for the correction. As they note, we wrote that “The time averages used in [the EE model] correspond to a situation where there is no measurable uncertainty—final wealth will almost always be what the time average predicts” (Ford and Kay 2023, 317). They are correct to note that this is not true: as they write, even though the growth rate converges, terminal wealth diverges, and this uncertainty is measurable.

Without wishing to downplay the importance of this correction, we do not believe that it invalidates our argument. Our claim was that both growth and wealth are uncertain and that this explains why the intuition that EE and EUT should, in the limit, give the same result is incorrect. As we will discuss below, EE’s justification relies on the growth rate’s convergence; and of course this only matters because of its implications for terminal wealth. In the authors’ words, “EE focuses on maximising the time-average (or expected) growth rate of wealth. In doing so, it guarantees that EE agents, unlike EUT agents, maximize not only wealth but also utility as time passes” (Hulme et al. 2023, 340). Whilst the specific claim we made about the convergence of terminal wealth was wrong, the broader point we were making was correct.

**Unresolved issues**

We believe two important questions raised in our original paper remain: what is the justification for using EE; and which gambles is it appropriate to apply it to? The authors write that EE’s motivation is that agents who maximise the time average of the appropriately defined growth rate “maximize the long-term growth rate of their wealth. In the long run, agents who act in this way become wealthier than agents who act differently” (Hulme et al. 2023, 337–338). Elsewhere, they claim that EE “guarantees that EE agents, unlike EUT agents, maximize not only wealth but also utility as time passes. … As time passes, EE agents are guaranteed to do better than EUT agents, in terms of wealth and utility” and that “maximisation [of the time average] guarantees that we end up with greater wealth (and utility) in the long-time limit” (ibid., 340, 343). So the justification appears to be that using EE means your wealth grows faster and ends up being larger than using any other approach.

It is important to be clear on what the authors are saying here. Consider Peters’ bet, played for $t < \infty$ rounds. An EE agent will decline to play this bet and keep their initial wealth $w_0$. In contrast, an EUT agent with linear utility will accept this bet. As long as $t > 1$, they will probably end up with terminal wealth $w_t < w_0$; nevertheless it is possible that they will end up with $w_t > w_0$, for example if the coin
lands heads on every toss. “Guarantees” and “maximize,” therefore, do not mean that there exists any physically possible situation, i.e., a finite gamble, in which an EE agent will certainly end up with more wealth than an agent who takes the bet. Formally, they are saying \( P(w_t < w_0 \text{ for all sufficiently large } t) = 1 \). But for any finite value of \( t \), this probability will be less than 1.

In our original paper we quoted Henry Latané who explicitly discussed the “probability of adverse dominance” (Latané 1979, 310; quoted in Ford and Kay 2023, 322), and we believe that this is a much clearer perspective on the problem. His formulation also makes it clear that for an agent playing Peters’ bet \( P(w_t > w_0) > 0 \) for every \( t > 0 \). Given this, we do not think looking at the limit result provides much insight, and certainly does not justify claims that EE “guarantees” that EE agents do better than an agent who takes the bet.

In the light of this, we should update EE’s justification: using EE means your wealth probably grows faster and ends up being larger than using any other approach. But it is unclear why it makes sense to rank choices by their most probable outcomes. This is not how we generally act in life: many people fasten seatbelts and insure their house against fire even though the most probable outcome is that they are not involved in an accident and their house does not burn down; some, including some of the same people, buy lottery tickets for the pleasure of imagining a prize they do not expect to win. We apply for a good job which we know we will probably not get because the upside and chance that we will get it are sufficiently large relative to the costs of applying; we invest in a startup which we know probably won’t succeed because the potential payoff is enormous. There is no clear answer to the question ‘how probable must an outcome be before we can disregard all other outcomes?’ short of ‘it must be certain.’

In fact, as we note in our original paper, it is unclear why we would ever want to base a decision rule on any kind of average value, rather than examining the possible outcomes and their associated probabilities. Past work on EE (Peters and Gell-Mann 2016; Peters 2019) implicitly justified EE by arguing that EUT based decisions on an inappropriate average, whereas EE based them on an appropriate average. But as we explained in our original paper, this is not the justification generally made for EUT, which merely represents preferences. Under special circumstances, therefore, EUT’s use of an average is justified. But there is no equivalent justification for EE’s use of an average.

Our second question is about which finite gambles it is appropriate to apply EE to. Let us assume, for the sake of argument, that there is a good answer to why we want to use EE in the first place. Is there a rule with a convincing justification for which gambles are admissible?

The authors appear to address this: the model of EE they set out has no restrictions on \( t \), suggesting that it applies even to gambles of length 1. But as we
note in our original paper, this leads to bizarre results. Firstly, it implies that a gambler could develop a gamble where $t = 1$ and toss a coin to decide whether he describes it additively or multiplicatively. This is just an issue of labelling, of course: the gamble is exactly the same either way. And yet the EE agent will accept or decline the bet based on that description! The authors argue that EE preferences are complete as the dynamic in such cases is “a hidden variable” (Hulme et al. 2023, 341), but since this hidden variable has no impact on what actually occurs we do not find this a very persuasive defence.

In their appendix, the authors take another approach. They note that we can find a value, christened $T^*$, by equating the expectation and standard deviation of the growth rate of a particular gamble. Using this, they claim they can “quantify what we mean by ‘large’ time” (Hulme et al. 2023, Appendix p. 4). But it is not apparent why doing so tells us what we want to know. Indeed, they immediately undercut the idea that it does in three ways: they write that “large time’ corresponds to a number of iterations much larger than $[T^*]$” (without specifying what “much larger” might mean); that this number will also depend on “the required level of certainty of the growth rate”; and that in any case “standard deviation is not a robust measure of the dispersion” (Hulme et al. 2023, Appendix pp. 5, 7). The only thing this makes clear is that EE doesn’t guarantee a higher growth rate in real situations.

Secondly, whatever a gamble might look like in the limit, it seems clearly unpersuasive to say that we should take a certain gamble because doing so will maximise the long-term growth rate of our wealth, when in the time that actually elapses there is, for example, a 0.5 or 0.25 probability of it not maximising our growth rate.

Thirdly, having no minimum value of $t$ results in EE agents engaging in bizarre behaviour. It is easy to construct pairs of additive gambles where one has a marginally higher growth rate than the other and much higher variance. Thus we have a situation where gambles which are obviously very risky are taken in preference to gambles which are only marginally less generous and substantially safer. In our original paper we present an example and note that it implies EE may not apply even to gambles with very large values of $t$ (Ford and Kay 2023, 324).

Finally, we note in our original paper that EE violates completeness, meaning that there are gambles where it behaves inconsistently or offers no ranking. As we discuss above, cases where $t = 1$ are an example of the former issue. But, as we discuss at greater length in our original paper, there are other cases where EE certainly violates completeness (in both an EUT and EE sense) because it cannot rank gambles. Such cases include gambles with different dynamics and time-varying gambles, even in cases where these gambles do converge to a particular growth rate in the limit.
Experimental issues

As a result of the issues we discuss above, as well as others, we noted that two experimental studies on EE (Meder et al. 2021; Vanhoyweghen et al. 2022) had serious problems which meant that we did not follow their conclusions. The authors, several of whom were involved in those studies, accept at least some of these criticisms (Hulme et al. 2023, 345), and five of them have designed and preregistered a new experiment (Skjold et al. 2023).

The Skjold et al. (2023) experiment involves subjects making a series of choices between lotteries in additive and multiplicative dynamics; these lotteries change the number of experimental points they have, and their payout in real money is determined by their points. The choices are then analysed to estimate each subject’s coefficient of relative risk aversion, on the grounds that EE predicts these will differ between dynamics whereas EUT predicts that, for each individual subject, they will not. Unfortunately, this experimental design also appears to have problems, in large part because the issues we raised in our original paper—in general, and more specifically in our appendix—have not all been considered.

Two issues relate to our broader critique of EE. Firstly, as in Meder et al. (2021), the experiment uses myopic EUT to estimate the coefficient of relative risk aversion, resulting in the same issues we discussed in our original paper. In a different context, the authors discuss Peters’ bet under EUT and write that “we let the agent evaluate utility after one round, although nothing changes if the agent were to evaluate utility after an arbitrary number of rounds.” (Hulme et al. 2023, 342). It is true in the case they set up there that the two approaches are equivalent, but this is a special result which only occurs when the agent is staking their entire wealth on the gamble in each period and has a utility function from the constant relative risk aversion (CRRA) family. We mention this because, as we noted in our original paper, these conditions are not met in the experiments: firstly, agents may not have a CRRA utility function; secondly, they certainly have significant external wealth which is not subject to the experimental dynamics. Only considering CRRA utility functions and ignoring outside wealth also biases the experimental estimates for other reasons: readers interested in more detail about both of these issues should examine Appendix B in our original paper (Ford and Kay 2023, 328–332).

Secondly, the issues we noted above about when it is appropriate to use EE for finite gambles are relevant here. The experimental analysis appears to put equal weight on every data point gathered when estimating the coefficient of relative risk aversion. As we note above, gambles where \( t = 1 \), i.e., the last choice in each session, can be described as additive or multiplicative, so it is unclear what result
EE predicts for them. For gambles where $t > 1$ but still small, it is unclear whether EE should apply to them; at the very least it seems that proportionally less weight should be put on those choices. Furthermore, we note that subjects’ performance in the task is positively correlated with their payout in real money. In the additive session, we can think of each choice as being over an amount of money which will be taken out of the experiment into the real world, where the wealth dynamics the subjects will experience are presumably multiplicative. Applying the EE model, it seems that in this session they should treat every gamble as being part of a multiplicative dynamic and that their coefficient of relative risk aversion should equal 1, rather than the 0 they predict.

There is a further problem with the experimental design in Skjold et al. (2023). Similarly to Vanhoyweghen et al. (2022), gambles in the experiment are over experimental points rather than money. At the end of a session, subjects receive a share of a ~365USD prize pool based on their performance relative to 9 other subjects. (For example, if the subject had finished with 3,000 points and the other 9 subjects had collectively finished with 12,000 points, the subject would receive ~73USD.) The utility functions estimated are therefore over experimental points, not wealth. They would only be equivalent if there was a linear relationship between points and wealth (and if, as mentioned above, outside wealth was considered). This is not the case. If a subject has a very large point total, then (even holding the points earned by others in their group equal) they gain much less from getting additional points than they do when their point totals are lower. This asymmetry creates an incentive to be risk averse (when measured in terms of experimental points) even if a subject is really risk neutral (when measured in terms of wealth). The specific issue here is distinct from that in Vanhoyweghen et al. (2022), where the use of a payout based on experimental points and relative performance incentivised risk seeking, but the general issue is the same: the experimenters have not considered how, in EUT terms, the problem should be set up, and so make mistakes in their estimation.

**Conclusion**

We hope that this exchange has resulted in the debate making some progress. For our part, we are glad to have the opportunity to clarify what we meant in various passages about the practical differences between EUT and EE, and are happy to affirm that EE and EUT are only exactly equivalent when the conditions for the specific mapping between them is met—a result we relied on in our original

---

3. As we explained above, we do not find the claim that EE includes the dynamic as an additional variable persuasive. Regardless of one’s view on this, the experiment does not specify to the subjects that they are in an additive or multiplicative dynamic, so appealing to this additional variable does not deal with this issue.
paper. Similarly, we are grateful for the correction regarding diverging terminal wealth. Fortunately, our mistake did not have any serious implications for our analysis. Whilst we cannot make assumptions about what the authors leave unstated, we hope that the lack of discussion of the metaphysical issues with EUT means that this particular concern has been alleviated, and the comparison of the two theories can take place on less exalted territory.

Nonetheless, our original argument—hopefully clarified by this exchange—remains unanswered. It has three main elements: we do not believe that there is a compelling justification for EE; we do not believe it is at all clear when it is appropriate to use EE as a model; and we believe that EUT has been seriously misrepresented by proponents of EE, which undermines one of the justifications offered for EE. We further believe that, as a direct consequence of these issues, experimental work planned on EE is seriously flawed.

No single model can explain human behaviour and economists should use a toolbox of different models. Very few decisions—from what detergent to buy to which career to follow—are purely financial; most of the risks individuals face—such as those of disappointment or injury—are non-financial. EUT, which derives from axioms of consistent but idiosyncratic preferences, accommodates this; EE, by contrast, seems to have nothing to say here.

We remain confident in our original conclusion. Adapting George Box: all models are wrong, but some are useful; some are more wrong and less useful than others. The limited use case for EUT (and similar theories) is clear: it is a convenient mathematical representation of some sets of preferences which in some cases serves as a helpful way of representing the behaviour of modelled agents. Because it is axiomatic, it is clear when the situation we want to analyse does not match the model, and in those cases we must use our judgement to establish both how helpful and how limited its application will be. In contrast, the use case for EE is not obvious, and it remains unclear whether it is supposed to be a normative theory offering individuals advice, a descriptive theory, or some combination of the two.

References


Matthew Ford read History and Economics at St. John’s College, Oxford, and then worked with John Kay and Mervyn King as a research assistant on their book *Radical Uncertainty*. Having got interested in decision theory, he returned to St. John’s to write *Intellectual Capital*, an academic and financial history of the College, and to investigate how useful decision theories were in understanding the actual situation the College found itself in. His email is matthewcford@icloud.com.

John Kay has been a Fellow of St. John’s College, Oxford, since 1970. In 1979 he became research director and then director of the Institute for Fiscal Studies. In 1986 he founded London Economics, an economic consultancy. He was the first dean of Oxford’s Saïd Business School and has held chairs at London Business School, the University of Oxford, and the London School of Economics. He is a Fellow of the British Academy and the Royal Society of Edinburgh. His email address is johnkay@johnkay.com.
Counter-Reply to Naumenko on the Soviet Famine in Ukraine in 1933

Mark B. Tauger

L E I N T O A B S T R A C T

Econ Journal Watch in its September 2023 issue, after several revisions, published my article-response to a series of papers by the economist Natalya Naumenko on the famine of 1932–1933 in Ukraine. I wrote this article because Naumenko’s papers presented many false and misleading claims about Russian and Soviet history and about some of my earlier publications, and also presented detailed statistical calculations on the causes and character of that famine that were based on invalid and incorrect evidence. She also in one of the papers appropriated data and arguments that I had presented many years earlier in an article in an important American journal, yet did not cite my article and instead made it appear as if these data and arguments were exclusively her work. All of my criticisms were carefully documented, in contrast to many of her claims that had inadequate, misrepresented, or no documentation.

I am grateful to Econ Journal Watch for publishing my article, but they also published with it a brief reply by Naumenko, in which she tried to discredit my paper by using some of the same problematic approaches that she used in her earlier papers. After reading her reply, I decided that I had to write at least a partial counter-reply that explained these problematic approaches, so that readers will have the opportunity to understand the problems with her reply.

This counter-reply addresses the following issues:

---

1. West Virginia University, Morgantown, WV 26506.
1. The general importance of accurate background information. Naumenko dismissed it as unimportant for her arguments, but in fact the background information is important on its own and as an indicator of the writer’s knowledge and reliability.

2. Peasant and kolkhoz private food trade, one of the main background issues. On this topic Naumenko in her response again appropriated one of my arguments without acknowledgment, and on which she ignored substantial evidence I presented that contradicted her arguments.

3. The structure and functioning of the collective farms, especially the peasants’ private plots and the existence of incentives.

4. Most important, the relative importance of collectivization and agro-environmental factors in the causation of the famine. My points here are:

   ◦ Naumenko’s analysis of collectivization is based on inaccurate data.
   ◦ The main way she argues that collectivization caused the famine, high grain procurements, is not correct because as I documented the Soviet government sharply cut grain procurements for Ukraine in 1932.
   ◦ The agro-environmental factors that sharply reduced grain harvests in Ukraine and other regions were more important than Naumenko admits, and such disasters had caused many crop failures and famines in Russia and the early Soviet Union with no collectivization.
   ◦ Attributing the famine to collectivization would lead to the expectation that more famines should have occurred as collectivization increased in the 1930s, yet the collective farm system responded to the famine crisis by increasing grain production greatly in Ukraine and most other primary grain regions in 1933.

Background information

Naumenko (2023, 305) thanked me “for pointing out inaccuracies and oversimplifications in the Background section,” and promised to be “more nuanced” on the Civil War, food trade, and collectivization. She argued, however, that the Background section had no new findings and did not affect her paper’s conclusions.
This claim overlooked the problem that I emphasized in my article: that her “arguments and analysis are based on major historical inaccuracies and falsehoods, omissions of essential evidence contained in her sources or easily available, and substantial misunderstandings of certain key topics. All of these characteristics reflect a biased approach to the history and issues that these papers address” (Tauger 2023, 255). Naumenko failed to deal with this point: an author presenting an interpretation of Soviet history who gets basic aspects of this history wrong in an ideological manner is an author who is biased, and this makes her conclusions questionable.

**Peasant food trade**

Then she turned to one of her major mistakes in a background topic, her claim that the Soviet regime prohibited private trade in food. In response to my evidence and arguments against this claim, she argued in her reply that food trade “was in legal limbo until the ‘neo-NEP’ reforms of May 1932.” She asserts that “peasants were allowed to sell their produce there [at collective farm markets] only after the whole region where they lived fulfilled its compulsory grain procurement quota” (Naumenko 2023, 305).

There are two major problems with Naumenko’s responses here. First, all of the arguments that she made in this section repeated arguments that I made in my article, yet she does not acknowledge this duplication at all, instead presenting her claims as if they were exclusively her findings and implying that I did not make these points.

One example of this was the May 1932 decree that explicitly authorized kolkhoz and peasant private food trade after procurements were fulfilled. I documented this decree and this restriction the decree imposed on private trade in my article (Tauger 2023, 278). In my discussion I provided much more information than Naumenko did, including the date that the regime set for allowing trade—15 January—and references to two other related laws issued in the same month. Naumenko’s failure to acknowledge that I made this point in my article repeated one of her problematic practices that I described in my article when she cited the kolkhoz annual reports without acknowledging that I had published that information first (Tauger 2023, 259). Here again she made this point about the May 1932 law as if I had never mentioned it and as if it was exclusively her finding, when I was the scholar in this exchange who documented it first in my EJW article (ibid., 278), as well as in my first article in 1991.

The second problem with Naumenko’s response here is that she ignored the substantial evidence I presented from Soviet published and archival sources
and from Western eyewitness sources that contradicted her arguments in her reply and that showed that peasant food trade continued extensively despite these restrictions (Tauger 2023, 277–279).

On her “legal limbo” argument, she referred to a “slogan” at a grain conference in June 1931 that opposed “speculation,” implying that the regime opposed private trade (Naumenko 2023, 305). Yet she ignored the evidence I presented that while Soviet economic publications in 1930 argued that because of high inflation (in other words, speculation) and predicted “planned exchange” in the future, by May 1931 Sovnarkom reversed this and promoted private trade to replace rationing—in other words, before that “slogan” she quoted came out—and later the regime gave up trying to stop that trade and instead decided to tax it (Tauger 2023, 277). She treats that reference to a “slogan” as the only viewpoint at that time, yet I unambiguously documented that the regime was rejecting the viewpoint of that “slogan,” and she ignored my points and my evidence.

Additionally on this point of “legal limbo,” Naumenko ignored the GOSPLAN studies I cited that showed that peasants in 1930–1932 made billions of rubles on private food trade. This evidence clearly indicated that this “legal limbo” status she emphasizes had virtually no effect on peasant food trade (Tauger 2023, 277).

Then, Naumenko’s statement about the May 1932 decree’s restriction on private trade until after fulfillment of procurements ignored my arguments and evidence that peasants and most officials ignored that restriction. In particular, she ignored the archival document from the secret police (OGPU) in October 1932 that admitted that peasants were trading grain “everywhere” despite the 1932 law and that no officials were enforcing that law but were instead tolerating this trade. She ignored my point that this decree ordered confiscation of grain being sold illegally but also ordered no interruption of trade in other food products. This point in the OGPU document clearly showed that that May 1932 decree did not attempt to control all trade in food, but only in grain (Tauger 2023, 278–279).

The way Naumenko wrote suggested that she thought that if the Soviet regime issued a regulation, everyone automatically followed it, like robots, without any deviation. This viewpoint is biased and ignorant about Soviet history. As the evidence I presented in my article documented, those government statements and laws had virtually no effect, and trade continued, and certain Soviet leaders recognized this and did not act against it.

Naumenko’s “push back” against my points here constituted an attempt to mislead readers by hiding behind certain official Soviet statements and decrees while failing to acknowledge the much more substantial evidence I presented that shows that those statements and decrees were overturned and ignored, as well as failing to acknowledge that I had made her points earlier. I do not know whether
this failure on her part was intentional, the result of incomplete reading or forgetfulness, but either way it is clearly incompetent, and readers should be aware of this.

### Structure and functioning of collective farms

Naumenko (2023, 305–306) challenged my points about kolkhoz private plots and livestock. She cites the claim by Soviet historian Stephen Kotkin that Stalin and other leaders demanded formation of kommuna kolkhozy during the first collectivization drive in early 1930, referring to a “politburo resolution” of 5 January 1930.

I checked in my three-volume publication of Soviet Politburo agendas, and the decree on collectivization from 5 January has the same title as the Central Committee decree of 5 January, the same date: “On the tempo of collectivization and measures of government aid to kolkhoz construction” (Anderson et al. 2001, II:7; Sharova et al. 1957, 258–260).

This decree does not have any statement demanding formation of kommuna kolkhozy. Instead, in point 9 it states that the main form of kolkhoz at that time was the TOZ, which still had private property of the ‘means of production’ and that the new system should shift to the artel’ as a transitional form, in which the basic means of production would be under the control of the farm but would still allow a private sector. This point then commissioned the commissariat of agriculture and kolkhoz organizations to prepare a model artel’ statute.

If Naumenko is trying to use this decree to argue that Stalin was behind the “excesses” of promoting kommuna kolkhozy, this decree does not support that claim at all. The decree, which Stalin endorsed, unambiguously promoted the artel’ form. This decree was issued in January 1930, in the midst of the campaign, and two months later Stalin clearly endorsed the artel’ in his “Dizzy with Success” article, and simultaneously the government published that model statute of the artel’. Naumenko (2023, 306) cites Kotkin referring to “rabid members of the Yakovlev commission,” and I also referred to radicals who in the first campaign promoted kommuny (which again she did not acknowledge), but overall there were very few such kommuny formed in that campaign. Naumenko’s attempt to attribute all of those cases to Stalin is not supported by the evidence she cites. She also ignored the evidence and arguments I made that showed the dominance of the artel’ form in the history of collective farms in the USSR in the 1920s (Tauger 2023, 285–286).

Naumenko’s attempt to discredit my arguments is thus based again on her ignorance of the actual history and documents, and on what appears to be a bias.
She ignored the evidence I presented about private plots in Soviet kolkhozy from 1918 to 1929, a much longer period than the few weeks of the early 1930 campaign, during which the existence of a private sector in kolkhoz was firmly established, and it continued for the rest of Soviet history.

Naumenko (2023, 306–307) also attempts to challenge my arguments about the functioning of kolkhozy. First, she tries to undermine my point that peasant village communes in some ways anticipated collective farms, by arguing that peasants in communes were still “independent production units” and that “communes provided more economic freedom.”

These points are not quite correct. Peasants in village communes had to farm the crops that the village decided on in the fields defined by the village, and therefore all the peasants had to farm their strips in the spring crop field with the spring crops that the village decided on, and the same with the other fields. They were only “independent” in that each peasant household was responsible for farming their own strips, and if they did not farm their strips, they would not get any harvest from them. While communes did have certain economic freedoms that kolkhozy did not have, my research found that kolkhozy also had certain economic freedoms, both within the laws and by ignoring the laws, as they did in conducting private food trade long before the official date in the 1932 decree discussed above. These points are issues that would require detailed discussion with much more evidence than the unsupported assertions that Naumenko makes (on these points, see Tauger 1991). Nonetheless, peasants in village communes before collectivization were already working as groups and not as totally “independent production units,” and in that sense village communes already had elements of collective farming.

Then Naumenko tries to challenge my argument that kolkhozy did have some incentives. She presents three quotes asserting that kolkhozy lacked material incentives, or that those incentives did not replace the incentives of the market. These sources are extremely problematic. First, these writers and their sources overlooked the fact that collectivization reduced the need for labor in farms, because it eliminated interstripping. Then, I cited studies of kolkhozy in 1930 showing that they increased their cropped land despite low labor turnout because they no longer needed so much labor. This was one of the basic premises for collectivization—that it would free labor for industry.

Finally, despite all of these issues, the kolkhozy and sovkhozy farmed very large areas in 1931–1932 and afterwards, and while in some years they had poor harvests, in others they had good harvests. All those critical quotes that Naumenko cites ignored these facts: their views reflected an anti-collectivization bias. They also ignored the fact that many if not most agricultural systems around the world provide inadequate pay and low incentives for farm laborers (Tauger 2021).
Relative importance of collectivization and agro-environmental factors in the causes of the famine

Naumenko (2023, 308–311) begins this section with the weather issue. First, she admits that she “misread” my 1991 article and that she agrees with my points about the exaggerated official harvest estimates.

Then she disputes my points regarding the total grain procurements from Ukraine in 1932. She refers to her original sources, which are official figures. Naumenko ignores the much more reliable archival sources that I cited showing that Ukraine did not fulfill the procurement quota she cites, and that the regime even specified that additional procurement collections in early 1933 were to be used for seed (Tauger 2023, 260). She described my point as “puzzling,” but it is only puzzling because she cannot accept that her official evidence was wrong. This is the case even though in her NBER paper she referred to a table from the work of Davies and Wheatcroft that showed the procurement quotas were not met (Markevich et al. 2023, 9).

Then Naumenko responded to my point that the drought in 1931 affected Ukraine. My point was to show that Naumenko’s assertion that there was no drought in Ukraine was incorrect (Tauger 2023, 265). Naumenko in her response did not admit her mistake, but instead evaded the points I made and argued instead that there were other droughts in 1934 and 1936 “without catastrophic famine,” and that “picking just one factor that deviated from the average as the cause of harvest failure and famine is wrong” (Naumenko 2023, 309). Here again Naumenko repeated a point that I made without acknowledging it. As I wrote: “any study of agriculture will show that in most cases one cannot reduce agricultural production exclusively to weather” (Tauger 2023, 267). There I showed that Naumenko was the writer who was “picking just one factor” to argue that the famine did not have natural causes, yet in her reply she failed to acknowledge that this criticism applied to her work and that I had pointed it out in my article. She attempted to evade this by asserting that “it is important to not just look at one selected factor (April–June rainfall), but at weather overall” (Naumenko 2023, 309), but the weather was still just one factor—she dismissed all the other environmental factors that affect farm production.

Naumenko attempted to revive her comparative argument that the 1901–1915 weather showed that weather did not cause crop failure in 1931–1932. She acknowledged my point that the 1901–1915 period had harvest failures and that she ignored them, but she asserted that she “never claimed (nor thought)”
that his period saw only good weather and harvests. Yet nowhere in any of her
three papers did she ever acknowledge any of the massive weather disasters, crop
failures, or famines of the 1901–1915 period. She even repeated her false claim
that the 1891–1892 famine was “the last large famine under the tsarist regime”
(2023, 310), which clearly indicated that she did not think or acknowledge that there
were any later climate disasters, even though I documented clearly that there were
several. Furthermore, as in dealing with 1931–1932, she relied on official general
weather information for 1901–1915, when in fact the very long and substantial
tsarist government reports on the crop failures in those years showed extremely
volatile weather conditions in those years which would not be reflected in
generalized data she cited and her generalized approach to them.

Then Naumenko discussed the other environmental factors I had discussed,
pests and infestations. First, she tried to discredit my points by asserting that other
scholars did not mention these factors. Yet she ignored the fact that I used scientific
evidence that no previous scholars ever cited or even mentioned before I published
it, and that was not fully understood by the few scholars who cited it after I
published it, such as Davies and Wheatcroft.

Next she manipulated my point by arguing that if weather helped spread
these factors, “the grain production function should capture it.” In a footnote to
this point, she cited a passage I quoted from an archival source that attributed
the ergot infestation to favorable weather, and again asserted that this “should
be captured by the weather-driven grain production function” (Naumenko 2023,
310). Yet when she used her “grain production function,” she clearly did not
include anything beyond weather conditions, mainly temperature and rainfall. One of the
points I was trying to make in that discussion in my article was that since “favorable
weather” also supported the expansion of infestations, therefore “favorable
weather” would not inherently lead to improved production but could and did
also lead to serious crop damage and reduced production. Her assertions about the
“grain production function” concealed the fact that this function, as she devised
it and used it, completely overlooked non-weather environmental factors in grain
production.

Naumenko again referred to the failure of published documents to mention
these environmental factors, and then at least she admitted that I was correct about
unpublished documents showing these. She failed to acknowledge, however, my
points about the ignorance of Soviet officials regarding agriculture. Then she
argued that even if this is true, that her correlation of famine mortality with
collectivization, according to the “estimates” in her paper, showed the connection
between pests and diseases and collectivization. This point again overlooks what I
wrote. While to a small degree collectivization may have worsened some of these
environmental factors, many of them would have occurred without collectivization
as well, and in total they represented a major agro-environmental crisis that Naumenko, given her failure to reference any primary or scientific sources on these topics, is not in a position to deny.

Then Naumenko tried to argue that weather or infestations and pests could not have been the causes of the 1933 famine on the basis of reference to the 1891–1892 famine, which she had described as “the last large famine under the tsarist regime,” which it was not. She argued that Russian agricultural productivity slowly developed afterward, so that by 1928 the “Soviet economy” recovered to the pre-WWI level. She then proposed that agricultural technology returned to the level of 1913 or even 1891, and asked how natural disasters could have “killed five to ten million people—an order of magnitude more than in 1892?” (Naumenko 2023, 310). She asserted that the environmental factors had to have been “of Biblical proportions (and should be easily spotted in the data, which they are not)” or were exacerbated by collectivization, so that the famine was ultimately caused by government policies.

Yet as I showed, Naumenko did not use any of the archival evidence on the environmental factors, and Soviet officials were ignorant about these factors and did not understand them. Since Naumenko did not use any of the archival sources that documented these environmental factors, the fact that her data did not recognize those factors did not mean that they were not large-scale. As I documented in one of my archival citations, for example, “rust has the character of almost uniform infestation in the whole territory of the [Ukrainian] Republic” (quoted in Tauger 2023, 269). A total infestation of all of Ukraine is pretty close to “Biblical proportions,” but Naumenko again ignored, or overlooked, or forgot my arguments and evidence.

Her assertion that, during collectivization, agricultural technology regressed to the level of 1913 or 1891 was utterly false. In 1929 the Soviet government established VASKhNiL, the first national academy of agricultural sciences under the internationally recognized agronomist Nikolai Vavilov, and this agency immediately began taking a wide range of measures to improve agricultural production. Also in this period the Soviet government, as part of the first five-year plan, established many new factories to produce agricultural machinery, including tractors and combine harvesters, and also imported significant numbers of these (see Tauger 1991 and many other sources).

Then Naumenko turned to government policies (2023, 311–312). First, she indirectly insulted me by referring to a “gulf in training between quantitative fields like political science and economics and qualitative fields like history” which she claimed prevented “a more informed debate.” Then she referred to her Table 1 in her 2021 article (Naumenko 2021a, 180), Panel C Column 4, which she claimed compared mortality change from 1927–1928 to 1933 in regions with similar “pre-
famine characteristics.”

Yet she ignored the fact that this table claimed to relate mortality in 1933 to collectivization in 1930, which is absurd, as I documented using statistics that she did not use. As I documented clearly, there was no single level of collectivization anywhere in the USSR in 1930, especially in the Ukrainian Republic. She even asserted that she used for her calculations the level of collectivization in May 1930, which according to her article was 45 percent in Ukraine (Naumenko 2021a, 162). The archival data I presented in my article showed that the collectivization level in Ukraine in May 1930 was 41.5 percent, having declined drastically from 60.8 percent in March, and continued to decline to 28.8 percent into October and November (Tauger 2023, 290). She criticized my “training” as “qualitative,” implying that she was superior in her quantitative skills, yet in her reply she completely ignored the quantitative data I presented in my article. This was not a competent response to my points and evidence.

Naumenko claimed to compare similar regions in 1927–1928, but she never mentioned, either in her reply or in her original papers, the famine of 1928–1929 that was most severe in Ukraine. As I noted in my article, in her previous papers she explicitly denied that any famines took place in the later 1920s, which was completely wrong and which I had documented in previous published articles based on extensive archival research (Tauger 2023, 261–263).

Then Naumenko discussed issues of probability in order to reassert her earlier arguments that higher levels of collectivization were too strongly associated with higher mortality to be “just a coincidence” (2023, 312). This assertion is invalid for several reasons. First, she tried to connect a collectivization level in 1930 with mortality in 1933, which, as I explained above and in my article, is false because there was no single level of collectivization in 1930. Second, since collectivization changed significantly by 1932–1933, any connection between 1930 and 1933 omits those changes and is therefore invalid.

Third and most important, her calculations again omit any consideration of the agro-environmental disasters that harmed farm production in 1932. In her appendices, Table C3, she does the same calculation with collectivization data from 1932, which she argues shows a closer correlation between collectivization and famine mortality (Naumenko 2021b, 33). Yet, as I showed, those agro-environmental disasters were much worse in the regions with higher collectivization—especially Ukraine, the North Caucasus, and the Volga River basin (and also in Kazakhstan)—than elsewhere in the USSR. As I documented in my article and other publications, these were regions that had a history of environmental disasters that caused crop failures and famines repeatedly in Russian history.

The 1931–1933 famine was much more complex, and the agro-environmental factors had a much larger role in its causation, than her narrowly
focused and incomplete calculations accommodate. Furthermore, Naumenko’s emphasis on collectivization as a cause of the famine was based in part on her assumption that collectivization subjected peasants to higher procurements, but in 1932 in Ukraine this was clearly not the case. She ignored the fact, which I documented from archival sources, that the Soviet regime reduced grain procurements for Ukraine four times to approximately half their level in 1931 and, as I documented, that grain procurements both total and per-capita were much lower in Ukraine than anywhere else in the USSR in 1932 (Tauger 2023, 280ff., and table on 284). Clearly, if grain procurements were so greatly reduced, yet a severe famine followed, the agro-environmental factors were more important than collectivization in causing this famine.

Finally, Naumenko also overlooked a point I made at the end of my article: that I did not mean “to exonerate the Soviet regime completely for the famine” (Tauger 2023, 297). These agro-environmental disasters took place in a highly complex and politicized context, so they were not the only causes of the famine. At the same time, however, we cannot dismiss or minimize these factors, because if the harvests in 1931 and 1932 had actually been as large as the official figures indicated, in other words as large as the harvests of 1930 and 1933, especially given the reduced grain procurements in 1932, there would have been, as Naumenko herself recognized, much more food left in the villages, and no famine would have occurred. The agro-environmental crises have to be the central part of the explanation of the famine, even if government policies also played a role.

To summarize the points in this section:

• Naumenko’s attempt to reassert the connection between collectivization and mortality relied on inaccurate evidence and omitted the correct evidence that I presented in my article. In particular, in her reply she ignored the evidence I presented that showed that per-capita procurements were much lower in Ukraine than in any other primary grain region in 1932, as a result of the government’s repeated reduction of grain procurements for Ukraine in 1932.

• Naumenko ignored the evidence I presented that Russia and the early USSR had numerous serious crop failures and famines, not just one in 1891–1892, all caused by environmental disasters. Naumenko’s reply and earlier publications, by ignoring and suppressing this history, clearly constituted an attempt to misrepresent the real history of the 1931–1933 famine. These disasters and the crop failures and famines they caused occurred so many times in earlier Russian and Soviet history that this pattern was clearly sufficient to have been the primary cause of the 1931–1933 famines.
Finally, one further point on collectivization and famine. If collectivization was the cause of famine in 1931–1933, then one might expect that as collectivization levels increased after 1933, there should have been more famines. One might argue against this that the peasants adapted to collectivization afterwards. But the problem with this argument is that the collective farm system already reached more than 20 percent of peasants in 1930, yet they did not have a disastrous crop failure and famine in 1930, but instead increased cropland, farm production, and procurement of grain and other crops in that year. If collectivization caused famine, there should have been a disastrous famine that year, but there was not.

To understand this, we must remember that collectivization involved the same peasants farming the same crops in the same land as they had before, just in larger consolidated fields rather than dozens of strips scattered over the landscape. The peasants had done such large-scale farming for decades before the 1917 revolution when they worked on the big estates of the Russian landlords, so it was not new to them. The peasants, moreover, had wanted more land for decades before collectivization and the Bolshevik revolution, so they were quite ready to farm more land in 1930. This was one of the reasons, as I documented in my article, why during the first collectivization campaign in winter 1930, fewer than 5 percent of peasants put up any overt resistance to collectivization: many of them must have understood the advantage of having more land to farm.

All of these considerations imply that collectivization was not the disaster for Soviet peasant farming that Naumenko attempted to argue, and that the agro-environmental disasters of 1931–1932 played a central role in the famine crises of those years. Again, I acknowledge that some Soviet government policies and practices, especially their misunderstanding of these environmental factors, also played important roles in the famine. Still, if the harvests of 1931 and 1932 had been as good as the official figures showed, there would not have been any famine in these years.

I hope this counter-reply makes clear the problematic approaches in Naumenko’s reply, and I hope it encourages anyone who has read only her reply to read my initial article.

**References**


of Economic Research (Cambridge, Mass.). Link


About the Author

Mark Tauger grew up in Southern California. He earned a BA and MA in historical musicology at UCLA, then accepted a major fellowship there to do history, focusing on the history of agricultural development in the USSR. After he finished his Ph.D. on collective farms, he became a professor at West Virginia University, and a fellow at the Institute for Advanced Study at Princeton. He has written and published extensively on famines, agriculture, and agricultural history. His email address is mtauger@wvu.edu.

Go to archive of Comments section
Go to March 2024 issue
The false discovery rate in empirical research within economics and other social sciences is much higher than most researchers think. Apart from the well-documented problem of extensive specification searches (actual or potential ‘data mining,’ or what Gelman and Loken 2014 call the “garden of forking paths”) that naturally characterizes analysis of non-experimental, multi-dimensional, and extremely complex social science data, a less recognized reason comes from within the statistical framework itself. The classical, or frequentist, hypothesis testing framework that is a core element in most empirical analyses has some inherent conceptual and interpretational problems that lead to widespread misunderstandings and misuse. These problems are so prevalent that the American Statistical Association in 2016 issued an official statement on the proper use and interpretation of statistical tests (Wasserstein and Lazar 2016).

The subtleties of classical hypothesis testing are so deep that even professional statisticians often have difficulties understanding them correctly (McShane and Gal 2017). A root cause of the problem is the way of conditioning in such tests. All classical tests are based on distributions of test statistics under the null hypothesis, $H_0$. For example, the $p$-value is the probability of observing a test value at least as high as in the sample, given that $H_0$ is true. If the $p$-value is below some threshold value (typically 5 percent), $H_0$ is rejected in favor of an alternative hypothesis, $H_1$. The underlying rationale is that if the probability of observing the actual data (or more extreme data) is low under the null, then the null is probably false and, instead, the alternative hypothesis is probably true. As the
British statistician Ronald Fisher, who popularized the \( p \)-value, argued when seeing a low \( p \)-value: “either an exceptionally rare chance has occurred, or the theory [stated in \( H_0 \)] is not true” (Fisher 1959, 39).

However, if one is not careful such reasoning quickly leads to the ‘fallacy of the transposed conditional,’ by which the conditional probability of an event \( A \), given another event \( B \), i.e., \( P(A \mid B) \), is mistaken for the reverse conditional probability, \( P(B \mid A) \). All researchers in their first probability and statistics course have learned not to make this mistake; nevertheless, this is exactly the mistake researchers make when they interpret the \( p \)-value, i.e., \( P(D+ \mid H_0) \) where \( D+ \) stands for ‘the data or more extreme data,’ as giving a probabilistic assessment of the null hypothesis in light of the data, i.e., as \( P(H_0 \mid D) \). It is generally not recognized that the probability that \( H_0 \) is true, given the data, can be very high (and even higher than the probability that \( H_1 \) is true) when the \( p \)-value is low.

Similarly, the ‘significance level,’ \( \alpha \), in classical hypothesis testing is a conditional probability where the conditioning is on \( H_0 \). It gives the Type I error probability, i.e., the probability of rejecting a null hypothesis that is true: \( \alpha = P(H_0 \text{ is rejected} \mid H_0 \text{ is true}) \). If \( \alpha \) is prefixed at 5 percent, then in repeated application in different samples at most 5 percent of true null hypotheses will be erroneously rejected. So it comes stumbling close to mistakenly interpret \( \alpha \) as the fraction of false rejections in repeated application of the test, i.e., as \( P(H_0 \text{ is true} \mid H_0 \text{ is rejected}) \). However, this latter conditional probability is not controlled by the significance level. \( P(H_0 \text{ is true} \mid H_0 \text{ is rejected}) \) is denoted the ‘false discovery rate’ (FDR). The FDR depends on \( \alpha \), but it also depends on the power of the test, and on the unconditional probability of the null, \( P(H_0) \); see Equation (1) below.

A scientific empirical finding is often based on rejecting a null of no effect or relationship (e.g., between two variables). If the academic literature within a given field has produced a lot of such effects or relationships, i.e., a lot of null rejections, a natural question is: How many of these rejections are a mistake? In other words: What is the false discovery rate? To assess FDR, we need to assess \( P(H_0) \). Economists often implicitly (or explicitly, e.g., Abadie 2020) assume that \( P(H_0) \) is low. I will argue, however, that for both statistical and economic reasons,

---

2. The reliance on tail area probabilities in classical tests is an additional source of confusion. As noted above, \( D+ \) in the definition of the \( p \)-value is ‘the data or more extreme data.’ Thus, the \( p \)-value depends on hypothetical data that could have been observed but was not observed. The implications of this particular feature of classical tests are far-reaching but generally not recognized by empirical scientists. For example, correct computation of the \( p \)-value in principle requires knowledge of the researcher’s subjective experimental intentions; see Berger and Wolpert (1988) and Wagenmakers (2007) for details and illustrations of the paradoxes that this involves. See also Engsted and Schneider (2024) for a detailed discussion and survey of the foundational literature on \( p \)-values and hypothesis testing in the context of non-experimental data and empirical modeling within the social sciences.
this unconditional probability needs to be set at a high level, in general higher than 50 percent. As a consequence, since FDR depends positively on $P(H_0)$, then the FDR within economics is quite high and, in any case, substantially higher than the 5% Type I error rate implied by the 5% significance level chosen in most studies.

It is worth emphasizing that computing the FDR is not the same as correcting the significance level and $p$-value for multiple testing. The conventional 5% threshold for $\alpha$ applies for a single test in each sample. When more than one test is performed in a given sample, the significance level for each individual test needs to be adjusted downwards in order to control the overall $\alpha$ level. For example, if two independent tests are conducted, a significance level of 2.53 percent must be chosen for each test in order for the overall Type I error probability to be 5 percent ($1 - (1 - 0.0253)^2 = 0.05$). In most empirical papers multiple tests are carried out (e.g., several explanatory variables, often in various combinations, are tried), but it is seldom to see correction for multiple testing. Usually the 5% level is applied for each test, whereby the overall Type I error rate will be higher than 5 percent, and often substantially higher than 5 percent. Of course, this will only intensify the problem of a high FDR (as seen in Equation (1) below, an increase in $\alpha$ leads to an increase in FDR). But, as will become clear in the fourth section, even under the assumption that the true overall Type I error rate is 5 percent, the FDR will in general be substantially higher than 5 percent.

The issues discussed in this paper are well-known from the statistics literature and have been much discussed in other fields such as epidemiology and psychology. In economics, however, the issues are less well recognized and, hence, in my view deserve more discussion.

In the next section I formally define the FDR and show its dependence on the prior null probability, $P(H_0)$. In the third section I discuss how to assess $P(H_0)$, followed by, in the fourth section, an assessment of the FDR based on $\alpha$, test power, and $P(H_0)$. In the fifth section I discuss whether $p$-value < 0.05 is a sufficiently strict hurdle rate for declaring evidence against $H_0$ for a single test in a given sample. The final section contains some concluding remarks on the special problems that non-experimental social science data pose for hypothesis testing.

**The false discovery rate**

As described above, in a classical hypothesis test the significance level, $\alpha$, is the Type I error probability defined as $P(H_0$ is rejected $| H_0$ is true$)$. i.e., the probability of rejecting the null hypothesis $H_0$, given that $H_0$ is true. If $\alpha$ is set at 5 percent, then over many repetitions of the test, 5 percent of true null hypotheses will be erroneously rejected. The reverse conditional probability, however, $P(H_0$ is
true $\mid H_0$ is rejected), is not controlled by the significance level. This conditional probability is what defines the false discovery rate, or FDR, and is given as (Storey 2003; see Appendix A for a formal derivation):  

$$
\text{FDR} = P(H_0 \text{ is true } \mid H_0 \text{ is rejected}) = \frac{\alpha}{\alpha + \frac{(1 - \beta)(1 - P(H_0))}{P(H_0)}}.
$$

(1)

The FDR depends on:

- $\alpha$, the significance level
- $1 - \beta$, the power of the test, where $\beta = P(H_0 \text{ is not rejected } \mid H_0 \text{ is not true})$ is the Type II error probability, and
- $P(H_0)$, the prior probability that $H_0$ is true.

In the fourth section I will assess the FDR for tests at a 5% significance level and various levels of power. These concepts (significance level and power) are well-known among empirical researchers applying hypothesis tests. The prior probability, $P(H_0)$, however, is not a concept that is well-known and appreciated by most empirical researchers. I will therefore, in the next section, discuss it in detail.

**Prior probability of the null hypothesis**

As seen in Equation (1), the FDR depends on $P(H_0)$, i.e., the unconditional (prior) probability of the null. How should we assess this prior probability? I will discuss this in the context of a simple regression model,

$$
Y_i = a + bX_i + u_i, \quad i = 1, \ldots, n,
$$

(2)

where $Y_i$ and $X_i$ are economic variables, $u_i$ is the stochastic error term, and $a$ and $b$ are regression coefficients. The sample size is $n$, and subscript $i$ refers to observation number $i$. The data can be measured either across time (time series data) or cross-sectionally (e.g., across individuals at a point in time). The researcher

---

3. Storey (2003) calls it the “positive false discovery rate” because it is conditional on at least one ‘positive’ finding, i.e., null rejection. Equation (1) holds exactly for independent tests and asymptotically for weakly dependent tests.

4. The Type I error rate is sometimes called the ‘false positive rate,’ which is not to be confused with the FDR. Unfortunately, the very informative and widely cited study by Benjamin et al. 2018, denotes the conditional probability in Equation (1) as the “false positive rate”.

is interested in examining whether there is a relationship between $X_i$ and $Y_i$, and therefore conducts a classical test of the point null hypothesis $H_0: b = 0$, against the composite alternative $H_1: b \neq 0$. Assume the researcher applies the conventional 5% significance level such that $H_0$ is rejected in favor of $H_1$ if the $t$-statistic $> 2.0$ (or, equivalently, $p$-value $< 0.05$).

Assume that $H_0$ is rejected such that a statistically significant relationship between $X_i$ and $Y_i$ has been established. What is now the probability that this is a false discovery? To compute this probability we need to assess the prior probability $P(H_0)$ which is most surely context dependent and ultimately a subjective (personal) matter. But what can be said about $P(H_0)$ more generally?

It seems to be a widely held view in academic economics that point nulls should not be attached strong prior probability weight. After all, economists typically develop models of relationships between economic variables, not models of no relationships. Alberto Abadie states that the common situation in economics is one where “there are rarely reasons to put substantial prior probability on a point null” (2020, 193). In a Bayesian setting where researchers report classical tests while journal readers have priors over the parameter of interest, Abadie (2020) shows that, provided the prior probability of the null is low, a statistical non-rejection is generally much more informative than a rejection at the 5% level. Informativeness is measured by the discrepancy between the prior and the posterior distribution. When the null value $b = 0$ is assigned a low prior probability, a statistical rejection does not substantially alter this prior belief. By contrast, a non-rejection generally leads to a substantial change in beliefs. Intuitively, if $P(H_0)$ is low, you expect $H_0$ to be rejected. So, a non-rejection is surprising and, hence, informative. Abadie (2020) thus argues that the usual practice of conferring point null rejections a higher level of scientific significance than non-rejections is unwarranted.\(^5\)

Nowadays, statistically significant findings certainly get more attention than findings that are not statistically significant, cf. the discussions of ‘$p$-hacking,’ the ‘file drawer problem,’ and ‘publication bias,’ that occupy much of the current literature on the replication crisis in empirical research.\(^6\) This is a natural consequence of the fact that an empirical finding is almost always based on rejection of a null hypothesis, as in Equation (1) where finding a relationship between $X_i$ and $Y_i$ (the working hypothesis) requires rejection of $H_0: b = 0$.

However, not long ago it was not uncommon to have one’s working hypothesis (economic model) stated in the null and, so, failure to reject the null was

---

5. If the reader is uncomfortable with the (Bayesian) notion of a prior null probability, $P(H_0)$ may alternatively be thought of as simply the overall proportion of true nulls in the population of null hypotheses within a given field (see the next section).

6. See, e.g., Elliott et al. 2022 and the references therein.
taken as evidence in support of the model. Based on a sample of published articles in economics from the 1980s, Brad DeLong and Kevin Lang (1992) analyzed the fraction of unrejected null hypotheses that are true. In their sample, 78 of a total of 276 substantive economic hypotheses, formulated as a statistical null hypothesis, failed to be rejected. The 1980s saw a plethora of tests of, e.g., financial market efficiency (return unpredictability; overidentifying restrictions in rational expectations models), the unimportance of anticipated variables in monetary economics, and unit roots in macroeconomic variables (e.g., a unit root in real GDP was thought to have serious consequences for macroeconomic theory), all of which having the substantial model formulated as the null hypothesis. Here it is naturally interesting to analyze the fraction of false (or true) non-rejections.

Using the property of the $p$-value that it is uniformly distributed under the null, DeLong and Lang (1992) derived an upper bound to the fraction of unrejected null hypotheses that are true, and they found that none of the unrejected nulls in their sample was true. They concluded that “$\pi$ [the fraction of null hypotheses that are true] is essentially zero: that only a very small fraction of the null hypotheses in published articles are true. Failures to reject nulls are therefore almost always due to lack of power in the test, and not to the truth of the null hypothesis tested” (DeLong and Lang 1992, 1261).

Such analyses, where the null is the working hypothesis, have become less common. Nowadays, as discussed above, it is more common to have the economic model under consideration stated in the alternative hypothesis, so that rejection of the null is interpreted as support for the model. For example, in financial economics, instead of testing for market efficiency, researchers now develop models for new risk factors in the cross-section of asset returns and then test their empirical relevance by putting them in the statistical model stated in the alternative hypothesis, thus hoping for rejection of the null that the new factors are not relevant (I will elaborate on this example in the next section). In such analyses the (implicit) assumption is that the null has a low prior probability, cf. Abadie’s (2020) statement as referenced above.

From a statistical point of view, however, this line of reasoning is problematic. Singling out a point null, $H_0: b = 0$, against a continuum of values in the composite alternative, $H_1: b \neq 0$, must imply that this particular point value should have a substantial prior probability assessment; if not, why single it out? Statisticians James Berger and Thomas Sellke (1987, 115) point out that for such a test it will rarely be justifiable to choose $P(H_0)$ less than 50 percent. Furthermore, if the point null does not have a substantial prior probability, what is then the rationale behind choosing a low significance level, like 5 percent? As it is often stated in statistics textbooks, the null constitutes the maintained hypothesis that we strongly believe in and, hence, there has to be strong evidence against the null in
the data before we are willing to reject it; this requires a low significance level, i.e., a low probability of rejecting a true null. As E. L. Lehmann and Joseph Romano put it in their classic text on testing statistical hypotheses: “If one firmly believes the hypothesis to be true, extremely convincing evidence will be required before one is willing to give up this belief, and the significance level will accordingly be set very low” (2008, 58). The classical null hypothesis significance testing (NHST) paradigm is intended for settings where \(H_0\) is the maintained hypothesis that is held to be true unless there is strong evidence against it in the data, i.e., \(P(H_0)\) is taken to be high. It can be argued that in economics point nulls are rarely exactly true. However, \(H_0: b = 0\) should not be interpreted as an exact zero effect, but rather as a ‘negligible’ effect. Berger and Mohan Delampady (1987) show that, unless the sample size is very large, a point null will often be a good approximation to a small interval null.

Thus, for statistical reasons there are solid arguments for putting a high prior probability weight on the null hypothesis. If we beforehand believe that \(P(H_0)\) is low, testing a point null at conventionally low significance levels is almost by definition unsuited for the analysis. In the next section we shall see that there are also solid economic reasons for operating with a high prior null probability, although it goes against the (implicit) notion of many economists.

**Assessing the false discovery rate empirically**

Publication of comprehensive and progressively growing lists of statistically significant effects or relationships is a common phenomenon in many fields, and although several causes for an effect may be present, typically only a few causes contribute substantially. It is well-known that a multitude of risk factors for lung cancer have been identified, but cigarette smoking is the substantial factor accounting for 90 percent of lung cancer diagnoses (link). Similarly, Jonathan Sterne and George Davey Smith report that “by 1985 nearly 300 risk factors for coronary heart disease had been identified, and it is unlikely that more than a small fraction of these actually increase the risk of the disease” (2001, 227–228). For epidemiology they accordingly suggest to set the prior probability that the null hypothesis (no effect) is true to 90 percent. For experiments in psychology, Daniel Benjamin et al. (2018) state that the prior odds of \(H_1\) relative to \(H_0\) may be

---

7. In pure Neyman-Pearson hypothesis testing (where prior beliefs and the \(p\)-value are absent) the choice of a low significance level is due to Type I errors having more serious consequences than Type II errors. However, in the social sciences, meaningfully assigning ‘costs’ to these errors is very difficult, if not impossible, and is hardly ever done in practice.
about 1:10, corresponding to $P(H_0) \approx 90\%$, similar to Sterne and Smith’s (2001) suggestion.

Economics is no different. As was already realized by Edward Leamer (1978; 1983), scientific findings in empirical economics are typically based on passively observed, non-experimental data and extensive specification searches with the purpose of finding statistically significant effects. A given economic phenomenon naturally has several causes, but often the bulk of variation in the ‘dependent’ variable is explained by only a few of these. As alluded to in the previous section, in financial economics, several hundreds of risk factors for the stock market have been identified—the so-called ‘factor zoo’—and statistical significance at the 5% level seems to have been an important and necessary condition for publication of a new risk factor (Harvey et al. 2016; Harvey 2017; Harvey and Liu 2019). According to the Capital Asset Pricing Model (CAPM), only one systematic risk factor explains variation in expected returns in the cross-section of assets, and although CAPM is known to be empirically inadequate, it is generally accepted among financial economists that only a few additional factors are needed to account for the major part of the cross-sectional variation in asset returns. Thus, the vast majority of the hundreds of financial risk factors found statistically significant in the academic literature are not substantially important, just as the vast majority of the hundreds of statistically significant risk factors for lung cancer or coronary heart disease are not substantially important.

For predicting asset returns, Campbell Harvey asks the question: “Among the many variables that researchers have explored, how many do we believe have 1:1 odds of being true return predictors before we look at the data?” and he answers “Very few” (2017, 1419). For empirical analyses in financial economics, Harvey (2017) generally suggests to (explicitly or implicitly) operate with a prior null probability, $P(H_0)$, substantially above 50 percent. That seems to be a sober recommendation for empirical analyses in economics more generally.

Table 1 shows the false discovery rate (FDR), computed as in Equation (1), for testing at the 5% level and for various levels of statistical power and prior null probabilities. With a neutral a priori assessment of the hypotheses, i.e., $P(H_0) = P(H_1) = 0.50$, a test with maximum power (100 percent) gives an FDR = 4.8%, close to the Type I error rate of 5 percent. But for a low-powered test FDR is somewhat higher than 5 percent. Empirical researchers often pay scant attention to the power properties of their tests and tend to ignore (or are not aware) that many empirical studies are under-powered. In a recent meta-study, John Ioannidis et al. (2017) find that the typical power in empirical economic research is just 18 percent.

---

8. Statistical significance at the 5% level has been and continues to be the standard hurdle rate for publication in economics more generally (Andrews and Kasy 2019).
Table 1 shows that with neutral prior odds and such low power, 21.7 percent of all significant findings are false.

<table>
<thead>
<tr>
<th>Power (1 − (β))</th>
<th>0.18</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P(H_0) = 0.10)</td>
<td>0.030</td>
<td>0.011</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td>(= 0.25)</td>
<td>0.085</td>
<td>0.032</td>
<td>0.022</td>
<td>0.016</td>
</tr>
<tr>
<td>(= 0.50)</td>
<td>0.217</td>
<td>0.091</td>
<td>0.062</td>
<td>0.048</td>
</tr>
<tr>
<td>(= 0.75)</td>
<td>0.455</td>
<td>0.231</td>
<td>0.167</td>
<td>0.130</td>
</tr>
<tr>
<td>(= 0.90)</td>
<td>0.714</td>
<td>0.474</td>
<td>0.375</td>
<td>0.310</td>
</tr>
</tbody>
</table>

The FDR naturally increases with increasing \(P(H_0)\). As suggested above, \(P(H_0) = 90\%\) is not an unreasonable choice for many areas within epidemiology and the social sciences. If, in addition, the typical test has power as low as 18 percent (as shown by Ioannidis et al. 2017), Table 1 shows that the FDR is 71.4 percent. No wonder that there is a replication crisis in empirical research! The intuition behind the 71.4% false discovery rate is straightforward: Assume that 10,000 tests are conducted. In 9,000 of these tests \(H_0\) is true (\(P(H_0) = 0.90\)). With power equal to 18 percent, \(0.18 \cdot 1,000 = 180\) of the false nulls are correctly rejected. With a 5% significance level, \(0.05 \cdot 9,000 = 450\) of the true nulls are incorrectly rejected, and so among the total of 180 + 450 = 630 rejections, 450/630 = 71.4 percent are false discoveries.9

In some fields within economics, data samples are very large with thousands or even millions of observations (e.g., high-frequency financial data; register-based micro-data). In such analyses lack of power is not a problem. Instead, the problem in very large samples is that tiny and economically unimportant effects become statistically significant at conventional significance levels.10 Statistics and econometrics textbooks sometimes suggest to lower \(α\) when the sample becomes (very) large, but standard hypothesis testing theory generally contains no formal procedures for how to optimally relate \(α\) to \(n\), and reliance on conventional significance thresholds like 5 percent continues to dominate empirical research (as shown by, e.g., Harvey 2017; Andrews and Kasy 2019; Brodeur et al. 2020). There is no indication that empirical researchers generally lower the significance level when

9. Equation (1) also shows the danger of the intuitive notion that rejection of the null with a low-powered test is a strong result. If the null has a high prior probability assessment, such a rejection has a high probability of being wrong.

10. A \(t\)-statistic, for example, for \(H_0: b = 0\), is defined as the estimate of \(b\) divided by its standard error, where the latter automatically shrinks with the sample size, \(n\). Thus, for an arbitrarily small deviation from \(b = 0\), the \(t\)-statistic automatically increases with \(n\).
the sample size becomes big. Note also that, as seen in Table 1, even with maximum power, FDR is somewhat higher than the Type I error rate when the prior null probability is higher than 50 percent.

Thus, while low-powered studies naturally lead to a high FDR, in studies with large sample sizes and high power, tiny and economically insignificant findings that, as a consequence, have a high risk of being unreplicable in new samples, will often be statistically significant at conventional significance levels. Both cases thus contribute to the replication crisis.

A few decades ago, as noted in the previous section, it was not uncommon to have the proposed economic model stated in the null hypothesis, and not rejecting the null was evidence in support for the model, i.e., the ‘discovery’ was associated with non-rejection. In this case, specifying a low prior null probability would often be more reasonable. For example, the non-linear cross-equation parameter restrictions of highly stylized rational expectations models stated as the null hypothesis (see, e.g., Hansen and Sargent 1980) could reasonably be viewed as not likely to be true a priori (probably the implicit prior view of DeLong and Lang 1992). If we condition on $H_0$ not being rejected and compute the probability that the null is true, we obtain (see Appendix B):

$$P(H_0 \text{ is true} \mid H_0 \text{ is not rejected}) = \frac{(1 - \alpha)P(H_0)}{\beta + (1 - \alpha - \beta)P(H_0)}$$

(3)

Table 2 reports this conditional probability for the same $\alpha$ (5 percent), power levels, and prior null probabilities as in Table 1. Naturally, with maximum power (100 percent, i.e., $\beta = 0$), the null is certainly true if it is not rejected, independent of $P(H_0)$ and $\alpha$. However, if both power and $P(H_0)$ are low, the probability that $H_0$ is true if it is not rejected, is low. Table 2 is an alternative way of expressing DeLong and Lang’s (1992) insight that when none or only very few non-rejected nulls in a given sample of published articles are true (as they found), $P(H_0)$ must be low and the non-rejection must be due to low power of the test. Note, however, as argued above, this situation (low $P(H_0)$) is rarely relevant when we look at the more contemporary cases where an empirical finding is associated with rejection of the null. In such cases, the FDR in Table 1 with high values of $P(H_0)$ is more relevant. And, as argued in the previous section, from a statistical point of view, meaningfully

11. With 100 percent power, all false nulls will be rejected. Thus, if $H_0$ is not rejected it must be true, since if $H_0$ was not true it would have been rejected. More generally, one minus the conditional probability in Equation (3) is $P(H_0 \text{ is not true} \mid H_0 \text{ is not rejected})$ which can be denoted the False Non-Discovery Rate, FNDR. When the power of tests is low, this rate can be quite high. Thus, power has a substantial effect on both FDR and FNDR.
testing a point null against a composite alternative with a low \( \alpha \) level in any case requires a high \( P(H_0) \).

<table>
<thead>
<tr>
<th>( P(H_0) )</th>
<th>0.10</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power ( (1 - \beta) )</td>
<td>0.18</td>
<td>0.279</td>
<td>0.537</td>
<td>0.777</td>
<td>0.912</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.588</td>
<td>0.851</td>
<td>0.945</td>
<td>0.912</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.59</td>
<td>0.792</td>
<td>0.919</td>
<td>0.972</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( P(H_0) )</th>
<th>0.10</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power ( (1 - \beta) )</td>
<td>0.114</td>
<td>0.279</td>
<td>0.572</td>
<td>0.855</td>
<td>0.912</td>
</tr>
<tr>
<td></td>
<td>0.174</td>
<td>0.388</td>
<td>0.655</td>
<td>0.945</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>0.297</td>
<td>0.559</td>
<td>0.792</td>
<td>0.972</td>
<td>0.972</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

I have discussed \( P(H_0) \) in the language of a (Bayesian) prior probability. I believe this is the most natural way to consider it (see Engsted and Schneider 2024 for elaboration). However, \( P(H_0) \) can alternatively be given a purely classical interpretation as the overall proportion of true nulls in the population of null hypotheses within a given field, and there exist classical ways of estimating this proportion. Interestingly, such estimates often indicate that the proportion is high, thus lending support to the recommendation of setting a high value of \( P(H_0) \). For example, Laurent Barras et al. (2010), in separating skill from luck in assessing mutual fund performance across more than 2,000 mutual funds, estimate the proportion of truly neutral funds in the population (i.e., funds that deliver zero risk-adjusted excess returns, so-called ‘zero alpha’ funds) to be 75 percent, corresponding to \( P(H_0) = 0.75 \) associated with the null of zero alpha.

12 As in DeLong and Lang (1992), the estimate of \( P(H_0) \) in Barras et al. (2010) is based on the property of the \( p \)-value that it is uniformly distributed in the interval \((0, 1)\) under the null.

**Is \( p \)-value < 0.05 a sufficiently strict hurdle rate?**

As argued in the previous section, in those areas where a scientific finding is associated with rejection of a statistical null hypothesis, the fraction of false findings across the many rejections is probably quite high, and thus is a contributing factor behind the replication crisis. As a ‘quick fix’ in order to reduce this fraction, Benjamin et al. (2018) have suggested to redefine statistical significance by lowering the conventional 5% level to 0.5%, corresponding to raising the \( t \)-statistic threshold from 2.0 to 2.8.

Working with a particular project, i.e., a particular model and a particular
sample of data, many researchers will undoubtedly feel that a $t$-statistic threshold of 2.8 is too high. Researchers have been trained and accustomed to believing that $t = 2.0$ is a suitable hurdle to pass for having confidence in a discovery. It is generally not recognized that if—as I have argued above should be the case—substantial prior probability weight is put on $H_0$, the 5% significance level ($p$-value threshold of 0.05) is in fact a quite low hurdle rate.

To see this, let us look at the conditional probability $P(H_0 \mid D)$, i.e., the probability that the null is true, given the data, associated with a $t$-statistic of 2.0. Define $BF$ as the relative data likelihood under $H_0$ and $H_1$, i.e., $BF = P(D \mid H_0)/P(D \mid H_1)$. It can be shown that (see Appendix C):

$$P(H_0 \mid D) = \frac{P(H_0) BF}{1 + [P(H_0)(BF - 1)]} \tag{4}$$

In classical statistics, $BF$ is a likelihood ratio statistic given as the data likelihood evaluated at the parameter value under the null divided by the data likelihood evaluated at the maximum likelihood estimate of the parameter. In Bayesian statistics $BF$ stands for the ‘Bayes factor,’ where the numerator is the same as in the classical likelihood ratio, whereas the denominator is a marginal likelihood computed by positing a density function for the parameter and then averaging (integrating) over the parameter space.

Let the parameter of interest be $b$ (e.g., the regression slope parameter in Equation (2)). If we apply the $BF$ that for a given $t$-statistic gives maximum evidence against the null, i.e., the lower bound for $BF$, the result is $BF = \exp\left(-\frac{1}{2} t^2\right)$, where $t$ is the usual $t$-statistic associated with $H_0$: $b = 0$ (see Berger and Sellke 1987). This particular $BF$ is in fact identical to the classical likelihood ratio since maximum evidence against $H_0$ is obtained for a parameter density that is completely concentrated at the maximum likelihood estimate.

For $t = 2.0$ we obtain $BF = \exp\left(-\frac{1}{2} \cdot (2.0)^2\right) = 0.135$, and with neutral prior odds for the hypotheses $P(H_0) = P(H_1) = 0.50$, we get, by inserting in Equation (4), $P(H_0 \mid D) = 0.50 \cdot 0.135/(1 + [0.50 \cdot (0.135 - 1)]) = 0.119$. For $P(H_0) = 0.90$, that I argued in the fourth section would be a reasonable choice in many cases, $P(H_0 \mid D) = 0.549$, that is, although the null would be rejected at the 5% level with a classical test, $H_0$ is more likely true than $H_1$!

$P(H_0 \mid D)$, computed as in Equation (4), is denoted the “Bayesianized $p$-value” by Harvey (2017), because it is the Bayesian equivalent to the classical $p$-value. As I noted in the introduction, the classical $p$-value is often misinterpreted as giving a probabilistic assessment of the null hypothesis in light of the data. We often informally—and incorrectly—associate ‘statistical significance’ with a low
probability that the null is true. \( P(H_0 \mid D) \) in Equation (4) is the correct statement of this probability and, as seen, it depends on the prior probability assessment of the hypothesis, just as we have seen in the previous sections that the false discovery rate depends on this prior probability.

The above calculation shows that with a neutral prior stance on the null and alternative hypotheses, the Bayesianized \( p \)-value is an order of magnitude higher than the classical \( p \)-value (11.9 percent versus 5 percent) and, mind you, for a choice of \( BF \) that gives absolutely most probabilistic evidence against the null. Of course, the two types of \( p \)-values are not directly comparable because they measure different things. However, since many empirical researchers tend to associate the classical \( p \)-value with a probabilistic assessment of \( H_0 \), such researchers will undoubtedly be surprised to learn that what they traditionally consider quite strong evidence against \( H_0 \) based on ‘objective’ statistical criteria, in fact implies a probability assessment of \( H_0 \) of 11.9 percent or higher. If researchers, in addition, actually think more deeply about the choice of a 5% significance level and consider it consistent with a prior probability assessment of \( H_0 \) even higher than 50 percent (cf. the previous sections), the Bayesianized \( p \)-value \( P(H_0 \mid D) \) associated with a classical \( p \)-value of 5 percent, will be so high that no one would even consider rejecting the null when seeing a \( t \)-statistic of 2.0.

Of course, in reality, researchers may (implicitly) operate with a very low prior probability assessment of the null (as, e.g., Abadie 2020), in which case the 5% level may be regarded a sufficiently strict threshold value. If researchers (implicitly) take a rejection at the 5% level to imply that \( P(H_0 \mid D) < 5\% \), then their (implicit) prior null assessment is \( P(H_0) < 28\% \) and, thus, \( P(H_1) > 72\% \).13

Richard Startz (2014) provides a concrete example of the implicit prior probabilities of the hypotheses implied by the choice of a 5% significance level in a classical test, and he concludes: “We usually think that our standards for significance are chosen precisely to point in the direction of the null unless we have strong evidence to the contrary. But as this example illustrates, our usual standards do not accomplish that goal. In other words, in this example the \( p \)-values we usually regard as providing strong evidence against the null and in favor of the alternative do not in fact provide such evidence unless the econometrician already leaned strongly toward the alternative” (Startz 2014, 141).

13. From Equation (4) we obtain \( P(H_0) = \frac{P(H_0 \mid D)}{P(H_0 \mid D) + BF} = \frac{0.05}{0.05 + 0.05 \cdot 0.135 + 0.135} = 0.28 \), for the \( BF \) that gives maximum evidence against \( H_0 \) \( (BF = \exp(-\frac{1}{2} \cdot (2.0)^2) = 0.135) \). Thus, 28 percent is the highest possible prior null probability consistent with associating rejection at the usual 5% level with \( P(H_0 \mid D) < 0.05 \).
As argued in the previous two sections, there are both statistical and economic arguments against a low prior probability assessment of the null. Thus, classical hypothesis testing involves a paradox: the choice of a low significance level like 5 percent is due to the null being our maintained hypothesis that requires strong evidence against it in the data to be rejected, but, in fact, that same significance level has an implicit prior probability assessment that favors the alternative hypothesis! To avoid the paradox, an even stricter significance threshold is needed. Explicit new thresholds are presented below.

The above calculation of \( P(H_0 \mid D) = 0.119 \) for a \( t \)-statistic of 2.0 is based on conditions that resemble as much as possible the classical ideal of ‘objectivity’ (neutral prior odds: \( P(H_0) = P(H_1) = 0.50 \)) and a Bayes factor (\( BF \)) which is identical to the classical likelihood ratio. As seen, under these conditions the Bayesianized \( p \)-value is substantially higher than the classical \( p \)-value. But, in fact, Bayesian statisticians generally consider this particular \( BF \) too extreme and instead recommend Bayes factors that do not so strongly favor the alternative hypothesis.

One choice of \( BF \) that has gained general acceptance is based on the Bayesian Information Criterion (BIC) suggested originally by Gideon Schwarz (1978). BIC is widely used, also among classical statisticians and econometricians, as a model selection tool. BIC implies a distribution for the parameter under \( H_1 \) that can be approximated by a normal distribution centered around the maximum likelihood estimate and with a relatively large variance (see Raftery 1995; Kass and Raftery 1995). Thus, BIC is appealing if we have some, but not a very precise, idea of the range of variation for the parameter. With, again, \( t \) being the usual \( t \)-statistic and \( n \) the sample size, \( BIC = \log(n) - t^2 \), and the Bayes factor becomes \( BF = \exp(\frac{1}{2}BIC) \).

Table 3 reports \( BF \) based on BIC, and the associated Bayesianized \( p \)-value, \( P(H_0 \mid D) \), for a \( t \)-statistic of 2.0, various sample sizes, and the same prior null probabilities as in Tables 1 and 2. For a relatively small sample of \( n = 50 \) observations, the Bayes factor becomes \( BF = 0.957 \) which means that the data are slightly less likely under the null than under the alternative hypothesis. Thus, the data leads to only a very small downward revision of the probability that the null is true, and the Bayesianized \( p \)-value is substantially higher than the classical \( p \)-value of 0.05.

| \( n \) | \( BF \) | \( P(H_0 \mid D) \) | \( P(H_1 \mid D) \) | \( P(H_0 \mid D) \) |
|---|---|---|---|
| 50 | 0.957 | 0.096 | 0.131 | 0.252 |
| 100 | 1.353 | 0.242 | 0.311 | 0.502 |
| 500 | 3.026 | 0.489 | 0.575 | 0.752 |
| 750 | 0.742 | 0.802 | 0.901 |
| 900 | 0.896 | 0.924 | 0.965 |
When the sample size increases, the divergence between Bayesianized and classical $p$-values also increases and $BF > 1$ implies that the data lead to an upward revision of the probability that the null is true, a dramatically different conclusion than the rejection of $H_0$ that most researchers will do when seeing a $t$-statistic of 2.0.\textsuperscript{14}

If, as above, researchers (implicitly) take a rejection at the 5% level to imply that $P(H_0 \mid D) < 5\%$, we can reverse engineer the BIC based $BF$ to derive the implicit prior null probability. For example, for $n = 100$, we obtain (using the equation in footnote 13) $P(H_0) = 0.037$ and $P(H_1) = 0.963$, which are clearly not neutral. Alternatively, with neutral prior odds, $P(H_0) = 0.50$, we can ask what the $t$-statistic threshold needs to be in order for $P(H_0 \mid D) < 0.05$. The result is $t = 3.24$, corresponding to a $p$-value threshold of 0.0012; for $n = 500$, the required $t$-statistic is $t = 3.48$ and the $p$-value threshold is 0.0003. As seen, these thresholds are even stricter than the new 0.005 $p$-value threshold suggested by Benjamin et al. (2018). From a Bayesian perspective, the conclusion is that the conventional 5% significance threshold is a very low hurdle rate to pass in order to declare a significant finding; it needs to be raised markedly. Alternatively, and perhaps even better, statistical testing with fixed thresholds should be abandoned altogether. I discuss this in the concluding remarks.

**Concluding remarks**

In this paper I have argued that, for both statistical and economic reasons, when economists apply statistical hypothesis tests in their empirical work, and the substantive economic model appears in the alternative hypothesis, they should generally—explicitly or implicitly—put substantial prior probabilistic weight on the null hypothesis, in contrast to what many economists do in practice. A follow-up recommendation is that the conventional significance or $p$-value threshold should be lowered from the current 5% level. A natural consequence of applying a stricter threshold is to reduce the high false discovery rate that currently haunts empirical research.

However, transitioning from a $p$-value threshold of 0.05 to, e.g., 0.005 (as suggested by Benjamin et al. 2018) will—all else equal—lead to fewer null rejections and thereby more real effects not being discovered (lower test power), i.e., increase the false non-discovery rate. To alleviate this problem Benjamin et al.

\textsuperscript{14} As seen in Table 3, for a given prior null probability, $P(H_0 \mid D)$ increases when $n$ increases. In the statistics literature this is called the Jeffreys-Lindley paradox (Jeffreys 1961; Lindley 1957) that implies that for a fixed but arbitrarily high $t$-statistic, $P(H_0 \mid D)$ approaches one when $n$ goes to infinity.
(2018) propose—for experimental studies—to generally increase the sample sizes so as to maintain the conventionally required 80% power.

In economics and most other social sciences, however, empirical studies are usually based on non-experimental, passively observed data where increasing the sample size to obtain sufficiently high power will often be impossible. In some fields with an abundance of data, e.g., high-frequency financial data or register-based micro-data, power is in any case close to 100% and setting a stricter significance or p-value threshold should be unproblematic. But in analyses with small or moderately sized samples of non-experimental data, just raising the hurdle rate for significance will naturally lead to more missed discoveries.

The heretical question is whether the strong reliance on the null hypothesis significance testing (NHST) paradigm based on t-statistic or p-value thresholds, which continues to characterize empirical research, should finally be dropped. An increasing number of statisticians and scientists move in that direction, saying “abandon statistical significance” (McShane et al. 2019) or “retire statistical significance” (Amrhein et al. 2019), and suggesting “moving to a world beyond ‘p < 0.05’” (Wasserstein et al. 2019). In addition to the bad empirical practice that strong adherence to the NHST paradigm leads to, there is a growing recognition of the many paradoxes and built-in contradictions involved in this paradigm. Fisher’s ‘significance testing’ based on p-values and the Neyman-Pearson ‘hypothesis testing’ approach based on a cost-benefit analysis of Type I and II errors, are in fact incompatible (Hubbard and Bayarri 2003; Schneider 2015), a fact that is becoming more known but still not generally recognized among empirical researchers, and applied statistics and econometrics textbooks still often present statistical testing as a hybrid Fisher-Neyman-Pearson theory.

For economics and the social sciences in general, where passively observed non-experimental data are prevalent, and where empirical models are to be considered very crude approximations to reality, the NHST paradigm is particularly problematic. In the social sciences the data are often ‘convenience samples,’ and new samples from the population cannot be drawn. In fact, the underlying population is often not well-defined and researchers then need to rely on imaginary ‘super-populations’ in order to apply the traditional measures of statistical uncertainty (sampling error) based on the sampling distributions of statistics in (hypothetical) repeated sampling. There are a bunch of problems involved in applying the traditional statistical paradigm in such cases (Berk et al. 1995; Engsted

---

15. One could, for example, set a new threshold based on BIC and neutral prior odds for the hypotheses, and such that rejection occurs when $P(H_0 \mid D) < 0.05$, cf. the “Is p-value < 0.05 a sufficiently strict hurdle rate?” section above. With, e.g., a sample size of $n = 100,000$ observations, this leads to a t-statistic threshold of 4.17, corresponding to a p-value threshold of 0.00003.
and Schneider 2024), problems that are either ignored or not recognized by most empirical social science researchers. In essence, the problem with the traditional approach is that it focuses on sampling error and ignores model uncertainty, where the latter is what is important for most social science studies.

In empirical economics, alternative measures of fit not based on statistical significance have been developed since the 1980s (see Engsted 2002 for a survey of such measures). These measures focus on economic significance instead of statistical significance. For example, in financial asset pricing, measures of the economic magnitude of pricing errors of particular asset pricing models exist, but, unfortunately, in this literature many published papers continue to put more emphasis on whether pricing errors are statistically significant.

What is needed is a general recognition among empirical researchers that statistical significance at arbitrary significance levels should no longer serve as a de facto necessary condition for declaring a scientific result because it leads to an unfortunate preoccupation with passing certain statistical thresholds independent on the economic context and the nature of the data. Instead, focus should be on the economic magnitude and significance of estimated effects, the robustness of the results to changes in the data and variable definitions, functional form, estimation methods, etc., and on model uncertainty in general.

**Appendix A.**

**Derivation of the FDR in Equation (1)**

Using Bayes’ formula we can write the probability that \( H_0 \) is true, given that \( H_0 \) is rejected, as:

\[
P(H_0 \text{ is true} \mid H_0 \text{ is rejected}) = \frac{P(H_0 \text{ is rejected} \mid H_0 \text{ is true})P(H_0)}{P(H_0 \text{ is rejected})}
\]

and similarly:

\[
P(H_1 \text{ is true} \mid H_0 \text{ is rejected}) = \frac{P(H_0 \text{ is rejected} \mid H_1 \text{ is true})P(H_1)}{P(H_0 \text{ is rejected})}.
\]

Dividing the first with the second expression gives

\[
\frac{P(H_0 \text{ is true} \mid H_0 \text{ is rejected})}{P(H_1 \text{ is true} \mid H_0 \text{ is rejected})} = \frac{P(H_0 \text{ is rejected} \mid H_0 \text{ is true})P(H_0)}{P(H_0 \text{ is rejected} \mid H_1 \text{ is true})P(H_1)} = \frac{\alpha}{1 - \beta} \frac{P(H_0)}{P(H_1)}.
\]
where $\alpha$ and $\beta$ are the Type I and II error probabilities, respectively. Then imposing
that probabilities add up to one such that $P(H_1) = 1 - P(H_0)$ and $P(H_1$ is true $|$ $H_0$ is rejected) = $1 - P(H_0$ is true $|$ $H_0$ is rejected), Equation (1) is obtained.

**Appendix B. Derivation of Equation (3)**

Using Bayes’ formula we can write the probability that $H_0$ is true, given that $H_0$ is not rejected, as:

$$P(H_0 \text{ is true } | H_0 \text{ is not rejected}) = \frac{P(H_0 \text{ is not rejected } | H_0 \text{ is true})P(H_0)}{P(H_0 \text{ is not rejected})}$$

and similarly:

$$P(H_1 \text{ is true } | H_0 \text{ is not rejected}) = \frac{P(H_0 \text{ is not rejected } | H_1 \text{ is true})P(H_1)}{P(H_0 \text{ is not rejected})}.$$

Dividing the first with the second expression gives

$$\frac{P(H_0 \text{ is true } | H_0 \text{ is not rejected})}{P(H_1 \text{ is true } | H_0 \text{ is not rejected})} = \frac{P(H_0 \text{ is not rejected } | H_0 \text{ is true})P(H_0)}{P(H_0 \text{ is not rejected } | H_1 \text{ is true})P(H_1)} = \frac{1 - \alpha P(H_0)}{\beta P(H_1)},$$

where $\alpha$ and $\beta$ are the Type I and II error probabilities, respectively. Then imposing
that probabilities add up to one such that $P(H_1) = 1 - P(H_0)$ and $P(H_1$ is true $|$ $H_0$ is not rejected) = $1 - P(H_0$ is true $|$ $H_0$ is not rejected), Equation (3) is obtained.

**Appendix C. Derivation of Equation (4)**

Bayes’ formula implies that

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_0)}{P(D \mid H_1)} \cdot \frac{P(H_0)}{P(H_1)} = \frac{BF}{P(H_1)}.$$

Imposing that probabilities add up to one such that $P(H_1) = 1 - P(H_0)$ and $P(H_1 \mid D) = 1 - P(H_0 \mid D)$, Equation (4) is obtained.
References


Harvey, Campbell R. 2017. Presidential Address: The Scientific Outlook in Financial Eco-


Wasserstein, Ronald L., and Nicole A. Lazar. 2016. The ASA’s Statement on P-Values: Context, Process, and Purpose. \textit{American Statistician} 70(2): 129–133. \texttt{Link}

Wasserstein, Ronald L., Allen L. Schirm, and Nicole A. Lazar. 2019. Moving to a World Beyond “P < 0.05.” \textit{American Statistician} 73(sup1): 1–19. \texttt{Link}
About the Author

Tom Engsted is professor of economics at University of Aarhus, Denmark. He has numerous publications within economics, financial economics, and econometrics, in journals such as *Journal of Money, Credit, and Banking*, *Review of Economics and Statistics*, *Journal of Financial and Quantitative Analysis*, and *Review of Financial Studies*. His email address is tengsted@econ.au.dk.
Revisiting Hypothesis Testing with the Sharpe Ratio

Michael Christopher O’Connor

LINK TO ABSTRACT

If you manage a portfolio, or if you are invested in a portfolio that is managed by others, such as a fund, you would like it to show a high Sharpe ratio on an ongoing basis. The statistic is a popular device for assessing and rating portfolios and their managers. The Sharpe ratio is the mean of an investment’s per-period returns in excess of the returns on cash, divided by the standard deviation of those excess returns, during some interval of time that has been divided into periods (e.g., months).\(^1\)

Perhaps the ratio owes much of its popularity to the fact that it combines two familiar statistics that are each of great interest to investors, the mean and the variance (the square of the standard deviation), into one. Combining the two familiar statistics suggests a simple rating scheme whose use, in selecting investment alternatives, seemingly honors both high cash-beating returns and low variance.

But of course, the wisdom of basing investment decisions upon a single performance measure is doubtful. Here I focus on comparing the performance of one investment alternative with the performance of another—using the paired difference between the Sharpe ratios of the two portfolios as the performance measure.

When the Sharpe ratio of one portfolio substantially exceeds that of another there is a need to consider whether the exceedance is likely to be sustained. Did

\(^1\) If \(P_i\) is the value of the investment at the end of period \(i\) (which may consist of the market value of the asset and income that the asset has generated), and \(C_i\) is the value at the end of period \(i\) of cash held by anyone during period \(i\), then the investment’s excess returns of period \(i\) can be taken to be \(\ln(P_i / P_{i-1}) - \ln(C_i/C_{i-1})\), where \(\ln()\) is the natural logarithm.
it merely happen by chance? The statistical properties of the ratio that delimit the proper use of the test of the happened-by-chance hypothesis are hardly as widely understood and appreciated as they should be. There are entire categories of studies of Sharpe ratio differences whose authors, having disregarded limitations on the power of the test, reached erroneous conclusions regarding statistical significance. I demonstrate that. But first I review oft-cited mathematical publications about hypothesis testing with Sharpe ratio differences, publications that contain peculiar errors and omissions that may account for the deficiencies in understanding. I present a legitimate means of post hoc analysis of hypothesis test findings, for sorting out misclassifications of statistical significance. And I show, categorically, using simulations with pertinent Sharpe ratio difference examples and other analysis, that when the power of the test is low then the very best estimators that can be put to the task of determining statistical significance can hardly perform better than random number generators.

I typically have in mind one of the portfolios being a suitable benchmark—e.g., if the investment alternative of interest involves long positions in big-cap stocks, a suitable benchmark would be one that would replicate the returns of the S&P500® index. The burning issue then becomes the question of whether past outperformance of the benchmark by the investment alternative of interest is likely to be sustained. The hypothesis that said outperformance was due to chance, and not likely to be sustained is the null hypothesis. And it is that hypothesis that is tested. The hope is to be able to reject it.

And to do so reliably the probability of rejecting the null hypothesis when it is false, which is called the power of the test, must be high. The statistical properties of Sharpe ratio differences differ substantially from those of Sharpe ratios. That’s the rub. With differences, it’s far easier to find yourself in circumstances in which the power of the test is too low for hypothesis testing to be conducted.

It is therefore necessary to be especially wary when wielding the Sharpe ratio difference hypothesis test. And that’s really what this article is about.

The Sharpe ratio difference hypothesis test is the subject of the oft-cited Olivier Ledoit and Michael Wolf (2008) article “Robust Performance Hypothesis Testing with the Sharpe Ratio.” Although the authors seem to have produced a “robust” way of performing hypothesis testing, within the technical meaning of that word, they: (1) utterly fail to discuss the crucial matter of the power of the test; (2) improperly demonstrate the use of their method in a circumstance in which the power of the test is obviously too low to offer reliable hypothesis testing; and (3) claim superiority over competing methods when the number of observations is low, despite the obvious fact that the power of the test is too low for hypothesis testing when the number of observations is low.

Ledoit and Wolf (2008, 851) describe themselves as statisticians and refer to
others who might be interested in their work as financial practitioners. The article has been influential, so that there are users who are seemingly not as mathematically capable as the authors who have made the gross error of following suit and applying the method when the power of the test is obviously too low for use. Herein I deal with that and how it came about.

The very next section below reviews the basics of hypothesis testing and can be skipped by readers who are well versed on the subject. Further on, the section “The power of the test in practice” contains a discourse on post hoc analyses of hypothesis testing, which has been a controversial topic. In that section the use of a secondary hypothesis test is proposed, to better cope with errors that occur with findings of statistical insignificance when the adequacy of the power of the test is open to question. I have included a simple method of incorporating investment horizons when using Sharpe ratio differences, which you will find in the section “The recourse to confidence intervals.” Then further on, in the section “The last word,” I thoroughly demonstrate, using the very code that was provided by Ledoit and Wolf (2008), that it is imprudent to base portfolio selection on hypothesis testing when the power is truly low. Those are highlights. Complete derivations are provided for all of the mathematics, either directly in this article in easy-to-follow steps or straightforwardly in cited articles.

The mathematics of hypothesis testing

The operative null hypothesis when the test statistic is the difference between the Sharpe ratios of two portfolios is simply the claim that it will not be positive in the long run: Its true value is zero or negative. If the null hypothesis is true, then any positive difference that was computed from a sample of historical data was due to chance—merely analogous to ten tosses of a fair coin happening to turn up heads six times instead of five. To assess the risk that the measured difference is due to chance, one computes, using sampled historical data, the highest probability that the difference would be equal to or greater than the measured value that could be computed if the null hypothesis were true. That probability is called the ‘p-value.’ Low p-values, below some fixed cutoff value that the analyst is free to choose, indicate that it is likely that the null hypothesis isn’t true and that it should be rejected. When it is thus rejected, ‘statistical significance’ is declared. Would that it were that simple.

2. I’m limiting the present discussion to my preferred form of the hypothesis test, which is referred to as a ‘one-sided’ or ‘right-tailed’ test, because when the stated null hypothesis is rejected, the conclusion is that the value of the Sharpe ratio difference is almost certainly positive, to the right of zero. With either a right-sided or two-sided form of the test, the highest probability that the test statistic would be equal to or greater than the measured value if the null hypothesis were true is computed by taking the true value to be zero.
The first complication involves the chosen cutoff value of the p-value, which is called the ‘significance level.’ Given the definition of the p-value, it is the probability of wrongly declaring statistical significance when the null hypothesis is true. Doing that is dubbed a ‘type I’ error or a ‘false discovery,’ and to avoid such errors the significance level is set rather low (e.g., 0.05)—but not so low as to fix it so that there would seldom be findings of statistical significance. So, already it’s not so simple.

Then it takes some doing to compute the p-value. There is a head-first dive into that further on below. But it is equally imperative to heed another probability that is on the flip side of null hypothesis testing— the probability of erring by failing to reject the null hypothesis (due to the p-value being above the cutoff value) when the null hypothesis isn’t true. And 1.0 minus that probability is the probability of rejecting the null hypothesis when it isn’t true, the ‘power of the test.’

Whereas the p-value can be estimated using the sample of historical data, the power of the test is based upon the true value, called the population value by statisticians (and, hereafter, in this article), of the test statistic. And that value can never be known. The goal is to find out as much about it as possible, such as bounds within which it is likely to be found. And mere hypothesis testing does just that, because a finding of statistical significance is evidence that the test statistic is probably not negative, which means that zero has been shown to be a lower bound of sorts.

The population value of the chosen hypothesis test statistic is the most important of the parameters on which the power of the test depends. But surprisingly, when the chosen test statistic is the Sharpe ratio difference, there is a very strong dependence of the power upon the correlation coefficient between the returns of the two portfolios. With the Sharpe ratio difference, factors that bring about inadequate power include the population value of the difference being too small, the correlation coefficient being low, and the time duration of the sample being small.

Certainly high power is wanted, so that the null hypothesis will usually be rejected when it isn’t true. But exactly what sort of ill wind is it that blows if the power is low, not anywhere near 1.0? When there is a failure to reject the null hypothesis when it isn’t true, a ‘type II error’ has been sustained. The immediate harm is that the investor is thereby led toward failing to invest in the portfolio with the seemingly better Sharpe ratio, even though it is destined to continue to outperform the portfolio with which it was compared. That other portfolio with the inferior Sharpe ratio could be a benchmark, such as a stock market index portfolio. Missing out on beating the market certainly involves some kind of opportunity cost. But there’s more.

The low power can otherwise increase the likelihood of committing the error
of rejecting the null hypothesis when it is true, a ‘type I error’ or a ‘false discovery.’

How so? Consider an analyst who wants to go forward with an investment alternative that is found to be benchmark-beating, based upon Sharpe ratios compiled from sample returns of the alternative and of the benchmark. A single alternative, when considered alone with its benchmark as a Sharpe ratio difference, has a finite chance of generating a type I error. What happens if, owing to a type II error that is brought about by the low power of the test, the analyst is misled into having to sift further through several alternatives, looking for one whose benchmark beating was calculated to be statistically significant? The analyst encounters, with each alternative considered, the risk of a type I error. The overall probability of settling upon a false discovery is thereby enhanced because of the multiple tries.³

But even if the search doesn’t finish with a type I error there is still a problem that is brought about by having to search for an alternative with statistically significant outperformance. Seeking statistical significance (that was denied due to the low power of the test causing type II errors) is, in the main, seeking a high sample value of the Sharpe ratio difference because the two go together. And selecting the alternative that performs best on sampled data is a formula for inducing selection bias: Any preference for the alternative to have scored high in performance on the sample is automatically also a preference for chance outliers, for alternatives whose performance on the sample is considerably better than what can be expected in the long term. The Sharpe ratio differences of such alternatives may have population values that are no better than those of alternatives that were passed over because of type II errors. Cross-validation methods and walk-forward methods can be used to counter selection bias, but they have their limitations.

Relatedly, very low power can even diminish the probability, when there is immediately a positive finding of statistical significance, of the true Sharpe ratio difference actually being positive. This is explained by Katherine Button, John Ioannidis, et al. (2013, 2), building on some simple mathematics from Ioannidis (2005, 696–697).⁴

³. Remediation is available for this problem. See for example Benjamini and Yekutieli (2001, 1168–1169), wherein the word “conservative” means that the method errs on the side of dismissing more findings of significance than would be necessary to control the false discovery rate. It is still advisable to do the remediation.

⁴. Ioannidis (2005) finds that the probability that a finding of statistical significance is really true (that the population value of the test statistic is positive), which has been dubbed the positive predictive value (PPV), depends upon the product of the power of the test and the odds ratio being high enough. Loosely put, the odds ratio is the expected frequency of good performance outcomes divided by the expected frequency of bad performance outcomes. Often financial analyses involve circumstances in which the odds ratio might as well be estimated as being ≈ 1. For example, if an investor sets out to examine the past performance of a no-load mutual fund, one whose holdings seem to be of much the same character as those
Reviewing Sharpe ratio difference literature

Circa 2021, I was innocent of knowing anything of substance about the use of a Sharpe ratio difference as a test statistic. As for other test statistics involving financial time series data, I did at least know that the gold standard was that an analysis that properly establishes statistical significance is one that allows for investment portfolio returns not being independently and identically distributed (i.i.d.), having fat-tailed and skewed distributions, and being autocorrelated so that the current month’s return is influenced by the returns of prior months. ‘Heteroskedasticity’ denotes the circumstance of the variance of a statistic not being constant but varying with time. So yes, there is a need to try to allow for that too—a varying variance. I searched the literature and I found Ledoit and Wolf’s article (2008). It is all about computing the p-value when the test statistic is the difference between the Sharpe ratios of two portfolios. I had found the answer.

Or so I thought. Ledoit and Wolf (2008) have open-sourced their code (found here), in the computer languages Matlab and R, and their article has been cited something like 1,000 times. Thus their procedures and the code that they wrote to implement them have been well scrutinized and used. I opted for extending their checks on the validity of their code just a bit, in a relevant way. Their tests involved progressively more realistic simulated histories of portfolio returns but, unlike mine, their tests were limited to the case when the Sharpe ratio difference is actually zero (Ledoit and Wolf 2008, 857). I tested only with simulated returns randomly drawn from the not very realistic bivariate-normal distribution, so that the sampled values are i.i.d. My simple tests are of interest, notwithstanding the limitation to idealized circumstances, because the imposed limitation doesn’t leave us with just simulations to work with. The idealized circumstances make it possible to straightforwardly derive asymptotically valid expressions for the p-value and the power of the test. And I did not limit my tests to Sharpe ratio differences of zero, because only non-zero differences are of interest when hypothesis testing. Apart from the very small effects of the involved iterations not being infinite in number, of the benchmark, then probably the odds ratio is about 1. Low power tends to drive the PPV below 1, but when the odds ratio is not much below 1 the power of the test must be very low if the PPV is to be driven much below 1.

But, if the investor is considering a mutual fund with a substantial load or is considering an investment advisor of the kind that are affiliated with brokers, who professes skill and consequently charges an annual fee of, say, 0.5 to 1.0 percent, then the odds ratio may be drastically lower than 1. That’s with the mutual fund’s load or the advisor’s fee being fully accounted for in the performance measure that is the test statistic, thus lowering the odds ratio. In that circumstance the deleterious effect of low power on the PPV is greatly magnified by the odds ratio multiplier being low.
I found no disagreement between the p-values that I computed with Ledoit and Wolf’s code and those that I computed with either my own simulations or with the asymptotic expressions that were derived using calculus.

In the introduction to this article I delineated the wholesale and consequential neglect of the power of the test in Ledoit and Wolf (2008). With the goal of understanding how it might have happened that the 2008 article won acceptance despite its fairly obvious deficiencies, I read prior articles that Ledoit and Wolf had cited, by J. D. Jobson and Bob M. Korkie (1981), Christoph Memmel (2003), and John D. Opdyke (2007).

And indeed, Jobson and Korkie (1981) were sowing confusion. Their paper contains three glaring errors: One is their claim that they had concocted an alternative to the Sharpe ratio difference statistic that had more power as a test statistic than the Sharpe ratio difference; the second was to state an incorrect formula for the asymptotic variance of their own statistic; and the third was to state that the power of the test is chronically low and insensitive to the value of the correlation coefficient between the returns of the two portfolios. In only three pages, Memmel (2003) corrected the first two errors; leveraging off of Memmel’s work, Opdyke (2007) disclosed and fixed the third error, showing that when the correlation coefficient is high the power of the test is high (but also affirming that when it is not high the power is low).

But Ledoit and Wolf mention Opdyke’s (2007) contribution without noting that he affirmed that a low correlation coefficient means that the power of the test is very low. The inappropriate use that Ledoit and Wolf made of their method was to compute the p-value for the paired Sharpe ratio difference between two hedge funds, using a decade of data. The authors provided the returns of the funds and I have calculated the correlation coefficient between the returns of the two hedge funds as 0.00—meaning, in conjunction with the use of only a decade of data, that the power was unusably low. That the power was too low is confirmed below, at the conclusion of the “Hedge funds are different” subsection of the section “The power of the test in practice.”

Citations of these articles since the start of 2009 have been as follows: Ledoit and Wolf (2008) 981 times, Jobson and Korkie (1981) 794 times, Memmel (2003) 450 times, but Opdyke (2007) just 180 times. Those who see in Ledoit and Wolf (2008) that Jobson and Korkie (1981) and Memmel (2003) are cited, who are diligent enough to obtain and read those articles but missed Opdyke (2007), would really have to be on their toes when figuring out what to think about Jobson and Korkie’s power estimates because of all of their mistakes.

Furthermore, anyone who did read Opdyke (2007) would have been grossly

5. Google Scholar citations as of May 13, 2023.
misled by the author’s discussion of the power of the Sharpe ratio difference test being confirmed, by his analysis of mutual fund data, to be high when the correlation coefficient is high. Opdyke well knew, from his derivations, that the power can be high when the coefficient is high if the population value of the Sharpe ratio difference is not too small and if the duration of the historical data is sufficient. But in circumstances where the portfolios being compared are the result of very similar portfolio formation schemes, the differences in performance are entirely unforeseeable and for numerous pairs the true Sharpe ratio difference is small. And not only were the funds that Opdyke studied of that character, but he had only three years of data. The power of the test must be presumed to be low in those circumstances, whatever the value of the correlation coefficient. The relevant mathematics are reviewed in the next section.

These circumstances have led to numerous researchers using Ledoit and

---

6. Opdyke wielded the right-sided hypothesis test. He formed 190 ordered pairs out of the 20 mutual funds—every possible ordered pair, where ‘ordered’ simply means that the sample value of the Sharpe ratio of the first fund of each pair was bigger than that of the second. He seems to have maintained that the fact that the four or five out of the 20 funds that had the highest sample values of the Sharpe ratios ranked above about half of the other funds by statistically significant margins was indicative of the test generally having good power, because low power would mean that not many null hypotheses would have been rejected, whereas (of the pairs formed with the top four or five funds) about half were rejected. But on his Table IV we see that there was failure to reject the null hypotheses of 122 of the 190 pairs. Many of the other pairs, having small Sharpe ratio differences, simply didn’t have adequate power.

How do I know that many pairs did not have adequate power? Well, it’s obvious that the 20 mutual funds having very similar methods of portfolio formation within the same asset class rules out the true Sharpe ratio differences of the pairs being somehow fenced away from zero. If say, someone were to compare the performance of 20 hedge funds with that of a benchmark portfolio that tracked the S&P500®, then yes, it should be expected that the hedge funds would have higher Sharpe ratios than the benchmark—meaning that the distribution of the Sharpe ratio differences of the 20 hedge funds would be centered on some positive value of the Sharpe ratio difference.

But that’s hardly the case with pairs of similar mutual funds. Ordered pairs could be formed based on the population values of the Sharpe ratios (if only we knew those values), instead of on the sample values. Every pair has a population value of the Sharpe ratio difference that is the main determinant of the power of the test in application to the pair. Suppose that we sort the list of funds in descending order by the population Sharpe ratio. Then the Sharpe ratio difference between the top fund and the fund immediately below will be small, because they are only one place apart on the list. The same will be true of the pair that consists of the funds that are placed second and third on the list, and so on. There will be 19 such pairs with small Sharpe ratio differences due to the funds being only one place apart on the list. What about pairs of funds that are two places apart? Yes, they will have bigger Sharpe ratio differences but there will be 18 of them, not 19. We can continue. When the funds are three, four, five places apart, etc., the Sharpe ratio differences will be getting larger still, but the number of pairs will be getting reduced (e.g., with the funds 10 places apart there will be 10 such pairs with large differences). The point is that the distribution of the population values of the Sharpe ratio differences would be peaked near zero. And for those fairly numerous pairs with differences near zero the power of the test will be too low.

That low Sharpe ratio differences always mean low power, however large the correlation coefficient, is demonstrated in the next section of this article. It’s also shown there that Opdyke using weekly data instead of monthly data in his three-year backtest interval did not improve the power.
Wolf’s (2008) article and code to calculate p-values for use with hypothesis testing, when the test statistic is the difference between two Sharpe ratios (or the absolute value of that), without understanding that in many commonly encountered circumstances the power of the test is simply way too low.

**Mathematics of power—i.i.d. bivariate-normal returns**

Here the mathematics of how to compute p-values and the power of the test whose statistic is either the difference between the Sharpe ratios of two portfolios or the absolute value of that difference is summarized, with some of the analysis being specific to the case in which the returns of the two portfolios are i.i.d. bivariate-normal. One of the portfolios may be a benchmark.

**Dependencies of the power of the test**

If $S_h$ is the Sharpe ratio of the portfolio under study, and $S_h^b$ is the Sharpe ratio of a suitable benchmark for it (or some other portfolio of particular comparative interest), then four test statistics will be considered: $S_h$, $|S_h|$, $S_h - S_h^b$, and $|S_h - S_h^b|$. The statistics that involve absolute values bring about two-sided tests; the other statistics establish one-sided, right-sided tests.

Figure 1 plots the power of the test versus the Sharpe ratio $S_h$ of the portfolio under study, when the test statistic is any one of those four statistics—for various values of the number $T$ of observations and of the correlation coefficient $\rho_{ab}$ between the portfolio and the benchmark (when applicable). The dotted lines pertain to the two-sided test whose statistic is $|S_h - S_h^b|$; the dashed lines pertain to the two-sided test whose statistic is $|S_h|$. The thin gray lines that conform to the dotted and dashed lines at elevated values of the Sharpe ratio $S_h$ of the portfolio under study plot the right-sided powers when the test statistic is $S_h$ or $S_h - S_h^b$. And the figure pertains to the circumstance of $S_h^b = 0.10$ (when applicable).

Figure 1 is valid as-is, for any frequency of observations—e.g., daily, weekly, monthly, or yearly. If for example the raw data were monthly with the Sharpe ratios being computed monthly as well, then the annualized Sharpe ratios could be approximated as the monthly ratios multiplied by $\sqrt{12}$. Thus, where the benchmark’s Sharpe ratio $S_h^b$ is taken to be 0.10 on the figure, that would correspond to an annualized Sharpe ratio of about 0.10$\sqrt{12} \approx 0.35$—similar in

---

7. Lo (2002) cautions that if the returns are autocorrelated then the factor of $\sqrt{T}$ is not accurate.
magnitude to what we might expect of the long-term average Sharpe ratio of some capitalization-weighted portfolio of stocks belonging to some stock market index. The fixed value of the benchmark portfolio’s monthly Sharpe ratio on the figure, of 0.10, was also the choice of Jobson and Korkie (1981) in almost all their examples. The shown number of months, especially T=120 and T=360 on the top half of the figure, would then respectively represent often-achievable and sometimes-achievable backtesting intervals of 10 years and 30 years. 8

Figure 1. Dependencies of test power with Sharpe ratios and Sharpe ratio differences as test statistics

Notes: The dotted lines and the thin gray lines that conform to them as the portfolio Sharpe ratio $S_a$ is increased respectively pertain to the hypothesis tests that are based on the statistics $|S_a - S_b|$ and $S_a - S_b$, where $S_b$ has been set at 0.10. The dashed line and the thin gray line associated with it respectively pertain to the test statistics $|S_a|$ and $S_a$ that involve just the portfolio Sharpe ratio.

8. But if instead the portfolio under study were a hedge fund, with the Sharpe ratios of it and a benchmark forming a difference, then if the data were weekly the benchmark’s Sharpe ratio $S_b$ of 0.10 on the figure would very roughly become $0.10 \sqrt{52} = 0.72$. That would not be an unreasonable value for the annualized Sharpe ratio of a hedge fund. Then the shown number of weeks, T=720 and T=1200 on the bottom half of the figure, would respectively represent often-achievable and sometimes-achievable backtesting intervals of something like 14 years and 23 years.
I hasten to add that the Sharpe ratios in Figure 1 are not to be thought of as sample-derived values. The ratios are in fact the population values. It can be imagined instead that data have not yet been acquired but significance testing for some portfolio formation scheme that has been contrived is being planned. And one of the burning questions would be: How many months of data are needed? If the analyst were to have some idea of the population values of the ratios, even if it were just a rough idea, or even if the analyst were proceeding on a ‘what if’ basis, Figure 1 (or better yet the code that produces it) could be looked to for some helpful guidance concerning the very feasibility of hypothesis testing with the Sharpe ratio.

The returns of the two portfolios are assumed to be i.i.d. and drawn from a bivariate-normal distribution, as in the Jobson and Korkie (1981) article. The figure directly implements Memmel (2003)’s correct asymptotic expression for the ‘standard error’ of the Sharpe ratio difference. I present the math and the mathematical terms in the next subsection.

The horizontal axes pertain to the Sharpe ratio of the portfolio under study; the power is plotted vertically. The distinctions between the four kinds of power that are illustrated in the figure have been drawn above, where four test statistics were described. There was also imposed on the analysis that produced Figure 1 the value \( \alpha = 0.05 \), which is the ‘significance level,’ which was introduced above. Again, the analyst gets to choose it! If the test is two-sided, then if the two-sided p-value exceeds \( \alpha \) there is failure to reject the null hypothesis. If such an \( \alpha \) has been chosen for a two-sided test, then for purposes of side-by-side comparison with a one-sided test there is failure to reject the null of the one-sided test if the one-sided p-value exceeds \( \alpha / 2 \). So \( \alpha \) or \( \alpha / 2 \), depending upon the sidedness of the test, is the probability of rejecting the null hypothesis if it is true (because the p-value is based on the null hypothesis being assumed to be true). Again, that would be a type I error, a false discovery.

That sidedness lingo is something that you will encounter in academic articles that present statistical tests based on Sharpe ratios. And it may be important for a reader to take note of which kind of test has been used, because whereas the reader might be quite understandably concerned about the sign of the Sharpe ratio or Sharpe ratio difference, rather than just its magnitude, there are circumstances in which academics prefer to implement a two-sided test which could lead to a finding of statistical significance with the Sharpe ratio or Sharpe ratio difference being of either sign. For example, there might be a factor that influences the makeup of one of the portfolios that the academician suspects of being incapable of producing a substantial effect. With a two-sided test there are two ways to reject the null hypothesis, so that the two-sided power is always greater, at least a tiny bit greater, than the one-sided power.
A most clearly discernible thing about Figure 1 is that the dotted lines show the power of the test, whose statistic is the absolute value of the difference between the Sharpe ratio of the portfolio under study and that of its benchmark, increasing markedly as the correlation coefficient \( \rho_{ab} \) is increased from \(-0.20\) to \(0.00\) and on up to \(0.80\). That is what Opdyke (2007) noticed and called attention to.

Taking a closer look, suppose that the Sharpe ratio of the portfolio of interest is \(0.15\). That’s \(0.05\) above the benchmark value of \(0.10\), a seemingly substantial improvement: If the frequency of observations is monthly, that would be annualized as the benchmark having a Sharpe ratio of \(0.35\) with the portfolio’s ratio being \(0.52\). So, it might be supposed that the difference of \(0.05\) would be statistically significant just because of that. But looking, say, at the plot on the upper right-hand side that was prepared with the number of observations being \(360\)—three decades of months, which is a fair amount of data—it is seen that the dotted line for the two-sided test when the correlation coefficient \( \rho_{ab} = 0.80 \) indicates that the power is well below \(0.80\) (which is often adopted as a minimally satisfactory power). The power is only \(0.32\), which means that the null hypothesis would be rejected only \(32\) percent of the time when it’s false.

Now it’s true that if \( \rho_{ab} \) the correlation coefficient is allowed to approach \(1.0\), then yes, the power will approach \(1.0\). But if the correlation coefficient is \(0.80\), which is not a low value for comparisons between equity portfolios, how many observations would be needed to achieve acceptable power? Well, if there are \(T = 1,200\) observations, a century of months, the power is \(0.78\)—close enough, but that’s a prohibitively large span of time.

In comparison, starting again with \(T = 360\), what sort of power is achieved when the test is based on the portfolio Sharpe ratio alone, with no consideration given to the benchmark at all? If the portfolio Sharpe ratio is again \(0.15\), the dashed line and the underlying light gray line tell the tale: The power is about \(0.81\)—satisfactory. It’s largely because the tests that only involve the portfolio Sharpe ratio effectively use zero as the point of reference that the power is so much higher than what it would be with the testing being done with the difference from the benchmark’s Sharpe ratio. With tests involving that difference, the point of reference is moved up (all the way up to \(0.10\) on Figure 1). Of course, it’s tougher to beat the higher reference point. And that has implications for the power of the test.

Therefore, correlation coefficients not being high, the duration of the time interval for backtesting not being long enough, and Sharpe ratio differences not being sizeable are controlling factors when Sharpe ratio differences are used as test statistics. Standing alone, any one of them can cause the power of the test to be unacceptably low.
The mathematical framework

Here again the ‘portfolio Sharpe ratio,’ as it is called above in Figure 1 and the discussion of that figure, is $S_h$. And the Sharpe ratio of the other portfolio, which was referred to as a benchmark above, is $S_b$. The other portfolio may not be a benchmark but could be any other portfolio that is of interest. Let $\Delta = S_h - S_b$ and $\Delta = S_h + S_b$. These are not sample values of these quantities but can be thought of as typifying the long-term future performance of the portfolios. It is better to say that they are the population values. The values of these statistics will never be known, and it is only for limited purposes that they can be estimated as being approximately equal to the values that they take on in samples. Sample values are distinguished by $\hat{\Delta}$ and $\hat{\Delta}$ with a circumflex atop the symbols. In this notation the population value of the correlation coefficient is $\rho_{ab}$, but there will be occasion to refer to the sample value $\hat{\rho}_{ab}$ as well.

Equations (1a)–(1c) are straight from Memmel (2003, 23), which contains very clear derivations involving calculus. The term $SE(\Delta, \Delta, \rho_{ab}, T)$ in equations (1a) and (1b) is the standard error, the standard deviation of the distribution of the $\hat{\Delta}$ sample value about the population value $\Delta$. And equation (1c) just announces that in the limit of $T \rightarrow \infty$ the distribution will take on a Gaussian form, a normal form.

$$T \cdot SE(\Delta, \Delta, \rho_{ab}, T)^2 = 2(1 - \rho_{ab}) + \frac{\Delta_+^2(1 + \rho_{ab}^2) + \Delta_-^2(1 - \rho_{ab}^2)}{4} \quad (1a)$$

$$Z = \frac{\hat{\Delta} - \Delta}{SE(\Delta, \Delta, \rho_{ab}, T)} \quad (1b)$$

$$Z \sim N(0, 1) \quad (1c)$$

With either a two-sided or a one-sided test involving a comparison between the Sharpe ratios of two portfolios, when computing the p-value the null hypothesis is implemented by taking $\Delta$ to be 0. Thus equations (1b) and (1c) become equations (2a) and (2b). Only equation (1a) is dependent upon the returns of the

---

9. This assumes that when the null hypothesis is invoked the values of $\Delta$ and $\rho_{ab}$ are known. In this situation those parameters are called ‘nuisance parameters.’ That language refers to the fact that although they are dependencies, they are in fact not known. They are the true values of these parameters, not sample values. It is clear that if $\Delta$ is not large, say under 1.0, then there is only a weak dependence of $SE(\Delta, \Delta, \rho_{ab}, T)$ upon $\Delta$. And that suggests that it might be feasible to simply approximate $\Delta$ by $\hat{\Delta}$. 

---
two portfolios being i.i.d. and bivariate-normal. The other expressions are instead asymptotically valid if only the returns of the two portfolios are stationary, which roughly means that the bivariate return distribution does not change over time; see for example Andrew W. Lo (2002, 39) in that regard.

\[
Z_0 = \frac{\hat{\Delta}_-}{SE(0, \Delta_+, \hat{\rho}_{ab} T)}
\]  
\[
Z_0 \sim N(0, 1)
\]

In equations (3a)–(3c) \( \Phi \) is the cumulative distribution function (cdf) whose derivative is the standard normal probability density function. And \( p_2 \) is the two-sided p-value, with equation (3a) following directly from the definition of the p-value and the meaning of equations (2a) and (2b) regarding the distribution of \( \hat{\Delta}_- \) under the null hypothesis. More simply, \( p_2 = 2\Phi(-|Z_0|) \).

\[
p_2 = 1 - \Phi(|Z_0|) + \Phi(-|Z_0|)
\]  
\[
\pi_2(\alpha) = 1 - \Phi\left(\frac{Z_0}{\Delta_- - \Delta_-, r(\alpha)}\right) + \Phi\left(\frac{-Z_0}{\Delta_- - \Delta_-, r(\alpha)}\right)
\]

\[
\Delta_- r(\alpha) = SE(0, \Delta_+, \hat{\rho}_{ab} T) \cdot \Phi^{-1}(1 - \alpha/2)
\]

Equation (3c) calculates the critical value of the test statistic. If \( |\hat{\Delta}_-| \gtrless |\Delta_-, r(\alpha)| \) then with the two-sided test the outcome is declared to be statistically significant. It is just a re-expression of the definition of the right-sided p-value \( p_r \), which is \( p_r = 1 - \Phi(Z_0) \), evaluated with a particular p-value: Set \( p_r = \alpha/2 \), the critical value of the p-value, solve for \( \Phi(Z_0) \), and then apply the inverse of \( \Phi \). And \( \alpha/2 \) is the cutoff that is mentioned above that the analyst gets to choose. Often the significance level \( \alpha = 0.05 \) is chosen, in which case \( \Phi^{-1}(1 - \alpha/2) \) has the familiar value of \( \approx 1.96 \). Once again, \( \alpha \) and \( \alpha/2 \) are respectively the probabilities, with a two-sided test or a one-sided test, of rejecting the null hypothesis when it is true.

But there is a strong dependence upon \( \hat{\rho}_{ab} \). In Pearson and Filon (1898, 242) the standard error of \( \hat{\rho}_{ab} \) is calculated as being \( (1 - \hat{\rho}^2_{ab}) / \sqrt{T} \). That might suggest that if with a large enough sample of the returns of the two portfolios under consideration they are found to be highly correlated, then it may be assumed that \( \hat{\rho}_{ab} = \rho_{ab} \). In the next subsection it is explained that indeed, that is acceptable.
So, the meaning of equation (3c) is that it sets a critical value for $\hat{\Delta}_-$ that corresponds to setting the critical value $p_r = \alpha/2$. With that understanding $\pi_r(\alpha) = 1 - \Phi\left(Z\hat{\Delta}_- = \hat{\Delta}_-/(\alpha)\right)$, with $Z$ a function of $\hat{\Delta}_-$ as in equation (1b), would be the correct expression for the right-sided power because it is the probability, given the distribution of $\hat{\Delta}_-$ about $\Delta_-$, that $\hat{\Delta}_-$ is bigger than its critical value $\hat{\Delta}_-/(\alpha)$ —causing us to reject the null hypothesis. The $\Phi\left(Z\hat{\Delta}_- = \hat{\Delta}_-/(\alpha)\right)$ term of equation (3b) just adds the left-sided component of the power. And for it $Z$ is evaluated at $\hat{\Delta}_- = -\Delta_-/(\alpha)$ because $\Delta_-/(\alpha) = -\Delta_-/(\alpha)$ due to the symmetric form of the distribution.

Suppose that the data are monthly. If the power, computed with equation (3b) or the right-sided version of it, is too low, then the question arises whether it would help to instead make use of weekly, daily or even intraday data, as that would increase the number of observations. But the increased number of observations would be brought about by an increase in the frequency of observations within the same span of time. So, do the additional observations count the same as getting additional months of data to process? The short answer is no.}

10: The long answer: What exactly changes when the frequency of observations is increased but the duration of the backtesting interval is not? We’ll start with the assumption that the returns are i.i.d. If for example the frequency of observations is increased to daily rather than monthly, the value of the correlation coefficient $\rho_{ab}$ from daily returns would be the same as the monthly value. This follows if the returns of the portfolio are calculated, as they should be, as logarithms of the ratio of the value of the portfolio at the end of the given period to the value at the end of the prior period, so that compounding of returns is accomplished by summing. But with the same increase in the frequency of observations the daily Sharpe ratio difference would (assuming 21 trading days per month) become $\Sigma_{T=1}^{21} \Delta_{\text{monthly}}$ where $\Delta_-$ is the value from monthly returns, and likewise with $\Delta_+$. This is derived by Lo (2002, 40). And of course, the new number of observations would be $21 T$ in place of $T$ where $T$ is the number of months.

For securities that can be modeled as if the returns were i.i.d. bivariate-normal, begin by considering whether, with the starting monthly frequency of observations, the second term on the right-hand side of equation (1a) is sizable in comparison to the first term. The correlation coefficient $\rho_{ab}$ being near 1 and the Sharpe ratio difference $\Delta_-$ being large are circumstances within which the second term could be sizable when compared with the first. If the second term isn’t sizable relative to the first then further reducing its magnitude by resorting to a daily frequency of observations and thereby turning $\Delta_-$ into $\Delta_-/(\sqrt{T})$ and $\Delta_+$ into $\Delta_+/(\sqrt{T})$ would not significantly change the magnitude of the right-hand side of equation (1a). And then the left-hand side should be $21 T \cdot SE_{21}$ where the 21 subscript refers to daily observations, with the right-hand side being dominated by $2(1 - \rho_{ab})$. Thus $SE_{21} = \sqrt{SE^2}/21$ which scales with the number of days in the month in the same way as $\Delta_+/(\sqrt{T})$. This identical scaling factor has the consequence that the monthly value of the second term of $Z\Delta_- = \Delta_-/(\alpha)$, which is also the second term of $Z\Delta_- = -\Delta_-/(\alpha)$, which is $-\Delta_-/SE$, is not changed with the switch to daily values. Thus, the
Estimating the nuisance parameters from sample returns

By its very definition, the power of the test invokes a null hypothesis that has somehow been implemented. In equation (3c) above, which computes a quantity that is used in equation (3b) that computes the power, it’s assumed that to implement the null hypothesis it suffices to set \( \Delta = 0 \)… to compute the standard error using that value and the population values for \( \Delta \) and \( \rho_{ab} \). And we also need to set \( \Delta = 0 \) to compute the p-value as in equations (2a) and (3a).

It’s reasonable to worry a bit about whether that can properly be done. For one, can \( \Delta \) be changed to 0 without also changing \( \rho_{ab} \)? After all, both parameters pertain to the same portfolio returns. The answer is yes. With \( S_a = \mu_a / \sigma_a \) and \( S_b = \mu_b / \sigma_b \), subtract the constant \( (\sigma_a \mu_b - \sigma_b \mu_a) / (2 \sigma_a) \) from the returns of portfolio \( a \) and subtract the constant \( (\sigma_a \mu_b - \sigma_b \mu_a) / (2 \sigma_a) \) from the returns of portfolio \( b \). Because constants are being subtracted, \( \rho_{ab} \) the correlation coefficient of the returns is unchanged. And a bit of algebra shows that the new value of \( \Delta \) is 0, but that \( \Delta \) is unchanged.

In all, that’s a way with which the Sharpe ratio difference parameter \( \Delta \) could be dealt with to implement the null hypothesis when computing the p-value and the power—with null-restricted data. But what, say, should be used for the value of \( \Delta \), the sum of the population values of the Sharpe ratios? And what for the correlation coefficient \( \rho_{ab} \) of the returns of the two portfolios? Note that these two parameters only affect the standard error.

I have been referring to asymptotic approximations that are valid only when the returns are i.i.d. bivariate-normal. Happily, the truth is that to compute the standard error of the sampling distribution of \( \Delta \), the \( \Delta \) and \( \rho_{ab} \) parameters can simply be approximated by their sample values \( \hat{\Delta} \) and \( \hat{\rho}_{ab} \). And furthermore, that is generally applicable—not just to idealized distributions of portfolio returns.

It is a matter of the estimators for \( \hat{\Delta} \) and \( \hat{\rho}_{ab} \) being ‘consistent,’ which

power is not changed.

Now what happens if the second term on the right-hand side of equation (1a) is dominant over the first term? That could happen if \( \rho_{ab} \) gets very close to 1.0. If it is—see again equation (1a)—then
\[
T \cdot SE_{21}^2 = \Delta^2 / 2 \text{ so that } SE = \Delta / \sqrt{2T}, \text{ which yields } \Delta / SE = \sqrt{2T}. \]
In practice \( T \) is often 100 or more, so that the equation-(3b) power is already nearly 1.0 before we even consider upping the frequency of observations. Further improvement is scarcely possible. Things are a bit different with hedge funds, with portfolios whose returns are non-i.i.d. and both highly autocorrelated and heteroskedastic. Of course, increasing the frequency of observation would not improve the power of the test for such portfolios either. But such portfolios present the complication of the scaling factor that is the square root of the frequency of observation not being valid. That is shown by Lo (2002, 40).
roughly means that their values converge to those of $\Delta_+$ and $\rho_{ab}$ as $T$ the number of samples is increased without limit. For example, again with i.i.d. bivariate-normal returns, and beginning with $\hat{\rho}_{ab}$, if $\rho_{ab}$ is the true population value then $\hat{\rho}_{ab} = \rho_{ab} + \gamma \left(1 - \rho_{ab}^2\right) / \sqrt{T}$ can be written, where the added term consists of the adjustable factor $\gamma$ multiplying Pearson and Filon's (1898) calculation of the standard error of the correlation coefficient. Obviously every possible value of $\hat{\rho}_{ab}$ can be written that way, with a suitable choice of the value of $\gamma$. But that equation can be solved for $\rho_{ab}$ via the quadratic formula:

$$\rho_{ab} = \sqrt{T} \left[1 - \sqrt{1 - 4 \gamma \left(\hat{\rho}_{ab} - \gamma / T\right)}\right].$$

And for large $T$ that becomes $\rho_{ab} \approx \hat{\rho}_{ab} - 2\gamma / \sqrt{T}$ as $T \to \infty$. Of course $\gamma$ is to be expected to be of the order of magnitude of 1. If for example $\gamma$ is +2, then $\hat{\rho}_{ab}$ is two standard deviations above $\rho_{ab}$. It would be unlikely for $\hat{\rho}_{ab}$ to be higher still, so that to be very sure not to overestimate $\rho_{ab}$ while using $\hat{\rho}_{ab}$ in its place, $\hat{\rho}_{ab} - 2\gamma / \sqrt{T}$ could be substituted instead.

But the real point to make is that the standard error of the sampling distribution of the Sharpe ratio difference, as shown in equation (1a), is proportional to $1 / \sqrt{T}$ because it is the first term of an asymptotic expansion in the variable $1 / \sqrt{T}$. That's with it as a function of the population values of $\Delta_+$ and $\rho_{ab}$. If $\hat{\rho}_{ab} - 2\gamma / \sqrt{T}$ is substituted for $\rho_{ab}$ and the asymptotic expansion is formed again, the leading term will become the standard error as defined by equation (1a) but as a function of $\Delta_+$ and $\hat{\rho}_{ab}$. And the same can be done with $\Delta_+$ so that $\Delta_+$ can be replaced by $\hat{\Delta}_+$.\footnote{11,12}

This is actually standard operating procedure in the standard error business;

\footnote{11. I used the procedure of Memmel (2003), and the quadratic formula, and have assumed $\Delta_- = 0$ as for the null hypothesis case, to derive $\Delta_+ = \Delta_+ + \gamma \sqrt{2 \left(1 + \rho_{ab}\right) + \Delta_+ \left(1 + \rho_{ab}^2\right) / 4 / \sqrt{T}}$. Again, this permits us to substitute $\hat{\Delta}_+$ for $\Delta_- in the expression for the standard error of the sampling distribution of the Sharpe ratio difference because the last term, in $1 / \sqrt{T}$, has a numerator that is of order of magnitude 1 and the term can't contribute to the first-order term of the asymptotic expansion of the standard error. 12. Even with i.i.d. bivariate-normal returns, the mean of the sample values of $\hat{\rho}_{ab}$ is a biased estimator of $\rho_{ab}$ Lehman and Casella (1998, 96) report that the expected value is approximately $\rho_{ab} \left[1 - \left(1 - \frac{\rho_{ab}^2}{2T}\right)\right]$. But note that the correction term can be quite small if, say, $T \geq 100$. More to the point, no such correction term that varies in inverse proportion to $T$ could contribute to the first-order term in $1 / \sqrt{T}$ for the standard error. Similarly, Pav (2021, 5) finds that the expected value of the Sharpe ratio is $\mathbb{E}\left[1 + \frac{3}{4T}\right]$, which is similarly inconsequential.}
see for example equation (10) of Lo (2002, 38). But some scholars have concerned themselves with whether such a simple resort is the best that can be done with nuisance parameters; see for example Peter Reinhard Hansen (2005). So, I have dwelled on the matter.

To summarize, if the histories of the returns of the two portfolios of the Sharpe ratio difference test statistic are long enough, then those data alone can be used to estimate its standard error. No knowledge of the population values of the nuisance parameters is needed because the departures of the sample values from the population values are asymptotically insignificant. Suitable estimators of the standard error might be theoretically derived, based on assumptions about the statistical properties of the returns, as Memmel did. Or the analyst might resort to using resampling methods, bootstrapping, as in Ledoit and Wolf (2008). Therefore, with any backtesting procedure it can be assumed that the standard error \( \hat{SE} \) is known. And of course, \( \hat{\Delta}_- \) is known, leaving unknown only the population value \( \Delta_- \) of the Sharpe ratio difference.

**Confidence intervals and the standard error**

Given a sample value \( \hat{\Delta}_- \) of the test statistic, the population value can be written as \( \Delta_- = \hat{\Delta}_- + \gamma \cdot \hat{SE} \), with \( \hat{SE} \) being the standard error as inferred from the sample—by choosing \( \gamma \) to make that true. If there are grounds for supposing that \( \gamma \sim N(0, 1) \) then confidence intervals for \( \Delta_- \) can be defined, with a 95-percent confidence interval’s endpoints being \( \hat{\Delta}_- \pm 1.96 \cdot \hat{SE} \) where \( 1.96 \equiv \Phi^{-1}(0.975) \).

A better approach might be to derive the interval using quantiles of the bootstrap distribution of \( t^* = \frac{(\hat{\Delta}_- - \hat{\Delta}_-)}{\hat{SE}} \), where the asterisk refers to bootstrap samples derived from the original sample. For a two-sided test Ledoit and Wolf (2008) formed the bootstrap distribution of \( |t^*| \) via a complex procedure that involved doing bootstrapping twice, and in effect they used it to form a symmetric confidence interval. The confidence intervals of Table 1 of the “Hedge funds are different” subsection of the next section of this article are thus derived. Whereas the thus-found limits of the confidence intervals for the mutual fund pair and the hedge fund pair are respectively \((-0.020, 0.213)\) and \((-0.390, 1.283)\), the \( \hat{\Delta}_- \pm 1.96 \cdot \hat{SE} \) endpoints are \((-0.016, 0.210)\) and \((-0.388, 1.282)\)—not so very different.

That \( \hat{SE} \) can be substituted for \( SE \) must be considered to be a generally valid approach, and the best available method for deriving \( \hat{SE} \) should be used. But if
The power of the test in practice

It might be supposed that the power of the test be straightforwardly confronted, a priori. An analyst would like answers to questions such as: If the true Sharpe ratio difference $\Delta_-$ has approximately this value, then, with the available history of data, do I have satisfactory power? The problem that immediately arises is of course that of the analyst having hardly any proper basis on which to proceed to estimate the value of $\Delta_-$. 

But there are other ways to proceed. The analyst may want to indulge in a form of ‘a posteriori’ or ‘post hoc’ analysis regarding the power of the test, that is pursued subsequent to and on top of having assessed the statistical significance of a measured effect via the use of the p-value. The goal of the analyst might then be to further interpret any finding of insignificance by giving some sort of consideration to the power of the test. Is the negative finding trustworthy, or could the true effect be positive after all?

Those negative significance findings when the true effects are positive are of course type II errors. Whereas type I errors occur with the low frequency of the chosen significance level, analysts can easily stumble into circumstances in which the power of the test is so low that the frequency of type II errors exceeds 50
percent. And so, the general idea would be to give some sort of due consideration to the adequacy of the power, in order to control the type II errors—in a way that is analogous to the p-value having been used to control type I errors by setting the significance level $\alpha$ to a small value. Said control might just amount to labeling some findings of insignificance as suspect due to inadequate power.

Thus if I have been disappointed by seeing a portfolio formation scheme of mine produce a Sharpe ratio difference that was ruled to be statistically insignificant, and some form of post hoc power analysis suggests that the problem might have been that I simply didn’t have enough power, then that would mean that I should persevere and try to find uses for my scheme in circumstances in which more adequate data are available. That at least would be the hope.

Some researchers, statisticians being among them, have condemned post hoc power analysis as deeply flawed, even delusional. John Hoenig and Dennis Heisey (2001) has been cited about two thousand times and has been particularly influential in that regard. Others, such as Len Thomas (1997) and Magdalena M. Mair et al. (2020), have been less sanguinary. Hoenig and Heisey did offer an alternative approach. All of that will be sorted out next.

In search of a pro forma approach

Strikingly simple methods have been adopted by researchers, and even by some government agencies, that nonetheless look as though they should surely be helpful. For example, there is the idea of calculating a minimum detectable effect (MDE). That is the effect size—the Sharpe ratio difference—that would, with the sample value $\hat{SE}$ of the standard error, produce a test power of, say, 0.90 (which would be an adequate power). So, the natural thought is that a low MDE is wanted: The lower it turns out to be, the more evidence we have in support of the null hypothesis when there is a finding of statistical insignificance, because the power of the test will be quite high enough even when the population value of the test statistic is close to fulfilling the null hypothesis. That’s the thought, such as it is. And that is spelled out in Hoenig and Heisey (2001) and Mair et al. (2020). Or, more simply, the lower the MDE the better because a low MDE value means that the statistical significance of even a small effect can be accurately assessed with adequate power. Is that going to work out? That’s the question.

I’ll use the notation for the Sharpe ratio difference here, even though the findings will be generally applicable. And I’ll be dealing with the right-sided test, so the power function, cf. equation (3b), is

\[ P(\text{effect size}) = \Phi \left( \frac{\text{effect size} - \text{null hypothesis}}{\sqrt{SE^2}} \right) \]

\[ \pi_r(\alpha, \Delta) = 1 - \Phi\left(\Phi^{-1}(1 - \alpha/2) - \Delta / \hat{SE}\right). \]

Inverting that equation,

\[ \Delta / \hat{SE} = \Phi^{-1}(1 - \alpha/2) - \Phi^{-1}(1 - \pi_r). \]

Writing \( \Delta_{-0.90} \) for the MDE at 90-percent power, we see that it is

\[ \Delta_{-0.90} = \hat{SE} \cdot [\Phi^{-1}(1 - \alpha/2) - \Phi^{-1}(1 - 0.90)]. \]

It is simply proportional to \( \hat{SE} \).

The use that is made of the MDE is to set some threshold value for it, as a matter of policy, and to, in effect, affix trust labels to findings of insignificance if the calculated MDE does not exceed the threshold. Tests with MDE values greater than the threshold are labeled mistrust. If the label is trust, then the analyst trusts that the finding of insignificance is not a type II error. What could go wrong with that?

Well, because \( \Delta_{-0.90} \) is simply proportional to \( \hat{SE} \), and with the constant of proportionality being just that, a fixed number, filtering MDE values with a threshold is the same as filtering \( \hat{SE} \) values with a threshold. There is utterly no control of \( \Delta_- \), the population value of the test statistic, which means that there seems to be no actual control of the power of the test.

There is a related approach that consists of computing the power of the test using the minimum value of the test statistic that would be of practical interest, in lieu of the unknown population value of the test statistic. If the thus-computed power is at or above, again, say 0.90, then that warrants the trust label. But that would mean that the MDE would be at or below the minimum value of practical interest, so that we are again talking about imposing an MDE threshold. Thus, this related approach is really the same approach.

There is also a quantity that is called the minimum detectable difference (MDD), which is the critical value of the test statistic, \( \hat{SE} \cdot \Phi^{-1}(1 - \alpha/2) \), which is used in exactly the same way as MDE. But again, it is just proportional to the standard error and so, just as with the MDE and the related minimum value of practical interest, it does not provide direct control of the power of the test due to the lack of involvement of the population value of the test statistic (\( \Delta_- \) in this article).

The Hoenig and Heisey (2001) article refers to these efforts to determine the reliability of findings of insignificance as the power approach—though only one determinant of the power of the test is dealt with, the standard error. The authors then mention an alternative approach based on confidence intervals. And
O’CONNOR

in that discussion they remark that if neither limit of the confidence interval is far from the null value of the test statistic, which is zero in this article, then one has confidence that the null hypothesis should not be refuted. The reasoning behind the confidence interval approach outwardly seemed to be irrefutably correct, and it appears to have inspired the authors of Mair et al. (2020) to make use of the upper limit of the confidence interval in the very same way that MDD and MDE have been used: to affix the trust label to findings of insignificance when the upper limit of the CI is below some predetermined threshold.

Exploring that, let $CL_r = \hat{\Delta}_- + SE \cdot \Phi^{-1}(1 - \alpha/2)$ be the upper limit of the confidence interval. The finding of insignificance will be labeled trust if $CL_r < \Delta_{\text{max}}$ where $\Delta_{\text{max}}$ is the predetermined threshold at or above which the label is to be mistrust. By inspection of the expression for the right-sided power $\pi_r$, we then have that $CL_r < \Delta_{\text{max}}$ means that $\pi_r(\alpha, CL_r) < \pi_r(\alpha, \Delta_{\text{max}})$, because $\pi_r(\alpha, \Delta_\cdot)$ increases in a strictly monotone way with $\Delta_\cdot$. Thus, if we progressively decrease $\Delta_{\text{max}}$ so as to confine the upper limit of the confidence interval to the immediate neighborhood of zero, which should provide confidence-interval-approach assurances to the effect that the null hypothesis should indeed not be refuted, the power to determine the reliability of that conclusion is diminished. So is the confidence interval approach at odds with the the so-called power approach? This is indeed perplexing. Hoenig and Heisey brought this to light, but their examples and discussion differ from what I’m presenting here. Read on. This will be resolved.

Trying again, I then again explored the use of the upper limit of the CI, again with regard to the right-sided test, but with the studentized version of the upper limit: The studentized version of the upper limit is $\hat{\Delta}_- \Phi^{-1}(1 - \alpha/2)$, which is $\Phi^{-1}(1 - p_r) + \Phi^{-1}(1 - \alpha/2)$. In that form we see that trusting a finding of insignificance because the studentized upper limit of the CI is below some predetermined threshold is the same as trusting such findings when $\Phi^{-1}(1 - p_r)$ is below some predetermined threshold, which is the same as trusting the findings when $p_r$ is above some predetermined threshold.

Thus, the hope of being able to use the studentized upper limit of the CI to better determine the reliability of findings of insignificance went up in smoke. It boiled down to classifying the reliability based solely on the p-value, with higher p-values providing greater evidence of reliability. That is of course entirely consistent with prior expectations concerning the p-value, so that nothing new was discovered. And the computation and use of the p-value does not involve the population value of the test statistic ($\Delta_\cdot$ in this article). And so again there is no direct relevance to the power of the test because the power is so very dependent upon the unknown population value of the test statistic.
In short, the appearances seem to be that there is no post hoc way to bring about an improved understanding of the reliability of findings of insignificance. It’s all because the population value of the test statistic remains unknown and unknowable. According to Hoenig and Heisey (2001), one must leave what they call the power approach at that. But, I took a second look at the mutually-related MDE and minimum value of practical interest schemes, both of which amount to classifying findings of insignificance as trusted or mistrusted, respectively according to whether \( \hat{SE} < SE_{\text{max}} \). What should \( SE_{\text{max}} \) be? If \( \Delta_{-,PRAC} \) is the minimum value of the Sharpe ratio difference that is of practical interest, suppose that \( SE_{\text{max}} \) is chosen to be the value of \( \hat{SE} \) that would, with the value of the test statistic being \( \Delta_{-,PRAC} \), produce a right-sided power of 0.90. Thus

\[
SE_{\text{max}} = \Delta_{-,PRAC} / [\Phi^{-1}(1 - \alpha/2) - \Phi^{-1}(1 - 0.90)]
\]

\( \hat{SE} \) can be derived from the sample by inverting the definition of the right-sided p-value \( p_r \), getting

\[
\hat{SE} = |\hat{\Delta}_-| / [\Phi^{-1}(1 - p_r)].
\]

Or, rather than making use of a p-value estimator, an estimator for the standard error may be directly used. Therefore, a test based on \( \hat{SE} < SE_{\text{max}} \) could be implemented whenever the original hypothesis test has been concluded with a finding of statistical insignificance \( (p_r \geq \alpha/2) \)—with the goal being to characterize the finding of statistical insignificance with respect to its reliability, given that the power of the test may be suspect.

The starting point of this idea was the vague notion that an \( \hat{SE} \) value being known to be low, via \( \hat{SE} < SE_{\text{max}} \), seems to imply that the only way that the power could be low would be if \( \Delta_+ \), the population value of the Sharpe ratio difference, were to also be low—low but not necessarily zero or negative. If \( \Delta_+ \) were not low then the low \( \hat{SE} \) would mean that power would be high. But a high power with \( \Delta_+ \) not being low should mean that the finding of insignificance would not have happened in the first place. Thus, perhaps low \( \hat{SE} \) can be taken to truly mean low \( \Delta_+ \), so that one might try to turn \( \hat{SE} < SE_{\text{max}} \) into a provision that more or less guarantees that if the finding of statistical insignificance doesn’t mean that the population value of the test statistic is zero or negative then it at least means that the value is almost certainly less than \( \Delta_{-,PRAC} \).

But of course, that is just a sketchy idea. To further investigate, I also implemented a secondary left-sided hypothesis test whose null hypothesis is \( H_0: \Delta_+ \geq \Delta_{-,PRAC} \). Hoenig and Heisey (2001, 4) propose the use of a two-tailed version of this non-traditional casting of a null hypothesis, and present it as amounting to abandonment of the flawed power approach in favor of the confidence interval approach. A finding of statistical significance with respect to the secondary hypothesis test is then reached with rejection of the secondary null
hypothesis if \( \hat{\Delta}_- < \Delta_-\text{CRIT} \) where \( \Delta_-\text{CRIT} \) is de rigueur, defined by choosing it so as to make the significance level of the left-sided secondary hypothesis test be \( \alpha \), that hypothesis test’s type I error rate, as in \( \Phi(\frac{\Delta_-\text{CRIT} - \Delta_-\text{PRAC}}{\hat{\Delta}_-}) \). Thus \( \Delta_-\text{CRIT} = \Delta_-\text{PRAC} + \hat{\Delta}_- \cdot \Phi^{-1}(\alpha/2) \).

Does this look familiar? If we take the test that was introduced above, which was adapted from the formulation of Mair et al. (2020), which was \( CI_r < \Delta_-\text{max} \) with \( CI_r = \hat{\Delta}_+ + \hat{\Delta}_- \cdot \Phi^{-1}(1 - \alpha/2) \) being the upper limit of the confidence interval, the fact that \( \Phi^{-1}(1 - \alpha/2) = -\Phi^{-1}(\alpha/2) \) means that if we set \( \Delta_-\text{max} = \Delta_-\text{PRAC} \) then indeed the two tests are one and the same. But now it’s understood to be a bona fide hypothesis test of significance level \( \alpha\ ) with a well-defined attendant null hypothesis.

The simultaneous use of a secondary hypothesis test makes all the difference. For clarity in the moment, I shall affix accept and don’t accept labels, respectively according to whether \( \hat{\Delta}_- < \Delta_-\text{CRIT} \), in lieu of trust and mistrust. The difference is that by accept I mean accept the finding of statistical insignificance with the use of the original null hypothesis, but with the understanding that the attendant rejection of the secondary hypothesis test’s null hypothesis means that in reality \( \Delta_- \) might in fact be as large as \( \Delta_-\text{PRAC} \) (and not restricted to being as large as zero as in the primary null hypothesis). But the don’t accept label means much the same as the mistrust label—it is only to be taken to be an indication that it’s best to assume that the available data are not sufficient for reliable hypothesis testing. Any Sharpe ratio difference that earns the don’t accept label is to be regarded as not reliably replicable with regard to sign or magnitude; the accept label means that the population value of the difference very likely does not exceed \( \Delta_-\text{PRAC} \).

Consider what that means. The goal was to somehow dodge the adverse effect of the original hypothesis test having type II errors. With those type II errors there is a finding of statistical insignificance, even though the null hypothesis \( H_0: \Delta_- \leq 0 \) is false, meaning that \( \Delta_- > 0 \). With type I errors of the secondary hypothesis test there is a finding of statistical significance, appending the accept label to the original hypothesis test’s finding of insignificance, even though the secondary null hypothesis \( H_0: \Delta_- \geq \Delta_-\text{PRAC} \) is true, meaning that \( \Delta_- \geq \Delta_-\text{PRAC} \). So the type II error of the original hypothesis test occurs with \( \Delta_- > 0 \) and the type I error of the secondary hypothesis test occurs with \( \Delta_- \geq \Delta_-\text{PRAC} \). But these are similar inequalities. There is just a shift in the right-hand side and the analyst controls the magnitude of the shift.

And now here’s the big difference. With the original hypothesis test we have no way of knowing what the probability of occurrence of type II errors is, due to the power of the test not being known. But, notwithstanding the fact that the
secondary hypothesis test’s type I error is a doppelgänger of the original test’s type II error, we do know the probability of its occurrence: It’s \( \alpha/2 \). And as such, it’s controlled.

And, about the type II errors of the secondary hypothesis test, they consist of affixing the don’t accept label to findings of statistical insignificance under the original null hypothesis of \( H_0: \Delta_\leq 0 \), after having failed to reject the null hypothesis \( H_0: \Delta_\geq \Delta_{PRAC} \) of the secondary hypothesis test when in fact \( \Delta_\leq \Delta_{PRAC} \). Note especially that the secondary hypothesis test cannot be used backhandedly, in stand-alone fashion, to establish statistically significant outperformance. A finding of statistical insignificance under it only means that the possibility of \( \Delta_\geq \Delta_{PRAC} \) can’t be dismissed. And the power of the test being unknown, uncontrolled, means that any such finding would be confounded by the potentially high type II error rate. The left-sided secondary hypothesis test is only suitable for its ability to lend confirmation, of a sort, to some of the findings of insignificance under the original right-sided hypothesis test.

Now it might come to mind that some analyst might elect to just proceed with the original hypothesis test alone, and to take a finding of statistical insignificance under it to just mean that further study with more adequate data was needed—declining to make a spot judgment to the effect that the investment alternative of interest was not likely to become an outperformer. Yes, but the improvement that is brought about with the use of the secondary hypothesis test is the immediate resolution of the statistical insignificance findings in some cases, permitting the investment alternative to immediately be categorized as unlikely to be destined to offer outperformance of practical significance. That is not a huge victory, but it is beneficial.

I did apply both the \( \hat{SE} < SE_{max} \) test and the secondary hypothesis test to real data—to 47 findings of insignificance that I found in three published articles. Details about the findings of the three studies are to be found in the “More literature” section below. Of those 47 findings of insignificance, the standard error test rated nine as accept and the secondary hypothesis test rated 13 as accept. In no case was the standard error accept result overruled by the secondary hypothesis test; the latter awarded the accept label more liberally.

Of the three articles, one was based on a backtesting interval that was longer by far than that of the other two. And the frequency with which the accept label was awarded using the secondary hypothesis test was highest by far with that article—suggesting a positive correlation between the secondary hypothesis test’s frequency of acceptance of the original hypothesis test’s findings of insignificance and the duration of the backtesting interval. Such a correlation can come about in practice through the role played by the standard error \( \hat{SE} \) in the formula \( \Delta_{-,CRIT} = \)
\[ \Delta_{PRAC} + \hat{SE} \cdot \Phi^{-1}(\alpha/2) \] for the critical value of the secondary hypothesis test. Bearing in mind that \( \Phi^{-1}(\alpha/2) \) is negative, \( \Delta_{CRIT} \) is increased if \( \hat{SE} \) is reduced, meaning that accept labeling should be more frequent. \( \hat{SE} \) is reduced if the duration of the backtesting interval is increased. Hence the positive correlation.

Because the type I error rate of the secondary hypothesis test is known and controlled, I am inclined to use it rather than the standard error test. But on occasion power-approach-like consideration of the size of the standard error nonetheless suffices. For example, suppose that the analyst deliberately conjures up, out of whole cloth, a value for the population value of the test statistic that is implausibly high, and with it and the sample-derived standard error computes, using equation (3b) or its right-sided counterpart, the power that the test would have with that implausibly high value in place of the population value of the test statistic. If the thus-computed power were to be unacceptably low, then it would be reasonable to conclude that the true power of the test was unacceptably low. This would be much the same as finding the value of the MDE to be implausibly high for consideration as a population value of the test statistic.

The section “The last word” near the end of this article definitively demonstrates the folly of conducting hypothesis tests when the power of the test is really too low. It’s not easy to come up with an example of a portfolio pair that is such that the analyst can be confident in advance that the true Sharpe ratio difference test statistic, be it \( \Delta_{-} \) or \( |\Delta_{-}| \), is quite substantially greater than zero, with investors seriously interested in seeing the difference tested for statistical significance. For the most part, we are left with understanding only that the duration of the backtesting interval being too short or the correlation coefficient between the returns of the paired portfolios not being high enough (meaning that the standard error is not small enough), and the population value of the test statistic not being high enough, singly or together in concert, are factors that can readily cause the test to have inadequate power. But the secondary hypothesis test that I have described, which is an adaptation of Hoenig and Heisey’s (2001) confidence interval scheme as nearly emulated by Mair et al. (2020), can be helpful at resolving some outcomes as definitely being unpromising, even when it’s not known in advance that the power of the test is adequate for that purpose.

**Hedge funds are different**

Lo (2002, 44) provides data on actual funds that show that the returns of hedge funds are autocorrelated to a much greater degree than those of mutual funds. Benjamin R. Auer and Frank Schuhmacher (2013, 201) confirm strong autocorrelations in hedge fund returns. It’s hardly obvious as to why that might be
the case, but Lo and colleagues did present a feasible explanation: Mila Getmansky, Andrew W. Lo, and Igor Makarov (2004) investigated several mechanisms and found that only the effect of the illiquidity of assets held by hedge funds was enough to account for the observed degree of serial correlation. The fact that market prices are not frequently reestablished for illiquid investments means that there is effectively a sort of smoothing process affecting the returns, which acts like a trailing moving average. Since hedge funds hold securities both long and short, market risk is greatly diminished and that causes the funds to sport high Sharpe ratios. But the smoothing effect of the illiquidity further reduces the volatility and further boosts Sharpe ratios—by 73 percent in one studied example that involved smoothing over just two reporting periods.

Relatedly, Ledoit and Wolf (2008, 857) shows results pertaining to simulated portfolio returns. One pair of portfolios was given simulated i.i.d. returns drawn from a bivariate-normal distribution, as in the Jobson and Korkie (1981) work; other pairs were simulated so as to exhibit autocorrelations and other non-i.i.d., and non-bivariate-normal characteristics, such as might be exhibited by hedge funds. The Memmel (2003) corrected Jobson and Korkie model was configured to compute p-values, of iterates. And from the p-values the probability of rejecting the null hypothesis was computed, for the two-sided test whose statistic is the absolute value of the Sharpe ratio difference. The simulations emulated the null hypothesis being true, by forcing the two Sharpe ratios to be equal.

It is explained above that this rejection probability must be the significance level $\alpha$. Ledoit and Wolf’s findings, in their Table 1, are that the Memmel-corrected model brought about overestimates of the rejection probability when applied to simulated returns that had characteristics in common with those of hedge fund returns: For the five given flavors of departure from i.i.d. bivariate-normal returns the computed powers of the test when the assumed value of $\alpha$ was 0.05 were not 0.05 but were 10.7, 7.2, 7.4, 9.5, and 14.5. Understand that to get these particular estimates Ledoit and Wolf did not use their method for conducting hypothesis testing; they are the result of applying the Memmel-corrected model to simulated hedge-fund-like returns.

Auer and Schuhmacher (2013, 203) provide some confirmation of that result using real hedge fund data. Auer and Schuhmacher show that when the same Memmel-corrected model is applied to calculating the p-values of hedge funds, findings of statistical significance are reached much more often than with the method and code of Ledoit and Wolf which accounts for and compensates for non-i.i.d. and non-bivariate-normal characteristics. Assuming that the Ledoit and Wolf method and codebase doesn’t overestimate p-values of hedge funds, this can be taken to mean that the Memmel-corrected model underestimates p-values of hedge funds. These findings and Ledoit and Wolf’s rejection probability tests both
indicate that the Memmel-corrected model produces a standard error (a standard deviation of the Sharpe ratio difference) that is much too small for use with hedge funds. And that leads to p-values that are too low and power estimates that are too high.

And although it cannot always be counted on, by their very nature hedge funds tend to have returns that are weakly correlated with the returns of almost any other investment, including other hedge funds. As has been shown above, this has negative implications regarding the power of the test whose statistic is a Sharpe ratio difference.

In their own article Ledoit and Wolf (2008) show p-values for two pairs of funds. If there is a reason for supposing in advance that the true $\Delta_\cdot$ of either of the two pairs of is well removed from zero, I don’t know what it could be. So, it can’t just be assumed that power of the hypothesis test was sufficient. I therefore revised Tables 2 and 3 of Ledoit and Wolf (2008, 857–858), adding confidence intervals. To do so I made minor changes to Ledoit and Wolf’s R code for computing the p-value. My findings are shown in Table 1 below. The p-values and confidence intervals do not depend upon the significance level $\alpha$. That’s a hypothesis testing thing. If you don’t test the hypothesis with the computed p-value, then you don’t need and therefore don’t have a significance level.

<table>
<thead>
<tr>
<th>Fund pair</th>
<th>Sharpe ratio difference</th>
<th>P-value</th>
<th>Sharpe ratio difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fidelity Fidelity Aggressive Growth</td>
<td>0.092</td>
<td>0.097</td>
<td>[95.0% CI: −0.020, 0.213]</td>
</tr>
<tr>
<td>Coast Enhance Income JMG Capital Partners</td>
<td>0.447</td>
<td>0.294</td>
<td>[95.0% CI: −0.390, 1.283]</td>
</tr>
</tbody>
</table>

I did investigate the power of the test for the Coast Enhance Income–JMG Capital Partners hedge fund example—using the given sample values of the Sharpe ratios, the p-value as computed by Ledoit and Wolf, and the means described at the conclusion of the “Confidence intervals and the standard error” subsection of the “Mathematics of power—i.i.d. bivariate-normal returns” section above. I assumed various values of the population value of the Sharpe ratio difference, 0.40, 0.60, 0.80, and 1.0. These are conjectured monthly values, not annualized, and they are deliberately chosen to vary from being large to being very-very large (even for hedge funds). Surely the likes of one of those should produce adequate power, one might suppose. But no. The corresponding powers of the test are 0.16, 0.29, 0.47, and 0.65. That supports the claim that I made in the introduction of this article to the effect that it was certainly improper to have presented this particular example without disclosure of the likelihood of the power being too low. But also,
the secondary hypothesis test produced a don’t accept label for this pair, which is consistent with the power being too low.

More literature

I was curious to see what others had done with the Ledoit and Wolf (2008) procedure. Whereas the article has been cited 981 times since 2008, it was cited 136 times in 2022. It is not going out of style. I discovered that there is an entire category of studies, some quite recent, that use the Ledoit and Wolf (2008) method and codebase to determine if portfolio and benchmark Sharpe ratios differ by statistically significant margins, when at the outset there is no compelling reason to suppose that the ratios should substantially differ.

Why investigate when there’s no compelling reason to suppose that there would be substantial differences? Perhaps the authors of such studies privately stand in disbelief of the claims of others about some innovation or supposedly superior way of investing. So, their intention could be to dispassionately demonstrate that the true believers are wrong by showing that the observed Sharpe ratio differences lack statistical significance.

But if that is indeed their plan, then what should their thoughts be about the power of the test? They should be concerned that the power of the test would often be too low because there would be a systematic tendency for the magnitudes of the population values of the Sharpe ratio differences, the $\Delta_-$ values, to not be well removed from zero (cf. Figure 1). I have explained, and make especially clear in “The last word” section below, that it is foolhardy to attempt significance testing if there is every reason to suspect the power of the test to be quite low.

The aforementioned article by Auer and Schuhmacher (2013) is a study of hedge fund performance that involves 4,322 funds divided into 19 categories with a different benchmark for each category. For each of the funds the authors calculated the statistical significance of the difference between the fund’s Sharpe ratio and the benchmark’s. In their abstract the authors state that “Only a small fraction of hedge funds in our large dataset can significantly outperform passive investments in corresponding hedge fund indices.” And by “significantly” they did mean to refer to statistical significance. With that broad brush they condemn the entire hedge fund industry.

Remarkably, this study involves benchmarks that were each presumably designed to do a good job of representing hedge fund performance within each category. That should have caused Auer and Schuhmacher to have no reason to

---

suppose that the magnitudes of many of the $\Delta_-$ values were well removed from zero. Furthermore, the authors only eliminated funds having 24 or fewer months of historical returns data from their study, so that in some cases the number of observations was dismally low. Finally, unlike, say, long-only mutual funds, it can’t be assumed that hedge funds will generally be well correlated with their benchmarks. These are all foreboding indications that suggest that the power of the test was too low in many of the cases considered.

Since the low power of the test means that false null hypotheses will rather often not be rejected, it is understandable why few of the hedge funds were found to have outperformed their benchmarks in a statistically significant way—type II errors. The authors would have been better off speculating that the dearth of findings of significance was due to the designers of the benchmarks having done their jobs well, leading to generally low magnitudes of the $\Delta_-$ values and low power, rather than to lackluster hedge fund performance.\footnote{Curiously, the authors state on page 198 that “… the statistical power of the test is low, especially for small sample sizes. Thus, a significant test result can be seen as strong evidence of a difference in risk-adjusted performance.” But they say this only in reference to the Jobson and Korkie findings for i.i.d. bivariate-normal returns. And it is odd that they only mention an upside to the problem of low power. They may have concluded that Ledoit and Wolf’s method somehow improves the power of the test. And Ledoit and Wolf did claim that their method worked especially well with small sample sizes. That would explain why they are not concerned about low power and small sample sizes affecting their own Ledoit and Wolf test findings.}

I began to write something to the effect that in recent years dozens upon dozens of analysts have undertaken the task of estimating the efficacy of investing in stocks of corporations that are governed in such a way as to qualify as ‘ESG’ according to certain guidelines. I was a bit off! Google Scholar finds 484,000 hits on ‘ESG’ and 48 of them are among the articles citing Ledoit and Wolf (2008).\footnote{Accessed on May 6, 2023.} Generally, in one way or another these investigations involve substituting an ESG-qualified security for one of an enterprise that is otherwise similar in ESG-unrelated ways. The studies are based on the returns on the stocks of the companies, not upon environmental or social benefits brought about by ESG operation.

Of course, the authors of these studies don’t have good a priori reasons to suppose that ESG investing would be either beneficial or harmful to investors. Expecting a professor of finance to venture into the jungle to investigate whether growing coffee in the shade is going to be lucrative or not is asking too much. Then too, if ESG operations are good for investors and the stock market is efficient, an announcement of conversion to ESG operation by a company should immediately be greeted by a sudden large increase in the price of the company’s stock—with the gains to investors being realized at once and not necessarily occurring within the
study periods of these various ESG articles.

So, again, were the authors of such ESG studies to have considered the power of the test they would have had to assume that the population values of the $\Delta$-Sharpe ratio differences would be systematically of low magnitude—meaning that the power of the test would be low.

Consider Ricardo de Souza Tavares and João Frois Caldeira (2023, 61), who ask “Is replacing standard investments with ESG substitutes a good choice?” There was only about a decade of monthly data. Table 6 of that article shows that the procedure of Ledoit and Wolf (2008) assigned Sharpe ratio difference $p$-values under $\alpha = 0.05$ to only three of the 12 market indexes that were modified by ESG substitutions. And although the authors do state that this means that generally there is a lack of statistical significance, they don’t mention the power of the test. One is left wondering if the findings of insignificance can be trusted, due to the possibility of the power of the test possibly being systematically low.

Costanza Torricelli and Beatrice Bertelli (2022, 18) address “The trade-off between ESG screening and portfolio diversification in the short and in the long run.” In their Table 5, the significance of Sharpe ratio differences is assessed despite the data being monthly and there being only 15–18 months of data. The method of Ledoit and Wolf (2008) was used. If somehow in any of the studied cases the population value of the Sharpe ratio difference was sizeable, the power of the test would still be quite low due to the very low number of observations. With $\alpha = 0.05$ the significance level, the authors find significance with only two out of 24 endpoints. Some such outcomes would be expected given the low power of the test.

With all such considerations of ESG investment there are apt to be several alternatives that have been defined and tested—different guidelines for acceptability, substitutions made in different market indices, involvement with different alternative mean-variance optimization models, etc. Thus, the problems discussed above in “The mathematics of hypothesis testing” section of this article that are brought about by needing to choose among investment alternatives when the hypothesis tests are of low power come to the fore.

I had written all of the above, about the Opdyke (2007) article with his

---

17. A note at the bottom of their Table A1 states that the tests of that table, which do not include the Sharpe ratio difference test, were not applied to the short-term subperiods because the power would be inadequate given the small number of observations. Why then was significance testing conducted with the Sharpe ratio difference over the short-term subperiods on Table 5? Quite possibly it is because Ledoit and Wolf (2008) fails to mention the power of the test but contains statements that suggest that the Ledoit and Wolf method is especially good about working well with small sample sizes. Indeed, in the text Torricelli and Bertelli (2022) state that the method “is robust to non-normality, correlation and errors due to small samples,” which would seem to mean that it was assumed that the method could cure the problem of the low power due to the small sample size.
analysis of just three years of weekly data on certain mutual funds and about the research by Torricelli and Bertelli (2022) and by Tavares and Caldeira (2023), before I had settled on the secondary hypothesis test that is described in the “In search of a pro forma approach” subsection of “The power of the test in practice” section above in this article. Does the secondary hypothesis test help with the interpretation of the findings of these authors?

For the monthly Torricelli and Bertelli (2022) and Tavares and Caldeira (2023) data I set $\Delta_{PRAC} = 0.04$ which when annualized would amount to a Sharpe ratio improvement of approximately $0.04\sqrt{12} \approx 0.14$ which would be a small but meaningful improvement. And for the weekly Opdyke (2007) data I chose $\Delta_{PRAC} = 0.02$ which would be nearly equivalent on an annualized basis. Of all of the 15 Opdyke mutual fund matchups with his two best-performing funds (in the first two columns of his p. 20 Table IV) that produced p-values indicating statistical insignificance (with $\alpha = 0.05$), exactly none earned the accept label. That is entirely consistent with the fact that he had only three years of data, which would have led to low power of the original hypothesis test, hence the denial of accept labeling.

Similarly, the study of Torricelli and Bertelli (2022) that involved only 15–18 months of data would surely have been affected by low power, and of the 22 findings of statistical insignificance only six earned the accept label. But Tavares and Caldeira (2023) had a decade of monthly data, which is an amount that could in some circumstances be adequate. And seven of the nine findings of statistical insignificance (with $\alpha = 0.05$) earned the accept label. That seems to demonstrate a certain efficacy of the secondary hypothesis test at reducing the uncertainty in findings of insignificance when the power of the test is suspect. It would have made it possible for Tavares and Caldeira to have written, with a known measure of certainty, that with those seven outcomes there would be no improvement in the annualized Sharpe ratio difference of more than about 0.14. They were instead only able to write that “we cannot say that there is any difference between the Sharpe ratios,” which is decidedly less informative.

The recourse to confidence intervals

Larry Wasserman’s *All of Statistics* (2004) is a textbook by an academician, a statistician. And in it Wasserman wrote this: “There is a tendency to use hypothesis testing methods even when they are not appropriate. Often, estimation and confidence intervals are better tools.” But that was before the ball got rolling. Since then, others have chimed in, with statisticians that work in medical research leading the way. Valentin Amrhein et al. (2019) offered a letter that was signed by 800 researchers that appeared as a comment in Nature with the heading “Retire

To be sure, the objections are not all focused on the untoward aspects of hypothesis testing that appear when the adequacy of the power of the test is in question, because it is reasonable to simply complain about the black and white aspect of significance testing. That would be a reasonable complaint in and of itself. But John Ioannidis is a coauthor of Button et al. (2013), which is indeed particularly about the perils of testing when the power is too low. So yes, the complaints are in good measure about the damage done when the power is too low.

Means of computing confidence intervals have been presented above in the “Confidence intervals and the standard error” subsection of the “Mathematics of power—i.i.d. bivariate-normal returns” section. And it was shown that standard errors, and therefore confidence intervals, can be derived from published p-values. That means that whenever an article is encountered that lists p-values of Sharpe ratio differences and tells us whether the differences were found to be statistically significant or not, there is no need to worry about whether the power of the test was high enough. Approximate confidence intervals can be calculated and analysts and investors alike can be content with them.

I have been asked to show how ‘investment horizons’ might be worked into the topic of this article. There is nothing definite or exact about the bare notion of an investment horizon. That is, there are many entirely different ways of responding to the needs of investors who want to be assured that their money will be there when they need it at some point in the future, such as at retirement. But one way to implement an investment horizon is to modify the confidence intervals of the performance measures. The method that I show here is applicable to any performance measure that could serve as a hypothesis test statistic, not just to Sharpe ratio differences. But I’ll use the notation that was used for Sharpe ratio differences.

The formula is simple. Above, confidence interval endpoints for a 95-percent confidence interval were found to be \( \hat{\Delta}_- \pm 1.96 \hat{SE} \) with \( \hat{\Delta}_- \) being the sample value of the Sharpe ratio difference and \( \hat{SE} \) being the sample-derived standard error of that statistic. That would be for normally distributed values of \( \hat{\Delta}_- \).

\[ t^* = \frac{\hat{\Delta}_- - \hat{\Delta}_-}{\hat{SE}} \]

With, say, a bootstrap distribution of \( t^* \), the 95-percent confidence interval endpoints would be \( \hat{\Delta}_- - z_{97.5} \hat{SE}, \hat{\Delta}_- - z_{2.5} \hat{SE} \), as Tim C. Hesterberg (2015, 381) demonstrates. Here \( z_{2.5} \) and \( z_{97.5} \) are quantiles of the distribution of \( t^* \).
so that $z_{2.5}$ is negative. To implement an investment horizon of $T_{\text{hor}}$ periods into the future as a confidence-interval modification, simply multiply $\hat{SE}$ in the formula for the endpoints by $\sqrt{1 + \frac{T}{T_{\text{hor}}}}$ where $T$ is, as before, the number of periods in the sample.

How does this come about? It’s easier done than said. The random variable $\hat{\Delta} - \Delta$ is distributed with standard error $SE$, which pertains to $T$ observations of historical data, whose estimator yields the sample value $\hat{SE}$. The distribution’s 2.5 and 97.5 percentiles give us the 95-percent confidence interval endpoints. Similarly, the random variable $\Delta - \hat{\Delta}_{\text{hor}}$ is distributed with standard error $SE_{\text{hor}}$, which pertains to $T_{\text{hor}}$ observations of future data—for which an estimator is needed that somehow yields a $\hat{SE}_{\text{hor}}$ value that is derived from the only data that is available, the data of the historical sample.

To get horizon-adjusted confidence intervals, simply get the 2.5 and 97.5 percentiles of the distribution of $\hat{\Delta} - \hat{\Delta}_{\text{hor}}$, which is $\left(\hat{\Delta} - \Delta\right) + \left(\Delta - \hat{\Delta}_{\text{hor}}\right)$—after finding this distribution’s standard error. And what is it? Well, the squares of the standard errors of the distributions are all just variances. Since these are sampling distributions, pertaining to entirely different time intervals at that, the values of the Sharpe ratio differences would not be correlated. So, the square of the standard error of the distribution of $\hat{\Delta} - \hat{\Delta}_{\text{hor}}$ is the sum of the squares of the standard errors of the distributions of $\hat{\Delta} - \Delta$ and of $\Delta - \hat{\Delta}_{\text{hor}}$. That is, $SE_{\text{adj}}^2 = SE^2 + SE_{\text{hor}}^2$. And since the square of the estimator value of the standard error scales with the reciprocal of the number of observations, $\hat{SE}_{\text{hor}}^2$ may be taken to be $(T/T_{\text{hor}}) \hat{SE}^2$. Thus $\hat{SE}_{\text{adj}}^2 = \left(1 + \frac{T}{T_{\text{hor}}}ight) \hat{SE}^2$. Hence the investment-horizon confidence-interval adjusting factor of $\sqrt{1 + \frac{T}{T_{\text{hor}}}}$.

If $T_{\text{hor}}$ is very large then the adjustment factor is about 1, which is consonant with remarks made above to the effect that $\Delta$, the population value of the Sharpe ratio difference, can be regarded as the long-term future value. Or, if $T$ is much smaller than $T_{\text{hor}}$, which would not be a particularly unusual circumstance to encounter in practice, then the adjustment factor would be not much more than 1. That would be because the initial uncertainty that is brought about by $T$ being small would be dominant over that brought about by the larger $T_{\text{hor}}$, so that the initial uncertainty would not need much adjustment.

In practice, the horizon-conscious investor who needs to choose a safe-enough investment from a list of investments could rank the investments by, say, the lower bounds of the horizon-adjusted confidence intervals of the Sharpe ratio differences rather than by the Sharpe ratio differences. In general, the two rankings would be quite different. It’s interesting to note that whereas the Sharpe ratio is a
risk-adjusted measure of returns, the lower bound of the Sharpe ratio difference confidence interval is a risk-adjusted measure of the Sharpe ratio difference (because the term in $\hat{SE}$ is subtracted, amounting to a penalty for volatility).

Unfortunately, selection bias must still be dealt with when there are multiple investment alternatives under consideration and confidence intervals are in use. Cross-validation and walk-forward methods can mitigate selection bias. But also, a problem is encountered that is analogous to the problem of the false error rate accumulating with multiple hypotheses that is discussed in “The mathematics of hypothesis testing” section of this article. There is a similar kind of accumulation of errors, of the failures of the population values of the performance statistic to lie within the confidence intervals. But again, there are means of dealing with that. The confidence intervals can be adjusted (see Benjamini and Yekutieli 2005; Benjamini, Hechtlinger, and Stark 2019). These remedial measures for dealing with the multiple hypothesis problem have nothing in particular to do with Sharpe ratio differences but are quite generally applicable.

The last word

The power of any given properly stated hypothesis test must always be regarded as not being up for grabs. That is, the power of the test is intrinsic and immutable once the test statistic and the null hypothesis are both defined and the significance level $\alpha$ has been chosen. As Sir Ronald Aylmer Fisher (1935) himself put it: “It is evident that the null hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the ‘problem of distribution,’ of which the test of significance is the solution.” If the null hypothesis is indeed exact, then the power follows unambiguously for it is, by definition, a probability that can be straightforwardly thought of as a frequency of occurrence.

What remains that might be confusedly regarded as amounting to mutability of the power of the test is the struggle to invent improved estimators of the p-value, that are not grossly biased. Equation (1a) of this article is exact (in the limit of large $T$), provided that the returns of the two portfolios are i.i.d. bivariate-normal. That is, the estimator of equation (3a) is consistent if it incorporates (1a) via (2a). But if the returns are not distributed in that idealized way, then (3a) becomes a biased estimator. Auer and Schuhmacher (2013) confirm that it produces, with hedge funds whose returns are far from i.i.d. bivariate-normal, p-values that are very substantially lower than those computed using the method of Ledoit and Wolf (2008). Because of the way that it was derived it must be assumed that the Ledoit and Wolf estimator to be the least biased of the two estimators—able to cope with autocorrelations and heteroskedasticity.
But the p-value estimator of Ledoit and Wolf must in essence also be, and is, about as exact in application to i.i.d. bivariate-normal returns as equation (3a). That is confirmed by Figures 2 and 3 which appear further on below. Thus, for all its sophistication, if the power of the test with such idealized returns is low due to factors such as the duration of the backtesting interval not being long enough, the true Sharpe ratio difference not being high enough, or the correlation coefficient being too low, then the Ledoit and Wolf estimator can’t and doesn’t overcome that. The low power is innate.

One path to better understanding is to pay attention to the effect of the power of the test on the sampling distribution of the p-value. Figure 2 shows the sampling distribution of the two-sided p-value, computed via simulation of the histories of the returns of the two portfolios. Note that on the figure there are twenty histogram bars, so that the boundary between the first and second bars falls, by design, at a p-value of 0.05 which is the assumed significance level \( \alpha \). With each of the 1,000 trials the p-value was calculated using both the method and code of Ledoit and Wolf and the asymptotic method of equation (3a). Equation (1a) provided the estimate of the standard error for the asymptotic approach.\(^{18}\)

Note especially, in the legend, that the population value of the Sharpe ratio difference is non-zero, positive. Therefore, the fraction of all the p-values that exceed \( \alpha = 0.05 \) is the type II error rate \( \beta \). And from the legend, \( \beta \) is about 0.07 so that the power as computed from the sampling distribution of the p-value is about \( 1 - 0.07 = 0.93 \), which is very good. And that agrees nicely with what is independently computed with the asymptotic expression of (3b), which is 0.929.

The following is found in Wasserman’s text (2004, 158): “In other words, if \( H_0 \) is true, the p-value is like a random draw from a Uniform (0,1) distribution. If \( H_I \) is true, the distribution of the p-value will tend to concentrate closer to 0.” His \( H_0 \) can be taken to be the two-sided null hypothesis \( \Delta = 0 \), and then his \( H_I \) would be \( |\Delta| > 0 \).

On Figure 2 the distribution of the p-value is well-concentrated close to 0. But it is not a really a matter of one or the other—of the p-values being either uniformly distributed or concentrated closer to 0. Rather, the higher \( |\Delta| \), the greater the concentration near 0. Thus “if \( H_0 \) is true” can be taken to mean in the limit of \( |\Delta| \to 0 \), in which case there is no concentration near 0 but just an utterly

---

\(^{18}\) To simulate the return histories of two portfolios a sample a is prepared using random selection from \( N(0, 1) \) and then a second one b is prepared in the same way but is then altered by replacing it by \( \sqrt{1 - \rho_{ab}^2} \) times itself plus \( \rho_{ab} \) times the first, with \( \rho_{ab} \in [0, 1] \)—thus establishing a substantial amount of cross correlation between the first sample and the altered second sample. The samples are then parameterized: The returns of the first sample are multiplied by \( \sigma_a \) and then \( \mu_a \) is added; the altered second sample’s returns are multiplied by \( \sigma_b \) and then \( \mu_b \) is added.
flat histogram.

**Figure 2.** An example with 360 observations

<table>
<thead>
<tr>
<th></th>
<th>Ledoit &amp; Wolf</th>
<th>IID Asymptotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(p-value &gt; 0.05)</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>1000 simulation runs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

360 months, $S_{h_0} = 0.20$, $S_{h_1} = 0.10$

$\mu_0 = 1.60$, $\sigma_0 = 8.00$, $\mu_1 = 0.60$, $\sigma_1 = 6.00$, $\rho_{\mu_0} = 0.85$

**Notes:** Here $\Delta_0 = S_{h_1} - S_{h_0}$ is the a priori population value of the Sharpe ratio difference. And $P(p\text{-value} > 0.05)$ is the probability of a type II error, whose complement is the power of the test.

Wasserman states that it’s clear that the histogram must be flat when $H_0$ is true because that is consistent with the probability of committing a type I error being $\alpha$, the significance level (which is 0.05 in the figure), when the null hypothesis is rejected because the p-value is less than $\alpha$. That is, the area under the first bar of the histogram must be one-twentieth of the total area under the 20-bar histogram, because that would be 0.05, the value of $\alpha$. That is consistent with flatness. I have indeed rerun the code that produced the figure, forcing $\Delta_0 = S_{h_1} - S_{h_0} = 0$ which fulfills $H_0$, and the histogram is utterly flat.

Professor Wasserman might also have gone on to relate the fact that if $H_1$
is true, even if \(|\Delta_-|\) is quite substantially greater than zero, then the p-value distribution must also be flat if the power of the test is near \(\alpha\), which is as low as it can get.\(^{19}\) With \(\alpha = 0.05\) and the histogram having 20 bars it’s a matter of the area under the last 19 bars, which is the probability of a type II error, needing to sum to \(1 - \alpha = 0.95\). Flatness brings that about.

Figure 3 illustrates that tendency. By inspection of Figure 1, the power is decreased as the number of observations is reduced. To produce Figure 3, I lowered the number of observations to just 36, corresponding to just three years of monthly data, to deliberately lower the power. But I have retained the same \(|\Delta_-|\) value that is substantially above zero, and the same \(\rho_{ab}\) value of 0.85 with is rather high. Now the power has not been decreased all the way to \(\alpha\), which is 0.05; from the legend, it is instead approximately \(1 - 0.85\) or \(1 - 0.80\), which is 0.15 or 0.20. That’s why the histogram is not completely flat. And the discrepancy between the Ledoit and Wolf histogram and that of the asymptotic approximation, such as it is, may have arisen due to the asymptotic formula for the p-value not working as well with only 36 observations.

With 1,000 trials and 20 histogram bars an utterly flat histogram would show 50 p-values in each bar. I reran the code that produced Figure 3, again with 36 observations but with \(\Delta_-\) reduced to 0.05 and \(\rho_{ab}\) set to 0. That still means that \(H_1\) is true. The histogram became truly flat—with the number of p-values in each bar generally varying between 40 and 60 with no discernible downtrend in the direction of the higher p-values. And the power was about 0.055—very close to \(\alpha = 0.05\).

If this is not yet clear, if in the circumstances of Figure 2 the Ledoit and Wolf (2008) code was run and a p-value under 0.05 was found, there was little chance that a value higher than 0.05 might instead have been produced because the few bars above 0.05 that are not empty don’t contain many of the trials. But in the low-power circumstances of Figure 3 if the Ledoit and Wolf code had been run and some p-value had been found, there was every chance that a very different value might instead have been produced. So, in the circumstances of Figure 3 whatever is produced for the p-value is utterly unreliable.

Ledoit and Wolf (2008, 858) state the following: “We have discussed alternative inference methods which are robust. HAC inference uses kernel estimators to come up with consistent standard errors. The resulting inference works well with large samples but is often liberal for small to moderate sample sizes. In such applications, it is preferable to use a studentized time series bootstrap.”\(^{20,21}\) By

---

19. Wasserman (2004, 157) does state the following: “Warning! A large p-value is not strong evidence in favor of \(H_0\). A large p-value can occur for two reasons: (i) \(H_0\) is true or (ii) \(H_0\) is false but the test has low power.”


21. An identical statement appears in the subsequent article Ledoit and Wolf 2011. And in that follow-up
“studentized time series bootstrap” the authors are referring to the method of their article, which they do describe as being “quite complex.”

But it is clear, when Sharpe ratio differences are the test statistics, that small and moderate sample sizes automatically have low or moderate power—making p-values randomized and hypothesis testing infeasible, as shown by Figure 3. Thus, it’s hard to imagine circumstances in which it would be advantageous to use the quite complex Ledoit and Wolf method in lieu of one of the simpler HAC inference schemes.

Figure 3. A low-power example with just 36 observations

Notes: Here $\Delta = S_h - S_h$ is the a priori population value of the Sharpe ratio difference. And $P(p\text{-value} > 0.05)$ is the probability of a type II error, whose complement is the power of the test.

In the low power limit—due to some combination of low $|\Delta|$, low $\rho_{ab}$, or the duration of the backtesting interval not being long enough—the p-value estimator of Ledoit and Wolf and equation (3a) with the standard error derived article, which is about the variance rather than about the Sharpe ratio, the word power does not appear.
from equation (1a) both absurdly become random number generators.

The characteristic form of Figures 2 and 3 is not actually limited to i.i.d. bivariate-normal returns. Whatever the autocorrelations, heteroskedasticity, etc., in the returns of the paired portfolios, in the asymptotic limit the entire histogram is only dependent upon the power of the test and the significance level $\alpha$.

To derive $pdf_{p_2}$, the sampling distribution of the two-sided $p$-value, we can start by presenting it as an alternative means of calculating the power of the test:

$$\pi_2(\alpha) = \int_0^\alpha pdf_{p_2}(p_2) \, dp_2'.$$

But we don’t have to be limited to evaluating the power function $\pi_2$ at just $\alpha$. Since $\alpha$ can be given any value, give it the value $p_2$. Then

$$\frac{d\pi_2}{dp_2} = pdf_{p_2}.$$  

Writing $Z_{p_2} \equiv \Phi^{-1}(1 - p_2/2)$ and $Z_{true} \equiv \Delta / \hat{SE}$, then

$$\pi_2(p_2) = 1 - \Phi\left(Z_{p_2} - Z_{true}\right) + \Phi\left(-Z_{p_2} - Z_{true}\right).$$

And

$$\frac{dZ_{p_2}}{dp_2} = -\frac{1}{2} \Phi'(Z_{p_2})^{-1}.$$  

Putting it all together,

$$pdf_{p_2}(p_2) = \frac{1}{2} \left[ \Phi\left(Z_{p_2} - Z_{true}\right) + \Phi\left(-Z_{p_2} - Z_{true}\right) \right] / \Phi'(Z_{p_2})$$

from the chain rule of differential calculus. For every value of the power $\pi_2$ there is a unique value of $|Z_{true}|$. If the $Z_{true} > 0$ choice is made, as in Figures 2 and 3, then the equation for the two-sided power $\pi_2(\alpha)$ can be inverted to find $Z_{true}$, and that value can be substituted for $Z_{true}$ in $pdf_{p_2}$. Thus $pdf_{p_2}$ is only dependent upon the power $\pi_2$ and the significance level $\alpha$. Upon coding the expression for $pdf_{p_2}$ I find that, with the $\pi_2(\alpha)$ values that can be inferred from Figures 2 and 3, it closely agrees with those histograms.

**Conclusion**

I have shown that there are formidable limitations on the use of hypothesis testing with the Sharpe ratio difference between a pair of portfolios being the test statistic, that are innate to that statistic. Investors should accordingly be wary of claims by portfolio managers that their Sharpe ratio exceeds the ratios of other managers. It is advisable to avoid hypothesis testing when there is good reason to believe that the power of the test is too low—such as the sample covering a small interval of time, the correlation coefficient between the returns of the two portfolios not being high, or the reasonably expectable size of the true Sharpe ratio.
difference being small. If the power of the test is too low there is really no good fix for that, unless somehow pertinent additional data become available that span additional months. But confidence intervals do in any case provide a reasonable recourse. A secondary hypothesis test can be of some help in sorting out type II errors, but it’s no panacea.

**Code**

The source code and data that were used to produce the figures is available from the journal website (link).

**References**


Torricelli, Costanza, and Beatrice Bertelli. 2022. The Trade-Off Between ESG Screening and Portfolio Diversification in the Short and in the Long Run. Working paper. Link
About the Author

With a B.S. in physics and mathematics from Tulane University, **Michael O’Connor** worked for the United States Navy as a physicist and then went on to graduate studies. His Ph.D. is in physics from Stanford University. For 25 years he owned and operated a Silicon Valley business that conducted noise and air quality studies, often of major transportation projects. Today, O’Connor runs a startup consultancy, MO’C Portfolio Analytics in Washington State. His email address is mike@mocpa.com.
Classical Liberalism in Russia

Paul Robinson

Eamonn Butler (2015, 3–11) remarks that classical liberalism is “a presumption in favour of individual freedom,” which is guaranteed through “limited and representative government…the rule of law…spontaneous order…property, trade and markets…civil society…[and] common human values.” According to Butler, “classical liberal freedom is essentially negative” (ibid., 34). It is distinguished primarily by its support for freeing people from governmental incursions on their persons, property, and freedom of association. Positive definitions of freedom, by contrast, emphasize the individual’s ability to do or to be certain things. Classical liberals care about human flourishing, too, of course, but human flourishing is not their definition of freedom or liberty. Expansive positive definitions of freedom are often associated with modern liberalism or social liberalism, which is more favorable to activist government. In this article I speak of various liberalisms, but with a mind for distinguishing among them.

Butler’s definition of classical liberalism identifies two levels of analysis: one is that of basic principles, above all, freedom; the other is that of the institutions—representative government, property, markets, etc.—that are supposed to maintain those principles. When we examine the intellectual history of Russian thinkers identified as ‘liberal,’ we observe ongoing interest in ‘freedom,’ but we also find a diversity of views about institutions. Those in Russia who have self-identified or been identified by others as liberals have desired to expand personal liberty. But they have not always regarded representative government, free speech, free markets, and the like as the best mechanisms for doing so. They have supported some of those institutions generally considered liberal while rejecting others—promoting the expansion of civil liberties, for instance, while rejecting political liberties, or promoting political liberties while viewing free markets with

1. University of Ottawa, Ottawa, ON K1N 6N5, Canada.
some suspicion. Meanwhile, those considered ‘reactionary’ have in some respects on occasion been closer to the tenets of classical liberalism than those considered ‘liberal.’

In the 1860s and 1870s, for instance, the so-called ‘aristocratic opposition’ demanded the introduction of representative government (with a limited, property-based franchise), the abolition of the peasant commune, and the reorganization of the rural economy along free-market lines. All this, they reckoned, would guarantee the aristocracy’s continued dominance of Russian society. Liberals, meanwhile, rejected most elements of this program—for the same reason, that it would cement the power of the aristocracy (Robinson 2019, 94–95). Historian Daniel Field thus notes that in the circumstances of mid-nineteenth century Russia, “Doctrines naturally clustered together in Western Europe were in conflict in Russia. … The espousal of constitutionalism or laissez-faire economics was regarded, often correctly, as an attempt to perpetuate the dependence of the peasantry and the dominance of the nobility. … [A reformer] could not…embrace the whole bundle of liberal doctrines. Different men, grasping different parts of the bundle, naturally came into conflict” (1973, 60).

One can see a somewhat similar phenomenon in later periods. In Russia, liberal ideas have often combined with non-liberal and illiberal ideas in ways that are perhaps unique to it. The result is that there is considerable dispute as to who in Russian history deserves the liberal label. Some historians have claimed that key figures and institutions in the history of Russian liberalism such as political philosophers Konstantin Kavelin and Boris Chicherin and the pre-revolutionary Kadet Party were not actually liberal at all, while at the same time applying the liberal label to people and institutions who were not in their own time considered liberal but revolutionary or reactionary, such as the writer Alexander Herzen and the pre-revolutionary Octobrist Party. All this points to the difficulty of defining liberalism in a Russian context.

One reason for this confusion may be that from its earliest days Russian liberalism has rested on a rather different social base than its Western European and North American counterparts. As Pavel Miliukov, the leader of late Imperial Russia’s main liberal party, the Kadets, put it: “Russian liberalism was not bourgeois, but intellectual” (Milioukov 1906, 226). Liberal ideas arrived in Russia long before a large bourgeois class came into existence, and when the latter did emerge its members tended to be quite conservative due to the dependence of merchants and industrial producers on state orders and the consequent tight links between trade, industry, and the state. Liberalism, whether in the ‘classical’ sense or a more ‘modern’ or ‘social’ sense, became the ideology of that segment of the aristocracy that had taken up professional work—university professors, lawyers, doctors, and so on. Later, in the Soviet Union, liberal modes of thinking were associated
primarily with a narrow elite within the scientific community referred to as the technical intelligentsia (Lipovetsky 2013, 109–139). Nowadays they are commonly associated with urban professionals, or what are sometimes called the creative classes—journalists, artists, academics, IT workers, and the like. Russian liberalism, in the broad sense, reflects this group’s culture and interests, which have often been widely at odds with those of both the state and the mass of the Russian population.

The intellectual classes who have dominated Russia’s liberal movement have tended to be highly Westernized individuals, who speak Western European languages, have studied Western European philosophy, and go to Western Europe for their holidays. Such characteristics have separated them from the bulk of their fellow countrymen. Liberals have also often been positivist and rationalist in outlook, viewing human society as driven by scientifically determinable rules, regarding history as an inexorable process towards a known end—that being a liberal society in line with Western models.

Consequently, Russian liberalism, in the broad sense, has been, and still is, as much a cultural as a political or socioeconomic phenomenon. Its aim is to culturally transform Russia so as to make it what is often called a ‘normal’ country, by which is meant a Western one. “The Russian liberal is a thoughtless fly buzzing in the ray of the sun; that sun is the sun of the West,” said philosopher Pyotr Chaadaev in the early nineteenth century (quoted in Gutorov 2017, 11). Similar criticisms can be heard today. This pronounced Westernism has had some negative consequences. It has alienated liberals from the less pro-Western majority of the Russian population and saddled liberalism with the reputation of being an anti-national force. This helps to explain its political failure.

Another feature separating Russian and Western classical liberalism is the former’s attitude towards the state. Classical liberalism has sought to carve out a zone in which individuals can act independently of the state. It has therefore aimed if not to minimize then at least to severely limit the state’s competencies. Russian liberals’ relationship with the state has been more complex. On the one hand, liberals have often deeply disliked the Russian government of their time, regarding it as an oppressive force that must be opposed. At the same time, though, liberals have generally accepted the reality that in Russia it has been the state that has historically been the primary driver of reform and Westernization. While distrusting the government, Russian liberals have rarely shared classical liberalism’s distrust of the state per se, but instead have looked to the state as the means by which liberal values and institutions will be advanced. Furthermore, liberals have also generally supported the Russian model of a highly centralized system of authority, with power concentrated in the hands of the executive branch of government. Overall, Russian liberalism is considered to have a “statist orientation” (Poole 2015, 170).
Western liberalism has opposed political radicalism. It has generally presupposed an equality of subjection under a law-based government, and proposed mere policy reform (liberalization), not constitutional reformation. In Russia, liberals’ vision of the state has been a centralized but more law-based government. Referring to the Imperial period, Andrzej Walicki says that “the main concern of Russian liberal thinkers was the problem of the rule of law, and the most precious legacy of Russian liberalism was precisely its contribution to the philosophy of law.” Law, says Walicki, was “the core value in the liberal view of the world. … It is no exaggeration to say that the entire history of liberal thought in Russia revolved around the problem of the rule of law and the rule-of-law state” (Walicki 1992, 1, 3, 402). To a large extent this remains true today. Marc Raeff (1959, 223) notes, however, that the establishment of the rule of law in the Western sense has been impossible within the confines of the existing political order, given its rejection by successive rulers. Consequently, “introduction of real legality could only mean the overthrow of the existing regime.” The result has been a “blurring of the line between radicalism and liberalism,” and led some liberals in both the Imperial period and today into the revolutionary camp.

**Early Russian liberalism**

The origins of Russian liberalism date back to the reign of Catherine II in the late eighteenth century. An often cited starting point is the *Instruction* issued by Catherine early in her reign, which, it has been said, “opened the doors in Russia to the liberal ideas of the European Enlightenment” (Novikova and Sizemskaja 1993, 126). Catherine soon pulled back from the promises of reform made in the *Instruction*, but the document did provide an opportunity for Russians to come forward with ideas which might merit the liberal label. An example was a response to the *Instruction* written by a professor at Moscow University, Semyon Desnitsky, who had studied in the classroom of Adam Smith in Glasgow, and who recommended the establishment of a permanent advisory body, the Senate, that was to be elected by landowners, merchants, artisans, and teachers in higher educational institutions (Hamburg 2016, 589). Desnitsky also proposed the introduction of jury trials and suggested that judges be appointed for life (ibid., 590). As Desnitsky’s Senate was to be advisory only, his proposal left the absolute monarchy in theory intact. Nevertheless, it represented an attempt to blend elements of Russian absolutism with liberal institutions such as representative government.

Desnitsky was among those who propagated Smith’s ideas. Gary Hamburg (2016, 595) remarks that, “the degree of Smith’s influence on Desnitsky’s economic
thinking was striking: hostility towards consumption taxes, preference for progressive taxes, the desire to avoid a heavy state presence in economic life, and the concern for production as the source of national wealth.” Desnitsky did not propose abolishing serfdom, but did suggest that peasants be given certain property rights, such as the “right to buy and sell moveable property” (ibid., 692–693).

The first translation of Smith’s *Wealth of Nations* appeared in Russia in four parts from 1802 to 1806, and is said to have influenced Russian government thinking in the first decade of the reign of Alexander I (Bennett 2013, 20). Nikolai Mordvinov (1754–1845), appointed head of the Department of State Economic Affairs in 1810, drew on the model of England, where he had studied and had fallen under the influence of both Smith and Jeremy Bentham. Mordvinov wrote that “Property is the cornerstone. Without it, without the permanence of the rights that guarantee it, neither laws nor the fatherland, nor the state can be of use to anyone” (quoted in Leontovitsch 2012, 34, 36).

Another admirer of Smith was Alexander Kunitsyn, who taught at the Lyceum in Tsarskoe Selo, and whose students included the poet Alexander Pushkin (Berest 2011, 45). “Kunitsyn’s classes,” says his biographer Julia Berest, “undoubtedly contributed to opening the minds of his students to liberal ideas” (ibid., 42). According to Berest, “Kunitsyn taught his students that man’s primary natural right is the right to one’s person, by virtue of which ‘every individual can demand from others that they treat him not as a mere tool to their ends but as a person endowed with reason and will’” (50). To Kunitsyn, the purpose of the state was to secure peoples’ freedom. “Freedom,” he wrote, “is the right of each person to act according to his will in all matters that do not harm others” (quoted in Berest 2011, 78).

Both Desnitsky and Kunitsyn promoted the concept of natural rights. Desnitsky, for instance, argued in a 1768 lecture that justice required recognition of such rights (Hamburg 2016, 603). Meanwhile, in his 1818 book *Natural Law*, Kunitsyn argued that the most fundamental right “is the right to one’s person,” which presupposed “the right to exist,” “the right to act,” and “the right to achieve well-being” (quoted in Berest 2011, 147). Since human desires varied, so too did understandings of well-being. Kunitsyn concluded that “each person has a right to choose the way of life and occupation which he finds conducive for his well-being” (ibid., 148). Following from this, Kunitsyn argued that the state should limit its activity to providing security and justice. “The subjects agree to obey the supreme power only for the sake of safety; therefore in all their private matters they remain free,” he wrote (ibid., 157).
Kavelin, Chicherin, and conservative liberalism

In the mid-nineteenth century, the predominant strain in Russian liberal thought was what is often called ‘Right Hegelianism.’ The most notable Right Hegelians were Konstantin Kavelin and Boris Chicherin, both of whom followed Hegel in viewing history as a process involving the gradual expansion of liberty, culminating in the institution of the modern state. In 1847, Kavelin produced an essay entitled “A View of Juridical Life in Ancient Russia,” in which he argued that history involved the gradual development of the autonomous individual, or as Kavelin (1989, 23) put it, the “principle of personality.” Russian history, he claimed, passed through various stages—communal, tribal, and family—before reaching the era of the state. In the earliest stages, strong blood ties meant that people did not distinguish between themselves and others (ibid., 22). By contrast, “The appearance of the state was a liberation from an existence based purely on blood, and was the basis for the independent action of the person” (ibid., 48).

Kavelin (1989, 65–66) argued that as this process developed, “The Russian and the foreign have merged into one to carry Russia forward… The boundaries between the past and the present, Russian and foreign, are being destroyed.” The progression was not a matter of Russia copying the West. The process of moving towards the state and the principle of personality was a universal phenomenon, not something specifically Western. Thus, concluded Kavelin (1989, 66), “The difference [between the West and Russia] lies solely in the preceding historical facts; the aim, the task, the aspirations, the way forward are one and the same.” Certain universal goods were ‘Western’ only in the sense that the West was the first to approach them.

In the mid-nineteenth century context, representatives were nearly always elected by means of a property-based franchise that gave disproportionate influence to the wealthy (Walicki 2015, 453). Both Kavelin and Chicherin believed that in Russia representative institutions would serve the narrow class interests of the nobility rather than the interests of the people as a whole. In an 1875 article, Kavelin argued “the government has lost all our respect and trust” (Kavelin 1996, 89, 92). The solution did not, however, lie in parchment constitutional reforms, including the creation of an elected parliament. Kavelin wrote:

A constitution only makes sense when a well organized and authoritative wealthy class supports and protects it. Without such a class a constitution is a worthless scrap of paper, a lie, a prelude to the most dishonest and
dishonorable deceit. … By itself, a constitution doesn’t give or guarantee anything, absent these conditions it is nothing, but a harmful nothing because it deceives with the external form of political guarantees. (Kavelin 1996, 108)

Kavelin’s conservatism extended also to economic affairs. The abolition of serfdom in 1861 left intact the peasant commune as the main institution governing the lives of Russian peasants. Kavelin viewed the commune as a counterweight to the process of social radicalism that he believed resulted from industrial development. He also argued that progress did not require the transformation of communal property into private property. “Private property…is a source of movement, of progress, of development; but it becomes a source of death and destruction, it corrodes the social organism when its extreme consequences are not moderated and balanced by other landowning principles. … I see communal landowning as one such principle,” Kavelin wrote (Gorlov 2012, 47).

Kavelin’s beliefs may be described as a conservative liberalism that looked to the autocratic monarchy as the driver of progress and sought to initiate political change at the local level, only gradually moving up to the national level as the necessary political culture developed. This conservative liberalism is even more strongly associated with Chicherin.

Chicherin (1998a, 117), in an 1858 essay entitled “Contemporary Tasks in Russian Life,” which has been published in this journal (link), argued that, “Government activity must not preclude the autonomy of the people, for popular autonomy is a basic precondition of public life. … For the government to establish norms of behavior and opinion and to bend everything to these norms, to render the people voiceless and silent before the government, is to kill any life and to destroy one of the fundaments of society.” “We need freedom!” wrote Chicherin. “We want the opportunity to freely express and develop our thoughts, so the Tsar will know what Russia is thinking and can govern us with a clear understanding of social and economic conditions” (1998a, 133–134).

“Liberalism! This is the slogan of every educated and sensible person in Russia,” wrote Chicherin. In practice this meant, “freedom of conscience,” “emancipation from servile status,” “freedom of speech,” “freedom of the press,” “academic freedom,” “publication of all government activities,” and “public legal proceedings.” “In liberalism…is Russia’s future; it alone can awake Russia to new life,” he concluded (1998a, 134–139).

In an 1862 article, Chicherin identified three types of liberalism: “street liberalism,” “oppositional liberalism,” and “okhranitel’nyi liberalism,” the last of
which may be roughly translated as “protective liberalism.” Chicherin favoured the last of these. “The street liberal,” he complained, “feeds on irreconcilable hatred of everything that rises above the crowd, of all authority. It never occurs to him that respect of authority is respect of thought, of labor, of talent, of everything that gives mankind higher reason” (Chicherin 1996, 41–42). Meanwhile, oppositional liberalism “doesn’t seek to achieve any sort of political demands, but takes pleasure in the glory of the oppositional position” (ibid., 44). By contrast,

The essence of okhranitel’nyi liberalism consists of reconciling the principle of freedom with the principles of power and law. In political life its slogan is “liberal measures and strong government,” liberal measures that enable society to act independently, that guarantee the rights and personhood of citizens, that protect freedom of thought and conscience, that allow one to express all lawful desires; and strong government, the guardian of state unity, that connects and restrains society, preserves order, severely ensures obedience of the law, and punishes any breaches of it (Chicherin 1996, 49).

Chicherin did not believe that Russia was ready for constitutional government, and he therefore supported the retention of the autocratic system, albeit with expanded civil liberties (Chicherin 1998c, 365). In his book On Popular Representation, Chicherin (1998b, 162) argued that political liberty was dependent upon the existence of the appropriate political and legal culture. This culture, he believed, was a product of owning property. Thus, he wrote, “There is therefore nothing ethically troubling in denying political rights to poor people.” Chicherin (1998b, 206) concluded: “Under a given set of circumstances, and taking into account the political sophistication of a people, one must decide whether the advantages [of political liberty and representation] outweigh the disadvantages. The conclusion will not always be the same, and, for this reason, representative government is not always appropriate.”

Chicherin also played an important role in combating the dominant legal theory of the time—legal positivism—and promoting instead ideas of natural law. Chicherin argued that law was a means of limiting external freedom for social purposes. As such, it had no business intruding upon individuals’ inner freedom. Chicherin argued that “the essence of law…has a purely external character” (quoted in Yevlampiev 2009, 119). He wrote: “I am not in the least inclined to treat law as an expression of interests; on the contrary, I see it as a manifestation of the eternal principles of justice… I think, however, that the law of justice, which requires that everyone be given his due, should not be confused with the law of love, which demands sacrifice for the sake of one’s neighbours” (quoted in Walicki 1992, 154).

In other words, the law should enforce contracts but not moral obligations.
In Chicherin’s eyes, although the rich have a moral obligation to help the poor, they should not have a legal obligation to do so. Chicherin similarly argued that workers should not be forced to contribute to health or unemployment insurance, as this would infringe upon their inner freedom of choice. He supported laissez-faire economics and opposed government-sponsored social welfare schemes. “Inequality predominates in human communities,” he wrote. Inequalities were in part due to natural differences among humans, and in part due to the fact that freedom inevitably produces inequalities as a result of the different choices people make. According to Chicherin, “Freedom by its nature, leads to inequality. In the sphere of property this rule shows itself fully operative, yet one cannot destroy this inequality without destroying its root—that is human liberty” (quoted in Hamburg 2010, 122).

**New Liberalism in late Imperial Russia**

By 1900, both Chicherin’s faith in the reforming potential of the autocracy and his preference for economic liberalism had fallen out of favor among Russian liberals. In the decades leading up to 1900, the conservative reigns of Tsars Alexander III and Nicholas II, as well as international intellectual currents of the late 19th century, caused Russian liberals to demand that Russia be transformed into a constitutional monarchy with representative institutions and that the state play a more active role in the economy.

An important influence on liberal thought in this era was the writing of Russian religious philosopher Vladimir Solovyov. Like Chicherin, Solovyov helped to promote ideas of natural law, writing that “law is not determined by the concept of utility, but contains within itself a formal moral principle” (Soloviev 2000a, 136). “The rule of true progress,” he claimed, “is this, that the state should interfere as little as possible with the inner moral life of man, and at the same time should as securely and as widely as possible secure the external conditions of his worthy existence and moral development” (ibid., 459).

Liberal legal scholars such as Pavel Novgorodtsev and Leon Petrazycki picked up on these claims. Novgorodtsev, for instance, wrote of a crisis of legal consciousness that was the product of the domination of the theory of legal positivism. This had the effect of making people regard law as being law solely because the state had deemed it such rather than because it was based on any universal moral principle. Because people regarded the law as being entirely a product of force, they had no respect for it (Walicki 2015, 734). The solution, said Novgorodtsev, was a revival of natural law.

In line with the idea that law must be founded on morality, liberals concluded
that economic policy likewise must have a moral foundation, including providing everyone with the economic means to enable a dignified existence. This also owed much to Solovyov. In an 1897 essay titled “The Social Question in Europe,” Solovyov declared:

The principle of equality in its true sense is that all men are equal. … Each represents intrinsic value and possesses an inalienable right to an existence corresponding to his human dignity. The raison d’être of society in relation to its members is to assure for each not solely a material livelihood, but moreover a dignified livelihood. Now it is clear that poverty beyond a certain threshold…is contrary to human dignity and therefore incompatible with true public morality. Therefore, society must insure all its members against this degrading poverty in securing for each a minimum of material resources. (Soloviev 2000b, 32–33)

Members of the liberal Kadet party, founded in 1905, took up this mantra. Novgorodtsev, who became a member of the party’s central committee, wrote that:

Securing the right to a dignified human existence has in mind people who are suffering from economic dependency, from a lack of means, from unfortunate circumstances. … The use of freedom can be completely paralyzed by a lack of means. The task and essence of law is the protection of personal freedom, but to achieve this goal one must care for the material conditions of freedom, without which freedom can remain an empty word. … Thus, in the name of protecting freedom the law should be concerned with its material conditions; in the name of personal dignity, it must be concerned with defending the right to a dignified human existence. (Novgorodtsev 1911, 5–6)

Novgorodtsev (1911, 9, 11, 12) said the state should pass legislation to protect workers, for instance regulating sanitary conditions in the workplace, providing insurance for illness and old age, and legalizing trade unions. Novgorodtsev was a great admirer of British Prime Minister Herbert Asquith (1852–1928), whose Liberal Party government at the start of the twentieth century exemplified what was known as ‘New Liberalism.’ This abandoned the laissez-faire principles of classical liberalism in favor of a more interventionist approach. Kadets such as Novgorodtsev rejected state ownership of the means of production, but at the same time expressed an indebtedness to socialism as well as a belief that liberalism and socialism shared much in common. As Miliukov put it: “In studying the history of the liberal and socialistic currents, we have found that the chasm existing between them at their inception was perpetually narrowing…the utopian element…slowly but steadily vanishing from the socialistic programs” (Milyoukov 1906, 561).
In some respects, the supposedly reactionary late Imperial state was arguably closer to the tenets of classical liberalism than new liberals were. Under Prime Minister Pyotr Stolypin, the government introduced a reform allowing peasants to withdraw from the commune and consolidate their land into a single block that would be their own personal property. Stolypin argued that “The Government wants above all to promote and enhance peasant land ownership. It wants to see the peasant earning well and eating well… But for this it is necessary to give opportunity to the capable, industrious peasant… He must be given the chance to consolidate the fruits of his labor and consider them his inalienable property” (1964, 462). Some liberal economists, such as Boris Brutzkus, supported this reform. But most liberals rejected it. Fearing that the reform would create divisions between rich and poor peasants and so inflame revolutionary tensions, the Kadets accused Stolypin of imposing a foreign form of land ownership without respect of Russian traditions (Egorov 2010, 2010). They proposed instead the expropriation of noble and church property and its consolidation into a national land fund from which peasants could lease land for their own use. Novgorodtsev justified this by saying:

Adherents of the old dogma that derives from the principle of holy and inviolable property see this presentation of the problem [i.e., the expropriation of land] as a perversion of the idea of law. But the legal consciousness of our time places the rights of the human person above property rights and, in the name of this law, in the name of human dignity, in the name of freedom, rejects the idea that property is inviolable, and replaces it with the principle of public-legal regulation of acquired rights with necessary compensation for their owners in the case of alienation. (Novgorodtsev 1911, 10–11)

What made Russian liberals of this era ‘liberal’ was not, therefore, their economic policies. The meaning of ‘liberal’ had drifted.

The liberals of this era insisted on the need for the introduction of a constitutional order, meaning a new law-based constitution including representative institutions. Indeed, it is noticeable that late Imperial liberals generally referred to themselves not as ‘liberals’ but as ‘constitutionalists.’ Convinced by the conservative reigns of Alexander III and Nicholas II that the autocrat state could not be a source of liberal policies, they became full-fledged proponents of the need for constitutional reform and demanded a government that would be responsible to a representative assembly elected directly via a universal franchise. To pursue this goal, they allied with revolutionary forces and refused to condemn political violence, hoping that this violence would force the Tsarist regime to make the necessary concessions. At the same time, they refused to cooperate with the Russian state. In October 1905, for instance, the Prime Minister, Sergei Witte, met with leading members of the Kadet party and offered them positions in his
government. The Kadets turned down the offer, demanding that the government summon a constituent assembly and grant an amnesty to political prisoners (Enticott 2016, 152). Similarly, in January 1907, Stolypin offered to legalize the Kadet party (it had never been legally registered) if it publicly denounced revolutionary terrorism. The Kadets refused (Riha 1969, 140–141).

One of the most prominent Kadets, Vasily Maklakov noted that the Kadet party “was created to fight against autocracy” (Enticott 2016, 178). “We believed that condemning political murders would mean allowing the authorities to believe that they were right,” he added (Egorov 2010, 205). The result was that following the 1905 revolution, Russian liberals passed on the opportunity to take a share of power and chose instead to remain in opposition.

Underlying liberals’ insistence that the Russian state must adopt a democratic constitution was a belief that once the Russian people acquired civil and political liberties they would abandon any revolutionary inclinations. As one of the leading liberal thinkers of the period, Pyotr Struve, wrote:

> The only way to direct the enormous social movement presently stirring Russia’s urban and rural population into the channel of lawful struggle for their interest is to invite the entire population, on equal rights, to share in the political life—that is to institute universal franchise. Give political freedom and political equality, and life itself will freely sweep away all that which is premature and unrealizable in radical programs. … Universal franchise…will bring no horrors and no miracles; the masses, called on to participate in political and social construction, will astonish us neither with their obscurantism nor their radicalism. … Under the universal franchise, the masses, having become responsible members of their own destiny, will understand what is necessary and what is not. (quoted in Pipes 1970, 381–382)

### Responses to the Russian Revolution

The events of 1917 suggested otherwise. After Tsar Nicholas II abdicated in March 1917, liberals briefly held the reins of power. The Provisional Government that ruled Russia from March to November 1917 contained a number of prominent liberal politicians. These included Miliukov (who was foreign minister from March to early May 1917), Prince Georgy Lvov (head of the Provisional Government until July 1917), and Andrei Shingarev (who served first as agricultural minister and then as minister of finance). This first experience of Russian liberals in government did not end well. Although there was no general vote or referendum involved in those events, the fact is that, when the governing structures of the Imperial regime were
leveled, the masses seemed to fall in behind the Bolsheviks. In that sense, they chose not the path of political moderation but that of radicalism. Liberals suffered terribly, with liberalism being almost entirely extinguished under the communist government that took power in November 1917. Such liberals as survived fled into exile, where many reassessed their beliefs. Faith in democracy collapsed, replaced by a belief that a post-communist Russia must by necessity undergo a period of dictatorial rule in order to restore order. As Novgorodtsev said in May 1919, “If nothing remains of our democratism, then that is an excellent thing” (quoted in Rosenberg 1974, 410).

Contemplating their own experiences in 1917, and observing the collapse of democratic states such as Italy and Germany in the 1920s and 1930s, many Russian émigrés came to the conclusion that the roots of communism and fascism were the lack of the spiritual values required to maintain a liberal order, such as faith in God, patriotism (rather than class identity), respect for the law, and so on. Novgorodtsev stated that “Naïve and immature political thought usually supposes that it is enough just to overthrow the old order and proclaim freedom of life, electoral rights and the constituent power of the people, and democracy will come into being all by itself” (quoted in Gusev 2001, 110). Reality was very different. To create a liberal order, there “must be a people that has matured to govern itself, knowing its own rights and respecting other people’s” (ibid., 111). This was only possible, Novgorodtsev argued, if democracy was guided by a “Higher Will,” by which he meant religion. He wrote: “The whole world is living through a crisis of legal consciousness. And the most important and fundamental thing in this crisis is a crisis of non-belief, a crisis of culture, torn from religion, a crisis of the state, which has become disconnected from the church” (ibid., 113).

Other notable émigrés such as Nikolai Berdyaev, Georgy Fedotov, and Fyodor Stepun, all of whom wrote for the journal Novyi Grad (New Town), agreed. Fedotov, for instance, blamed the rise of communism on Russia’s “abandonment of Europe’s high humanistic traditions” (quoted in Kara-Murza 2009, 206). Rather than appreciating the humanistic traditions behind liberal ideas, Russian liberalism, Fedotov complained, “has long been fed on journeys abroad, on superficial rapture at the wonders of Western civilization, accompanied by a total inability to link its enlightening ideals with the forces propelling Russian life” (ibid., 207). Liberals lacked roots in their own country’s history, he said. The result, he concluded, was “the illness of antinationalism. … Russia itself became an object of hatred” (ibid.). The solution, he believed, lay in a revival of Christianity. Likewise, philosopher Semyon Frank, once a passionate advocate of constitutional reform, in exile turned against it. “In Russian liberalism,” he explained, “belief in the value of spiritual principles such as nation, state, law and freedom remains philosophically unexplained and lacking in religious inspiration” (quoted in Kantor 2007, 857).
Boris Brutzkus’s criticism of economic planning

In this way, classical liberalism came under attack as lacking in spiritual foundations and as such being unable effectively to protect liberty. Insofar as classical liberalism survived in exile, it was among economists, notably Boris Brutzkus. Brutzkus’s book *Economic Planning in the Soviet Union*, which contained a foreword by Friedrich Hayek, is seen by some as having influenced Hayek’s own denunciation of socialism *The Road to Serfdom* (Wilhelm 1993, 343–357).

In opposition to Marx’s labor theory of value, which argues that value is a product of the amount of labor expended on producing something, Brutzkus (1935, 25–26) argued that value is a product of social need, as reflected in prices. According to Brutzkus, to replace the guidance provided by prices the socialist state would have to resort to a huge bureaucratic apparatus dedicated to collecting statistics about the population. But this apparatus would never be able to determine the subjective tastes of millions of distinct individuals (ibid., 44). Lacking proper information, investment decisions in a socialist economy were bound to be divorced from economic realities. The result would a considerable waste of resources (Kojima 2008, 127).

Surveying the first Soviet five-year plan (1928–1932), Brutzkus (1929, 430) commented that the Soviets had destroyed the “spontaneous regulators of economic life.” He wrote that “it would be completely wrong to think that any sort of stable compromise can be found between communism and capitalism. … In order for Russia to escape from the dead end into which it has fallen, communism must be finally overcome and all traces of it eliminated from national life” (ibid., 474).

Perestroika: Socialism with a human face

In due course, market liberalization would return to the Soviet Union and thereafter help to guide the government of the Russian Federation after the collapse of the USSR. Prior to that, a form of what one might call ‘Soviet liberalism’ did manage to emerge following Stalin’s death in 1953, but it owed little if anything to classical liberalism. Instead it found its inspiration in socialist thought and its aim was ‘socialism with a human face.’ In the economic sphere, it took the form of ‘market socialism,’ a concept that retained state ownership of the means of
production but sought to improve its efficiency by giving enterprises more leeway to determine their own output, set their own prices, and so on.

One of the most prominent proponents of market socialism was Tatyana Zaslavskaya of the Novosibirsk Institute of Economics and Industrial Organization. Zaslavskaya (1989a, 162, 168) argued that the Soviet economy could no longer expand by means of increasing inputs but needed instead to improve productivity. The improvement, she said, would require that workers and enterprises be given “a sufficiently wide margin of freedom,” while the economy as a whole needed “far more active use of ‘automatic’ regulators in balancing production, linked to the development of market relations.” Zaslavskaya (1989b, 123) stated that “we must have economic pluralism—not just state ownership, but also cooperative and individual ownership.” Nevertheless, she rejected the idea of large-scale private ownership of the means of production and so fell far short of endorsing free-market capitalism.

Mikhail Gorbachev’s program of perestroika in the late 1980s constituted an attempt to put socialism with a human face into practice. Perestroika was chiefly a program of liberalization of speech, association, and political activity. Nearly all restrictions were removed. Contrary to Gorbachev’s hopes, however, such liberalization did not have the effect of boosting the economy and strengthening the state. On the contrary, the Soviet economy, which had not yet been significantly privatized or liberalized, imploded, while social, political, and interethnic tensions skyrocketed. According to one report, by late 1989, “only 11 percent of 989 consumer goods monitored by an economic research organization were readily obtainable. Almost entirely absent from stores were televisions, refrigerators, washing machines, most household cleaning products, furniture of all sorts, electric irons, razor blades, perfumes and cosmetics, school notebooks and pencils” (Taubman 2017, 450). Given this reality, by 1990 it had become clear to most people that socialism with a human face had failed.

**Shock therapy:**
**Lifting controls and privatizing resources**

Looking for an alternative model, Russian intellectuals moved *en masse,* and in a very short period, to what was seen as the most successful alternative available—that of Western-style liberal democracy and free-market economics. The suddenness of this intellectual shift meant, however, that Russians were in many cases intellectually unprepared for what lay ahead, having only the slightest education in liberal theory and practice, and their understanding of democracy
and free markets was sometimes rather unsophisticated. They tended to view democracy as a system in which “democrats” held power and to be relatively unconcerned with issues of checks and balances (Sauvé 2019, 226). As for free markets, they tended toward the view that all that one had to do was enact economic freedom and all the country’s problems would be solved. Concern for the institutional underpinnings of a free-market economy were almost entirely absent.

Perhaps the first Soviet intellectual to unequivocally declare herself a believer in free markets was economist Larissa Piyasheva. “My views are based on the theories of the Chicago school, and I believe that all attempts to find a ‘third way’ [between socialism and capitalism] are headed for a dead end,” she wrote (Piyasheva 1991, 293). Journalist Igor Svinarenko described her views as follows: “Piyasheva talked a lot about the invisible hand... I couldn’t at all understand what kind of mechanism it was. But she kept on and on about this hand, which would immediately bring order everywhere, and that everyone would begin living a happy and rich life. I asked Piyasheva then, how could happiness just suddenly appear—after all, don’t needs and troubles always intensify at the start of any new capitalist period? She explained that problems occurred if you didn’t start out building the system correctly, but that if you did it right, there wouldn’t be any problems” (Koch and Svinarenko 2009, 193).

Such optimism was widespread among the liberal reformers who took power in Russia in the early 1990s under President Boris Yeltsin. One-time deputy prime minister Boris Nemtsov, for instance, remarked that “I believed that our main job then was to kill communism. If we managed that, we thought that we would live like the Americans, maybe in six months, perhaps nine months. … We thought things would work out in a short time” (Nelson and Kuzes 1995b, 141–142).

Underlying this optimism was a belief that Western economic and political models were products of universally valid social rules and could therefore be planted in Russia without much, if any, consideration of local conditions. As Pyotr Aven, Minister of Foreign Economic Relations from 1991 to 1992, put it: “There are no special countries. All countries from the point of view of an economist are the same” (Appel 2004, 167).

On 1 January 1992, the post-Soviet Russian government began a policy of rapid economic reform known as ‘shock therapy.’ Led by deputy prime minister Yegor Gaidar, this policy began by freeing prices on the vast majority of products and by removing most restrictions on private trade. The positive result was that goods began to appear once again in shops. The negative results included hyperinflation and a huge rise in crime and corruption. Between 1992 and 1996, Russia’s Gross Domestic Product (GDP) fell by 40 percent, industrial production by 50 percent, and real wages by 26 percent (Kazakevitch 2010, 128). The sudden
appearance of an unregulated market provided enormous opportunities for organized crime to extort payments from new businesses. This in turn led to demands for a restoration of state regulation, which in turn created opportunities for corrupt officials to demand bribes in return for trading licenses or for withholding licenses from competitors (Åslund 1995, 143–145). Well-positioned individuals whose contact with state officials gave them preferential access to rare resources were able to become phenomenally rich almost overnight by purchasing the resources cheaply from the state and then exporting them at world market prices (Åslund 1995, 150, 169).

As for the privatization process that followed, most enterprises found their way into the hands of former managers or well-placed individuals who were able to exploit their position to buy valuable companies at rock-bottom prices (Freeland 2000, 87). A few acquired so much wealth in the process as to acquire the label of ‘oligarch.’ These then used their newfound wealth to acquire political power, further corrupting the system as a whole.

Some liberals opposed shock therapy and proposed instead a form of ‘social liberalism.’ An example was the leader of the Yabloko Party, Grigory Yavlinsky, who argued that the “neoliberal-monetarist doctrine” did not fit Russia’s circumstances, because it “assumes the presence of a functioning market economy. … The special feature of the economies of the countries in the former socialist camp are such that standard monetarist methods yield different results there than they do in a developed market economy” (Yavlinsky 1992, 10). According to Yavlinsky, freeing prices in a highly monopolized economy, with few small businesses and a lack of property rights and other free market institutions, was a recipe for inflation (Nelson and Kuzes 1995b, 95). “Earlier, Moscow fixed prices; now the monopolist does it,” he said (Nelson and Kuzes 1995a, 42). According to Yavlinsky, the government should have demonopolized and privatized before freeing prices.

Another point of difference between the shock therapists and the social liberals was their attitude towards democratic institutions, with the former often regarding them as an obstacle in the way of reform. Rather simplified, social liberalism went hand in hand with political liberalism, whereas free-market liberalism often went hand in hand with a type of liberal authoritarianism. This indicated once again how different aspects of classical liberalism have often not gone together well in Russia.

The basic conundrum was expressed in 1990 by the future head of privatization under Yeltsin, Anatoly Chubais. Chubais noted that “There is a fundamental contradiction between the aims of reform (the forming of a democratic economy and society) and the means of their achievement, including measures of an anti-democratic nature” (ASEN 1990). ‘Shocking’ reforms are...
bound to have very negative consequences for many people. Democratic processes may therefore be unlikely to reform effectively, and perhaps not reform significantly at all.

Despite their declared favor for market liberalization, the shock therapists in some ways reflected their Marxist upbringing in that they tended to view economics as the substructure on which everything else depended. Liberalize economic activity, and everything else, including democracy, would naturally fall into place. As one prominent liberal intellectual, Igor Kliamkin, put it, “If we pretend that economic and political reforms advance in parallel, we know nothing (or don’t want to know) about the entirety of world history” (Kliamkin and Migranian 1989, 126). This logic dictated that anything which stood in the way of economic reform had to be resisted, including, if necessary, democracy.

**Contemporary Russian liberalism**

By 1993, popular opposition to shock therapy had grown considerably, and the Russian parliament sought to slow it down. In the face of this opposition, in October 1993 Yeltsin issued an illegal decree dissolving the parliament. In response, the parliament impeached him. Yeltsin, however, retained the support of the army and sent troops to blast the parliament into submission. That done, he then issued a new constitution that concentrated powers in the hands of the president. Most liberals applauded. But as Gaidar (1996, 252) noted, “It immediately became clear that the first casualty was democracy itself. On the morning of October 3, President Yeltsin was still only one of many players on the Russian scene. … On the morning of October 5, all the power in the country was in his hands. We had leapt from the gelatinous *dvoevlastie* [dual power] into a de facto authoritarian regime.”

A handful of contemporary liberals still stick to liberal authoritarianism. That attitude stems sometimes from a suspicion that Russian people are inherently reactionary and not to be trusted with power. An example is outspoken journalist Iuliya Latinina who argues that “It’s not enough to be a dictatorship, it’s necessary to be a good dictatorship, like in Singapore or Chile and not like in the Philippines or Haiti” (quoted in Gel’man 2010). This, though, is nowadays very much a minority view. Highly centralized power was acceptable as long as it was in liberal hands. For the past 20 years, however, it has been in the hands of Vladimir Putin, whose government has gradually pushed liberals out of public life and restricted civil liberties such as freedom of speech and association, especially since the invasion of Ukraine in February 2022. This fact has convinced many in the liberal camp that it would be better to have a system with more checks and balances.
Consequently, a consensus has emerged among liberals in recent years in favor replacing the presidential system created in 1993 with a parliamentary one. As journalist Vladimir Inozemtsev (2017, 8) comments: “The only way to deal with the current situation is to dismantle it completely—to make Russia a parliamentary instead of a presidential republic; to restore federalism in its true form and delegate powers to regional and local authorities.” Liberal political parties share this view. For instance, Yabloko (2016) declares that: “Fundamental political reform is necessary, changing the balance between the executive, the president, and the parliament in favor of the latter.”

The desire to change the balance of constitutional power reflects most liberals’ dislike of the Russian state, and represents a shift away from the historical preference for strong centralized state power as a necessary prerequisite of liberal reform. Nonetheless, a few so-called “systemic liberals” have chosen to continue working within the state system, particularly in the realm of economic policy. The most prominent of these is Alexei Kudrin, who was finance minister from 2000 to 2011, and head of the Accounts Chamber (in effect the chief auditor of Russian government expenditure) from 2018 to 2022. Kudrin has long been critical of the political and economic course pursued by the Russian government, arguing that heightened tensions with the West deprive Russia of much needed invested funds (Litvina 2017). Prior to the 2022 invasion of Ukraine, he called for reductions in defense spending, increased investments in health care and education, and the privatization of remaining state-owned industries (Lomskaya 2017). The Russian government has not followed his advice. Kudrin’s support of large-scale state investments in health care and education reveal him to be not a proponent of minimal government but rather someone who sees an important role for the state in creating the conditions for economic growth. This reflects a more general shift among Russian liberals, who have generally turned away from the radical free-market ideas of the early 1990s towards the precepts of a mixed market economy and social liberalism.

Until its dissolution in 2008, the main political force promoting classical economic liberalism was Union of Right Forces party, led by Nemtsov. Its 2001 manifesto stated that “Rights and individual freedoms have no sense or value where people lack the means to secure themselves and their families by honest labor and profitable enterprise on the basis of private property, where the institution of private property is not recognized and respected… The liberal response to this challenge consists in affirming property rights as sacred and inviolable” (Soiuz pravykh sil 2002, 484).

Few voices can now be heard speaking of property as sacred and inviolable. Political activist and former world chess champion Gary Kasparov (2015, 150) remarks that “those like me who favor free markets and an open, Western-leaning
society, learned to accept the need for the social and economic stability programs touted by the left.” Likewise, exiled oligarch Mikhail Khodorkovsky (2020, 57) comments that “Russia must say goodbye to the dream of ‘a small state.’” Khodorkovsky advocates for social supports to be maintained or raised (ibid., 62). “It’s utopian to imagine that one can come to power by democratic means by proposing a right wing, even extreme right, partly libertarian agenda, advertising the charms of a ‘small state’ and the potential of the ‘free market,’” he says (65).

This shift reflects an understanding that the negative experience of the 1990s has thoroughly damaged the popular appeal of free-market ideas. While Russia’s economic reformers did manage to create the basis of a market economy, for many Russians this achievement involved personal hardship from hyperinflation, skyrocketing crime rates, and the like. This hardship discredited both liberals and liberalism in the eyes of the great majority of the Russian population.

Worsening Russian-Western relations are another factor standing in the way of a liberal recovery. Russian liberalism’s strong association with Westernism means that as tensions with the West have increased, popular dislike of Russian liberals has risen too, particularly because of the tendency of some liberals to take the West’s side in its struggles with the Russian Federation. This became particularly noticeable following the Maidan Revolution in Ukraine in 2014 and the subsequent annexation of Crimea. For many liberals, the Maidan Revolution’s talk of making a ‘civilizational choice’ in favor of the West exactly reflected their own aspirations. They therefore supported it, condemned the annexation of Crimea, and opposed the anti-Maidan protests that erupted in Donbass. In liberals’ eyes, the problem was not much that the annexation of Crimea was a breach of Ukrainian sovereignty as that it led to a collapse in Russia’s relations with the West, destroying their dream of a Western future for Russia. As Yavlinsky put it, “The main consequence of the current policy towards Ukraine is the strengthening of Russia’s course as a non-European country” (quoted in Golovchenko 2018, 202).

This attitude put such liberals in sharp opposition to the bulk of the Russian people, nearly all of whom welcomed the annexation of Crimea. They have suffered the political consequences. As a member of Yabloko’s political committee, Anatoly Rodionov, told his colleagues during a party debate on the topic: “Russian society has said ‘No, Crimea is ours, and Yabloko is not ours.’ You understand, this is what has happened. We shouldn’t fool ourselves. We have crossed a red line separating society’s understanding…from society’s hostility. … I think there’s been a sort of ethical glitch. We’ve taken the enemy’s side” (Redchenko 2017, 218).
Concluding remarks

In Russia, those who have been considered ‘liberal’ have generally adhered to certain key tenets of classical liberal theory, above all a belief in the rule of law and expanded civil liberties. But they have varied considerably in their adherence to other key tenets, such as representative government, free markets, and reluctance about government provision of services. Consequently, classical liberals in the pure sense have been rare. When large-scale liberalization has occurred, such as the emancipation of the serfs in 1861 and Gorbachev’s perestroika in the 1980s, it has been initiated by the state for reasons that have had little or anything to do with classical liberal theory. The one exception is the economic reforms enacted by Russian liberals under Yeltsin in the early 1990s. This was probably the only period in Russian history when those governing the country professed classical liberalism across the spectrum. Even then, though, liberals’ commitment to democracy was rather weaker in practice than it was in theory. Furthermore, the perceived negative consequences of their policies have served to discredit them ever since. As Gulnaz Sharafutdinova wrote in 2020: “The widely shared belief that the Russia of the 1990s was a place of disorder, criminality, impoverishment, and a very weak, collapsing state, functions today as a cognitive frame that colors political imagination and shapes Russian citizens’ political judgment. This frame underpins societal fears of liberal and any other reforms and shapes popular preferences for stability and non-revolutionary political change” (Sharafutdinova 2020, 105–106).

Liberalism—from classical to social—has not fared well in Russia. By the time of writing, in the midst of Russia’s war in Ukraine, Putin’s government has largely driven liberals either into silence or into exile. The government has forced liberal civil society organizations, many of which have depended on foreign funding, to register as ‘foreign agents’ and made it increasingly hard for them to operate. The government has also forced liberal media outlets such as the Novaia Gazeta newspaper and the Ekho Moskvy radio station to shut down, and has arrested some critics of the war in Ukraine.

Some unexpected event may occur that once again breathes life into the liberal cause. As things stand, though, Russian liberalism is in a very poor condition, repressed by the state and despised by most of the Russian people. As one survey concludes, the words most associated with liberalism in Russian eyes are: “West, transition, chaos, oligarchs, foreign, unpatriotic, non-conformists, artists, and no respect for Russia’s values, traditions and history” (Simionov and Tiganasu 2018, 142).
References


Vestnik gosudarstvennogo i munitsip'al'nogo upravleniia 3: 39–49.


Gutorov, V. A. 2017. Rossiiskii liberalizm v politiko-kul'turnom izmerenii: opyt svravneni't'noogo teoreticheskogo i istoricheskogo analiza (chast' II) [Russian Liberalism in Political-Cultural Measurement: The Experience of Comparative Theoretical and Historical Analysis (Part II)]. Politeks (Izdatelstvo Sankt-Peterburgskogo Universiteta) 131(2): 4–42.


Leontovitsch, Victor. 2012. The History of Liberalism in Russia. Pittsburgh: University of
Pittsburgh Press.


**Litvina, Anna.** 2017. “Domestic Demand Will Not Push Our Economic Growth to 3–4%.” *Realnoe Vremya* (Kazan, Russia), July 5. Link

**Lomskaia, Tatiana.** 2017. Kudrin predlozhil Putinu zaniat'sia obrazovaniem i zdorov'iem. *Vedomosti* (Moscow), September 6. Link


Paul Robinson is a professor in the Graduate School of Public and International Affairs at the University of Ottawa. He is the author of *Russian Conservatism* (2019) and *Russian Liberalism* (2023), both published by Northern Illinois University Press, as well as other books and articles on Russian and Soviet history and current affairs. He also writes regularly for the international press. His email address is Paul.Robinson@uottawa.ca.

**About the Author**

Paul Robinson is a professor in the Graduate School of Public and International Affairs at the University of Ottawa. He is the author of *Russian Conservatism* (2019) and *Russian Liberalism* (2023), both published by Northern Illinois University Press, as well as other books and articles on Russian and Soviet history and current affairs. He also writes regularly for the international press. His email address is Paul.Robinson@uottawa.ca.
Classical Liberal Think Tanks in Greece, 1974–2024

Constantinos Saravakos¹, Georgios Archontas², and Chris Loukas

In the contemporary Greek context, the term liberalism predominantly retains its original political meaning, emphasizing key principles such as individual liberty, free markets, and limited governmental intervention. Thus, the use of liberalism in this paper denotes political outlooks with salient classical liberal, as opposed to modern social liberal, characteristics.

Similarly, the term conservative in this article is used in a sense more akin to European interpretations rather than the American context. In Europe, conservatism often implies adherence to existing structures, which frequently have corporatist and mercantilist traits. This can be in partial contrast to the American understanding of conservatism, which might refer to preserving the Madisonian liberalism embodied in the U.S. Constitution.

Liberalism appears to be one of the most prominent ideologies in Greece currently. According to a survey conducted by Kappa Research (2020) for the Friedrich Naumann Foundation, one out of every two Greeks holds a positive view of liberalism; however, this has not always been the case.

Although liberal ideas can be identified in Greece at least since the establish-

¹. Ph.D. candidate, University of Macedonia, Thessaloniki 546 36, Greece; Research Coordinator, Center for Liberal Studies (KEFIM), Athens 104 38, Greece. The authors would like to thank Professor George C. Bitros, Professor Aristides Hatzis, Tasos Avrantinis, and Nikos Charalampous for their useful comments and sources of information they provided.
². Adjunct professor, European Communication Institute, Maroussi 151 25, Greece; Head of Educational Programs, Center for Liberal Studies (KEFIM), Athens 104 38, Greece.
³. Nationwide survey in October 2020, with a sample of 1214 individuals. The standard error of the survey was ±3 percent. Question: “Does each of the following mean something positive or negative to you?” (link).
ment of the modern state in 1821, the development and implementation of liberal and democratic institutions faced significant challenges and disruptions, at least until 1974. For example, in a period of only 20 years, between 1929 and 1949, Greece experienced intense political conflict, oppressive dictatorships, Nazi occupation, and a devastating civil war (Hatzis 2012, 21–22). The restoration of democracy in 1974, following a seven-year military dictatorship, enabled Greece to align with western liberal democracies. This shift initiated the Third Hellenic Republic (1974–present), a time marked by significant progress and prosperity, though not without fragility. Given this backdrop, an intriguing question arises: How has liberal sentiment evolved and grown in Greece over the past 50 years?

This article examines the rise and development of liberal think tanks in Greece starting from 1974, marking the first time such institutions were established in the country. Again, by classical liberalism or liberalism in this context we refer to the ideas that champion the principles of liberal democracy, individual and civil rights, and economic freedom, particularly in the form of a free-market economy characterized by deregulation and liberalization. As indicated above, in Greece, unlike in North America, liberalism does not typically imply support for a welfare state, a regulatory state, a nanny state, or a nudge state, although these interpretations can occasionally be relevant.

Furthermore, it is crucial to acknowledge that until 1991, in the light of the ideological clash globally between the liberal democracies of the Western world and the collectivist or socialist regimes of the Eastern Bloc, some Greek liberal think tanks placed more emphasis on advocating for the institutions of liberal democracy, in opposition to “democratic socialism,” rather than focusing on free markets and individual liberty. This distinction is key to understanding the varied priorities of these think tanks during different periods of their existence.

Our analysis concentrates on think tanks as the main vehicle for disseminating and popularizing ideas, as well as influencing policy-making agendas. Thus, our treatment is confined both in that it says little about matters prior to 1974 and, after 1974, discusses political parties, thinkers, books, journals, etc., only in so far as they inform the think-tank story.

The mid-1970s saw the proliferation of political think tanks not only in Greece but across Western countries. This trend was driven by two main factors: first, the rise of the “new right” think tanks, which aimed to challenge the prevailing Keynesian economic models and establishment leftism (Arin 2014, 36–38, 69–74); second, the growing importance of the European Economic Community (today’s EU) and the subsequent ideological debates and competition concerning its future policies and direction (Boucher et al. 2004).

4. For more regarding the classical liberal components of the Greek War of Independence, see Hatzis 2021.
In our study, we catalog the liberal think tanks in Greece, identifying their key players, affiliations, partnerships, projects, and their wider impact on the political landscape. To better understand the evolution of Greek liberalism since 1974, we suggest dividing this timeline into three distinct periods: 1974–1991, 1991–2007, and 2007–2024. Each of these intervals represents significant phases in the development and influence of liberal thought and policy within the country, reflecting the changing dynamics of Greek politics and society over the past fifty years.

Some background

Following World War II, Greece endured a brutal civil war from 1946 to 1949. In the aftermath, the country was governed by a paternalistic democratic regime notably deficient in the rule of law (Hatzis 2012). From 1949 to 1974 the Greek Communist Party (Κομμουνιστικό Κόμμα Ελλάδας, or KKE) was outlawed. Although proper elections were held, the period from 1961 to 1964 was marked by political instability. In the 1961 elections, the then-prime center-right Prime minister, Constantinos Karamanlis, was accused by his political adversaries of electoral intimidation and fraud. These allegations precipitated a period known as the “Unyielding Struggle” (Ανένδοτος Αγώνας), a time of intense political conflict and turmoil. The main opposition party during this period was led by the center-left former Prime Minister Georgios Papandreou (Clogg 1988, 43–44). Although Papandreou was elected prime minister in 1963 and re-elected in 1964, a significant dispute with King Constantine II led to his resignation in 1965. The dispute centered around the Aspida case, a supposed conspiracy in which Papandreou’s son, Andreas Papandreou, was allegedly involved.

Following Papandreou’s resignation, Greece experienced a period of heightened political instability. From 1965 to 1967, the country saw the formation of four consecutive governments, none of which was the result of elections. This persistent instability and governmental turmoil set the stage for the military coup in April 1967. This coup led to a seven-year period of dictatorship, which eventually concluded in the summer of 1974.

During the 1950–1974 period, ideas and policies in favor of increased state intervention gained traction across the Greek political spectrum, in accordance with the broader global trend. According to George Bitros and Anastasios Karayannis (2015, 190–191), the government implemented policies focused on infrastructure projects such as road improvements to lower production costs in

---

5. “A reputed conspiracy of left-wing officers within the Greek armed forces” (Anschuetz 1965).
industrial and agricultural sectors and to boost employment. The policies were also aimed at attracting foreign investments and establishing regional and central development programs. Governmental decision-making was centralized in political leadership rather than bureaucratic entities.

Monetary policy during this time was geared toward stabilizing the currency’s value, fostering economic development through controlled interest rates, and regulating capital flow from the financial sector, which was subject to state control. Fiscal policy looked to generate budget surpluses (primarily by addressing tax evasion), to finance infrastructure investments, and to reduce production costs for the private sector.

Therefore, following 18 years of a democratic but relatively illiberal regime and a subsequent seven-year military dictatorship, along with an economic shift toward state intervention and control, Greece in 1974 had an opportunity to reconstruct its political institutions and realign with the principles of liberal democracy. This transition became known as Metapolitefsi, and its key figure was Constantinos Karamanlis. Karamanlis, as noted above, had previously served as prime minister in the 1960s, assumed the role again from July 1974 to May 1980, and later served as president of the Republic from 1980 to 1985 and again from 1990 to 1995. Upon his return to Greece in the summer of 1974 from his self-imposed exile in France, where he lived during the dictatorship, Karamanlis did not attempt to reinstate a political regime like that of the 1949–1967 period. Reviving his pre-1967 party, the National Radical Union (ERE, Greek: Εθνική Ριζοσπαστική Ενωσις), or a similar party model, did not align with Greeks’ strong desire for change after enduring seven years of dictatorship. His actions most symbolic of this change were the legalization of the Greek Communist Party, which had been banned in the preceding political climate, and ending the monarchy in Greece by discouraging the return of King Constantine II to the throne (Clogg 1988, 61).

New Democracy

During that critical period of transition, Karamanlis recognized the need for trustworthy and experienced partners. To this end, he established a new political party, New Democracy (Νέα Δημοκρατία, or ND). This party began as a coalition of well-known political figures from the pre-dictatorial center and right, alongside a group of younger politicians. This mix marked a shift from the pre-dictatorial right-wing political landscape. ND appeared to move away from the staunch anti-communism that characterized the pre-dictatorial right. The party adopted a more moderate approach in its dealings with political adversaries. This change in attitude was instrumental in opening the door to liberal elements within the party, both in
The gradual opening of ND to liberal ideas, initially under the leadership of Constantinos Karamanlis and later under Konstantinos Mitsotakis, presented a significant opportunity for Greek classical liberals to voice their opinions and exert influence on Greek politics. Especially under Mitsotakis, between 1984 and 1993, ND transitioned from a center-right party with moderate economic interventionist policies to a robust proponent of free-market economics and a reduced role for the government. This evolution of ND in embracing liberal principles occurred in three main phases.

The first phase began with the party’s pre-congress in Halkidiki in 1976. This event was notable not only for the participation of ND members but also for the involvement of scholars from think tanks. These scholars endeavored to steer the party’s orientation toward a market economy and to pivot ND’s ideology in a new, liberal direction. This initiative came after three years of ideological ambiguity within ND, as it had distanced itself from the old, pre–1967 system of ideas but had not yet fully embraced a new set of principles (Katsoudas 1991, 221). Two years later, ND held its first official congress. Karamanlis’s objective was to establish ND as a party of principles and at the same time to organize it in a manner akin to the Western center-right political parties (Chatzivasileiou 2010, 492).

The ideological reorientation of New Democracy involved the expression “Radical Liberalism” (Greek: Ριζοσπαστικός Φιλελευθερισμός), a term coined by Constantinos Karamanlis. It did not actually signify radicalism but aimed to qualify ND’s specific ideological mix as a middle ground between “traditional liberalism and democratic socialism” (Katsoudas 1991, 226). On the other hand, Manolis Alexakis (2001, 125) describes ND’s “Radical Liberalism” as “managerial empiricism.” He suggests that the term was essentially devoid of substantive content, permitting the government to approach situations based on pragmatic considerations without being constrained by ideological boundaries. The phrase “radical liberalism” gradually faded from the party’s rhetoric, particularly leading up to the 1981 elections, at which point the party suffered a defeat. By this time, the influence and appeal of liberal ideas within ND were seen as relatively weak.

The second phase in the liberal development of New Democracy began when Georgios Rallis succeeded Constantinos Karamanlis as prime minister. This phase continued with Evangelos Averoff’s tenure as the party president starting in 1981. During this period, New Democracy’s association with liberalism diminished, except for the brief period when Rallis was in leadership (Andrianopoulos 6.

6. However, ND seemed to belong to the so-called ‘cadre’ parties during 1974–1981, until its transformation to a more ‘mass’ party, from 1981 to 1984, according to Duverger’s typology (Duverger 1974).
The rise of Rallis to the position of prime minister in 1980 marked a significant shift in Greece’s economic policy. In his new role, Rallis explicitly stated that the effort for growth would be primarily driven by private initiative. This approach was encapsulated in the slogan “less state, more political and economic freedom.” According to Yannis Loulis (1981, 24) this was the first time that profit was recognized as a legitimate motive in the economic process in Greece and that failed Keynesian recipes should be abandoned. While the Keynesian mixed economic model remained at the core of Greece’s economic framework, there was a newfound interest in enhancing competitiveness, addressing structural issues in agriculture and industry, and developing the service sector (Chatzivasileiou 2010, 525). Nevertheless, following his defeat in the 1981 elections, Rallis resigned as ND president.

Rallis was succeeded by Evangelos Averoff, who led the party during its time in the opposition from 1981 to 1984.7 Averoff sought to reorient ND back to its traditional conservative roots, reminiscent of the Greek right’s stance prior to 1967. This shift involved diminishing the ideological influence that liberal think tanks previously held over the party.

To reinforce this conservative direction and counterbalance the influence of the Center for Political Research and Education (KPEE), which was a primary advocate for liberal ideology within ND, Averoff established the Institute of Social Research and Studies (IKEM). The creation of IKEM was a strategic move to garner theoretical and political support for Averoff’s more conservative viewpoints, thereby steering the party’s ideological trajectory toward a more traditional conservative path (Papavlasopoulos 2004, 262). Under his leadership, the focus within ND shifted significantly from ideological matters to a much-needed reorganization of the party.

The third and arguably most significant period started when Konstantinos Mitsotakis became the party’s president in 1984, and it continued until 1993 when he resigned following a loss in the general elections. During this period, from 1984 to 1993, ND endeavored to reinvent itself as a liberal party, aligning with the characteristics of the European center-right.

From 1985 onwards, especially leading up to 1990 when ND eventually gained power after several consecutive elections, the party’s agenda increasingly leaned toward market-economy principles. Mitsotakis stated the pillars of his

---

7. As Papavlasopoulos (2004, 263) notes, speaking of the Averoff period: “We could briefly say that, in this new ideological-political and tactical folding of the conservative faction, nationalism absorbs the radicalism of the first post-dictatorial seven years and liberalism is adulterated by strong doses of authoritarianism and right-wing populism.”
economic policy in the following words:

Privatization of the economy and strengthening of market rules. Liberalization of markets and healthy competition. Modernization of markets. (Mitsotakis 2013, 35)

From 1985 to 1989, the party’s program shifted from moderate statist positions to a stronger emphasis on free enterprise, private initiative, and a more deregulated and liberalized market economy (Loulis 1995, 198). This period is particularly significant as it represents the most substantial effort by classical liberal think tanks to influence policymaking through a political party in power.

The ideological shift within ND presented an excellent opportunity for liberal think tanks to have an impact; this is why we consider this short introduction to the Greek political landscape crucial to comprehending the development of Greek classical liberal think tanks. In fact, several think tanks were born and died with the main aim of making ND more liberal. Consequently, many of the think tanks in our study were in constant interaction with ND, given its status as the main party of the Greek center-right. However, it’s important to note that there was significant internal resistance within the party. This conservative opposition within ND posed a substantial challenge to the efforts of think tanks aiming to promote liberal policies.

Cataloging the classical liberal think tanks in Greece 1974–2022

In this section, we detail the principal features of the classical liberal think tanks that have been active in Greece since 1974. The information presented here is largely based on interviews conducted with key figures from each organization. These interviews were carried out in the summer of 2014 as part of a master’s thesis by the first author of this article. The thesis, titled The Liberal Think Tanks in Greece from 1974 to 2010,8 was completed at the University of Athens and is written in Greek. The current paper serves as an English-language summary and expansion of the original thesis, extending the analysis up to the beginning of 2024.

For the master’s thesis, six semi-structured interviews were conducted.

---

8. Saravakos (2014); the thesis committee consisted of Alexandros-Andreas Kyrtsis (professor of sociology, Department of Political Science and Public Administration, University of Athens), Aristides Hatzis (professor of philosophy of law and theory of institutions, Department of Philosophy and History of Science, University of Athens), and Yiannis Tsirbas (assistant professor of political science and methodology of social research, Department of Political Science and Public Administration).
Thematic analysis was used to identify and interpret patterns that emerged from the interviews (Braun and Clarke 2006, 86–95). The rest of the information was collected from websites and official documents. To classify a think tank as ‘liberal’ for the purposes of our study, a key criterion was that the organization explicitly stated in its mission statement or any other official document its commitment to promoting liberalism or liberal ideas within Greek society. In Table 1, we present nine liberal think tanks that have operated in Greece since 1974.9

<table>
<thead>
<tr>
<th>Full name in English</th>
<th>Full name in Greek</th>
<th>Abbr.</th>
<th>Years founded and deactivated/suspended/ceased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center for Political Research and Education</td>
<td>Κέντρο Πολιτικής Έρευνας και Επιμόρφωσης</td>
<td>KPEE / ΚΠΕΕ</td>
<td>1975–2001</td>
</tr>
<tr>
<td>Center of Political Studies</td>
<td>Κέντρο Πολιτικών Μελετών</td>
<td>KEMEP / ΚΕΜΕΠ</td>
<td>1975–1984</td>
</tr>
<tr>
<td>Movement for Multilateral Disarmament, Peace and Security</td>
<td>Κίνηση για τον Πολυμερή Αφοπλισμό, την Ειρήνη και την Ασφάλεια</td>
<td>ΚΙΡΑΕΑ / ΚΙΠΑΕΑ</td>
<td>1984–1988</td>
</tr>
<tr>
<td>Adam Smith Club / Liberal Forum</td>
<td>Λέσχη Adam Smith / Φιλελεύθερο Φόρουμ</td>
<td>Φόρουμ</td>
<td>1985–1993</td>
</tr>
<tr>
<td>Friedrich Naumann Foundation</td>
<td>Ίδρυμα Friedrich Naumann</td>
<td>FNF</td>
<td>1989–1993; 2012–present</td>
</tr>
<tr>
<td>The Center for Liberal Studies – Adamantios Korais</td>
<td>ΚΕΦίΜ - Αδαμάντιος Κοραής</td>
<td>ΚΕΦίΜ / ΚΕΦίΜ</td>
<td>2008–2010</td>
</tr>
<tr>
<td>Forum for Greece</td>
<td>Φόρουμ για την Ελλάδα</td>
<td>-</td>
<td>2010–2014</td>
</tr>
<tr>
<td>The Center for Liberal Studies – Markos Dramanis</td>
<td>Κέντρο Φιλελεύθερων Μελετών</td>
<td>ΚΕΦίΜ / ΚΕΦίΜ</td>
<td>2011–present</td>
</tr>
</tbody>
</table>

We now proceed to discuss each of the nine, in order of its founding date. We give an English version of the full name but then usually revert to an acronym from the Greek name.

9. The Friedrich Naumann Foundation was the only classical liberal think tank that was re-established after it first ceased.
The Center for Political Research and Education was established in December 1975 with the inspiration coming from George Rallis, who would later become prime minister of Greece from 1980 to 1981. Timoleon Louis, a notable classical liberal intellectual and member of ND, was the one who presented the idea of founding this kind of political institution to Rallis. Louis founded KPEE with the objective of steering ND toward liberalism, amid concerns that the party might follow a more conservative path. Initially, KPEE’s main goal was to support the necessary institutional reforms for Greece’s entry into the European Economic Community. Over time, it expanded its mission to actively promote liberalism in the public discourse. KPEE achieved this by organizing seminars and lectures on various subjects, including politics, economics, and international relations. Additionally, it published several books on similar themes, contributing to the intellectual discourse on liberalism in Greece.

Although initially KPEE tried to avoid aligning with a specific party, this proved challenging in practice since many of its members, including Louis himself, were active members of ND. In its statute, KPEE outlined the following main objectives:

1. Further dissemination of the basic values of a pluralistic, multiparty, democratic society.
2. Reinforcement of liberal values, private initiative, and the basic principles of a free market economy.
3. Continuous study and research on social, political, and economic issues.

Since the ND congress at Halkidiki in 1977, KPEE functioned as a policy-making partner for the party, attempting to guide ND toward a more distinctly liberal orientation, away from its initial ideological vagueness. KPEE also published a journal titled *Epikentra* (Greek: Επίκεντρα), which contributed notably to the intellectual discourse on liberal ideas.

KPEE’s funding came from a variety of sources. In its formative years, up until Evangelos Averoff assumed the leadership of ND, KPEE received financial support from the Konrad Adenauer Stiftung, the think-tank branch of Germany’s Christian Democratic Party. This support illustrates the international connections and support that KPEE leveraged in its efforts to influence policy and promote liberal ideals within Greece (Papandropoulos 2011, 66). In some collaborations for events held abroad, KPEE received support from various foreign think tanks,
usually its local partners.

KPEE expanded its activities and influence to encompass almost all the dimensions of think tanks.\(^\text{10}\) The ideological vagueness of ND and the traditional conservative roots of the Greek right spurred reactions from party members in the political center, who aimed to reshape the party’s orientation. KPEE’s influence became noticeable in 1979 at the party’s First Congress. However, the extent of KPEE’s impact on the proceedings and outcomes of that congress was not as significant as anticipated (Katsoudas 1991, 43–44).

Under the leadership of Yannis Loulis, KPEE played a pivotal role in orchestrating the ideological and strategic reorientation of ND, an effort that contributed to the party’s electoral gains in 1985, which were indicative of the systematic and effective work undertaken by KPEE. ND remained an opposition party, however. In 1985, ND published *Declaration of Liberty*, a manifesto that clearly defined the party’s ideological stance. This document emphasized a shift toward “less state, more market economy,” signifying ND’s commitment to liberal economic principles (Alexakis 2001, 109–110).

From 1987 to 1990, KPEE was under the leadership of Dimitris Katsoudas. During this time, KPEE continued to exert significant influence on the ideological orientation of ND. Notably, Katsoudas played a crucial role in shaping the party’s ideological stance. He was primarily responsible for the formulation of ND’s key ideological texts from 1985 to 1991 (Katsoudas 1991, 43–44). The publication of the review *Epikentra*, initiated at the very start of KPEE’s operations, stood out as the most notable activity of the think tank. Initially, its circulation was limited to select circles, maintaining a level of confidentiality. However, over time, it evolved to reach a broader audience with a wide array of substantial articles, ranging from ideological critiques spanning various perspectives to in-depth political and electoral analyses.

For several years, starting in 1984, KPEE organized training seminars targeted at young Greeks (Katsoudas 1991, 43–44). The primary objective of these seminars was to establish a foundation of liberal principles, moving away from entrenched old-party mentalities. To achieve this, KPEE conducted an extensive series of educational events: 170 one-day seminars, 18 two-day seminars, and 12 conferences and meetings (ibid., 42). From then on, KPEE embarked on what can be described as a marathon of seminars, a continuous effort that lasted until 1991.

Another focus of action was the organization of conferences and lectures to promote liberal ideas in the public discourse. KPEE organized an array of events

---

\(^{10}\) The main tasks of KPEE, according to the Boucher et al. (2004, 4) taxonomy, were: to popularize classical liberal ideas, to offer expertise to political parties, to influence policymaking, and to produce academic research.
addressing a variety of topics. These included seven events focused on the economy, twelve on ideology, seven on politics, five on international relations, two on education, and one each dedicated to the environment, the arts, and local government. Additionally, there were five lectures that centered around guest speakers and honored individuals (Katsoudas 1991, 48–50). Publishing books was another significant facet of KPEE’s activities, although it demanded more time and resources. In total, eight books were published in this period, including four books on liberal ideology, two on economics, one on politics, and one on education (ibid., 51).

The influence of KPEE began to wane in 1991, coinciding with a significant shift in the circumstances of its executive members. Many of these key individuals, who constituted the liberal core within ND, were appointed to administrative roles within the ND government, which was in power from 1990 to 1993. This transition to governmental positions, paradoxically, reduced their capacity to exert the same level of ideological influence from within the party.

The decline of KPEE’s influence became more pronounced after 1994, especially under the conservative leadership of Miltiades Evert in ND. In this new phase of the party’s leadership, the ideological priorities shifted, leading to a significant reduction in KPEE’s impact on public opinion. Eventually, KPEE’s presence and influence in the political and public spheres faded away.

Center of Political Studies (KEMEP, Greek: Κέντρο Πολιτικών Μελετών)

KEMEP, founded by Andreas Andrianopoulos in 1975, emerged just a few months after the collapse of the military dictatorship in Greece. This think tank was established based on the models of influential think tanks in the United Kingdom and the United States, aiming to replicate their success and impact. The primary funding for KEMEP came from businessmen who shared Andrianopoulos’ perspective. KEMEP focused on promoting liberalism through two primary avenues. First, it organized seminars for the youth branch of New Democracy, ONNED, and, more significantly, its university branch, DAP-NDFK (Greek: ΔΑΠ-ΝΔΦΚ). Secondly, it extended its outreach beyond youth to general audiences by organizing lectures and events. This broader approach was complemented by the production of research reports and op-ed articles.

Between 1981 and 1985, KEMEP experienced a period of relative inactivity. It eventually ceased operations in 1986 due to lack of funding after Andrianopoulos took office as mayor of Piraeus. The establishment of another think tank in 1984, the Movement for Multilateral Disarmament, Peace, and Security (KIPAEA), also played a role in diminishing KEMEP’s prominence. KIPAEA’s emergence
provided an alternative platform for the discussion and promotion of similar ideas, making KEMEP’s role in the ideological landscape less essential.

**Movement for Multilateral Disarmament, Peace, and Security (KIPAEA, Greek: Κίνηση για τον Πολυμερή Αφοπλισμό, την Ειρήνη και την Ασφάλεια)**

The Movement for Multilateral Disarmament, Peace, and Security (KIPAEA) was established in 1984, again by Andreas Andrianopoulos. Its primary objective was to promote anti-war sentiments from a liberal perspective. KIPAEA attracted several prominent ND members, including Kostas Karamanlis, the nephew of Konstantinos Karamanlis. Kostas Karamanlis took over the leadership of KIPAEA in 1986 after Andrianopoulos became mayor of Piraeus. Notably, Kostas Karamanlis would later serve as president of ND from 1997 to 2009 and as the prime minister of Greece from 2004 to 2009.

In 1988, Andrianopoulos published *This Is Liberalism*, a book that served as an ideological manifesto for the liberal movement. KIPAEA’s funding primarily came from ND. This support was partly driven by the context of the period, as KIPAEA was established in response to the disarmament calls from the political left. While there was a consensus on disarmament involving major powers like the U.S.S.R. and the U.S.A., KIPAEA distinguished itself by aligning with Western liberal democracies, in contrast to the socialist orientation favored by the left. This positioning helped KIPAEA gain traction and support within ND and the broader liberal community.

KIPAEA’s most active period was from 1984 to 1988. It advocated peaceful initiatives and liberal ideas especially relevant in the context of the Cold War. It emphasized the classical liberal belief that freedom of choice is crucial for ensuring peace, security, and a high level of prosperity in a democratic society.

KIPAEA concentrated on organizing educational seminars across Greece, often in collaboration with the KPEE. The decline of KIPAEA began around 1986, coinciding with Andrianopoulos’ taking office as the mayor of Piraeus. The dissolution of the Soviet Union and the perceived retreat of leftism at the end of the Cold War also contributed to the reduction in KIPAEA’s relevance and activities.

**Adam Smith Club / Liberal Forum (Greek: Λέσχη Adam Smith / Φιλελεύθερο Φόρουμ)**

The Adam Smith Club was established in 1985, initially as a loosely organized group of individuals who shared a common interest in liberalism. George Bitros, professor of economics at the Athens University of Economics, was instrumental
in its formation, connecting with various university student circles passionate about liberal ideas. The club’s founding members included Bitros, Anastasios (Tasos) Avrantinis, Panagiotis (Panos) Evangelopoulos, Kostas Christidis, Athanasios Papanthropoulos, Stavros Petroka, Panagiotis (Takis) Michas, Iason Zafoila, and Evangelos Psigas, with Bitros serving as president, Avrantinis as secretary, and Evangelopoulos as Treasurer.

The Adam Smith Club operated on democratic principles, allowing all members to propose ideas and engage in discussions about the club’s activities. It officially registered as a nonprofit organization in 1989, with ambitions to evolve into a think tank that would influence New Democracy (ND). However, the club faced significant financial constraints, operating with minimal funding, and relying mostly on internal support without public or private aid. Despite its financial limitations, the Adam Smith Club occasionally cooperated with ND, mainly due to the absence of other political parties more closely aligned with its liberal stance. Nevertheless, the club did not take on the role of providing expertise to ND.

The Adam Smith Club focused its efforts on influencing people, particularly the younger generation, toward liberalism by connecting them with prominent intellectuals associated with the club. However, many of the targeted young people were put off by the club’s self-identification as “neoliberal.” Despite not having a formal affiliation with ND, many members of the Adam Smith Club maintained close contact with the party.

The club primarily advocated for economic liberalism and the principles of a free market, emphasizing these aspects over political and social issues. For example, when it addressed topics such as drug legalization, it approached them primarily from an economic perspective.

To disseminate and promote its ideas more broadly, the club turned to publishing articles in popular media outlets, aiming to garner public interest. It initially published a newspaper named *The Neoliberal*, with Michas serving as the manager and Papanthropoulos, Psigas, Evangelopoulos, and Theodoros Benakis acting as editors. However, after releasing only three issues, the newspaper underwent a transformation into a bi-monthly review titled *Free Choice*.

Between 1987 and 1988, the Adam Smith Club began to experience a slowdown in its activities, leading to discussions about a change in its orientation. Club members recognized their potential for individually exerting influence within ND, especially since club members had many personal connections with politicians in the party. The club particularly targeted those close to Prime Minister Konstantinos Mitsotakis, the father of later Prime Minister Kyriakos Mitsotakis (2019–). As a result of this strategic shift, a new entity called the Liberal Forum was established, aiming to provide expert services to ND. The first secretary-general of the Liberal Forum was Thodoris Vamvakaris, with Aristomenes Syngros and
A second direction was to take modern ideas on economics and economic policy from the economically successful countries. We tried to show the size of ideological change on this topic that we were able to achieve in the previous years. The result of the work of the team in the third direction was the “Liberal Contract.” … It was worked upon for months and we had arranged for discussions to take place, in order to decide upon it. A fourth direction was to offer a subscription service to any members of the parliament who wanted consultation. Last but not least, it was the excellent work that the members of our team did in order to introduce the ideas of the Liberal Forum to ND. (Saravakos 2014, 51)

The Liberal Contract that Bitros speaks about was a book by the scientific council of the Liberal Forum that articulated the ideological foundations of the Forum. Drawing inspiration from thinkers like Robert Nozick and Jan Narveson, the book’s approach aligned closely with what is typically known as libertarianism, emphasizing individual liberty and minimal state intervention.

The dissolution of the Liberal Forum came in the wake of the third ND Congress in 1993. A defining moment at this congress was Miltiades Evert’s statement that “New Democracy is not a neoliberal party.” This declaration marked a clear ideological stance for ND, one that did not align with the views of the Liberal Forum, which was arguably the most radical classical liberal think tank that has operated in the Greek political scene.

Society for Social and Economic Research (EKOME, Greek: Εταιρεία Κοινωνικών και Οικονομικών Επιστημών)

EKOME, established in 1986 as a nonprofit organization, was founded by Sotiris Papasotiriou, an investment consultant and economist. The core team of EKOME included Sotiris Papasotiriou and, from 1991 onwards, his son Charalampos Papasotiriou. They were joined by several collaborators such as Athanasios Diamantopoulos, Theodoros Palaskas, Marietta Gianakou,11 and Athanasios Papandropoulos. EKOME was mainly funded by Papasotiriou himself and, to a small extent, by funded research.

EKOME was created in response to what its members viewed as a regression

in economic policy, particularly during the tenure of PASOK (the Panhellenic Socialist Movement), led by Prime Minister Andreas Papandreou. Notably, Papasotiriou was influenced by the economic reforms of Margaret Thatcher in Great Britain and the role that British think tanks played in facilitating these changes.\textsuperscript{12}

At the time of EKOME’s inception, the New Democratic Party (ND), despite having strong liberal elements, had a predominantly statist core. Conservative Miltiadis Evert, who later became the party’s president, was known as the leader of the anti-liberal group.

EKOME’s initial activities focused on translating key works of liberal thought, including those by Frederic Bastiat. Additionally, the organization developed several policy proposals, which were disseminated to various ministries, politicians, banks, and business leaders, aiming to influence economic and political discourse in Greece. In 1992, it published the book \textit{Liberal Social Contract}. A summary of its main thesis follows:

The fundamental problem of the Greek economy is underproduction, the lack of international competitiveness and stagnation. The causes are:

1. The counter-productiveness of the oversized public sector.
2. The crowding out of private enterprise by the public sector.
3. The interventionism of the state. (Saravakos 2014, 54)

EKOME published research on international relationships, education, economics, classical liberal ideas, and politics. Despite its efforts and contributions, EKOME ceased operations in 2001. This closure reflected waning interest in liberal ideas within Greek society at the time.

\textbf{Friedrich Naumann Foundation (FNF)}

The Friedrich Naumann Foundation, a German think tank, established a branch in Greece in 1986. This move was part of a tradition among German political parties to associate with organizations in Germany which then set up affiliated foundations in foreign countries, particularly those undergoing periods of political transition. Greece, having emerged from a military dictatorship in 1974, presented a prime example of such a nation in transition.

FNF in Greece aimed to consolidate the country’s liberal voices, recognizing the existing affiliations of political parties with other German think tanks: ND members were generally aligned with the Konrad Adenauer Stiftung (associated

\textsuperscript{12} The Institute of Economic Affairs was the leading example of how think tanks influence policies and spread the ideas of classical liberalism.
with Christian Democratic ideologies), while the socialist party tended to be closer to the Friedrich Ebert Stiftung.

FNF’s strategy involved engaging with the youth and building connections with existing Greek liberal think tanks such as KPEE. By providing these organizations with resources, FNF facilitated the co-organization of seminars and public discussions, fostering a collaborative environment among liberal entities.

Additionally, FNF distinguished itself from other think tanks by showing a keen interest in engaging with centrist political parties. It sought to address social and political issues with the same emphasis as economic ones. It published several books in Greek such as *Liberalism in Greece: Liberal Theory and Practice in Politics and Society in Greece* (Katsoudas 1991), and *Conservatism, Liberalism, Socialism: The Basic Political Movements and Their Answer to the Important Problems of the Age* (Meinardus 1992).

In 1993, FNF ceased its operations in Greece as part of a broader plan to reallocate its resources and efforts to other countries, particularly those that were deemed to have a greater need for liberal support at the time. As said before, FNF came to Greece in 1986 to help promote much-needed reforms at the time. By 1993, with the collapse of the former Soviet regimes, FNF leaders felt that Greece was fully integrated into the Western world and the foundation’s objectives in the country had largely been achieved.

In 2012, amidst the severe economic crisis, FNF returned to Greece, accompanied by other German organizations, to promote and advance liberal ideas and reforms particularly relevant in the context of the economic challenges Greece was facing. Since its reestablishment in Greece, FNF has concentrated its efforts on supporting the youth liberal movement. It has organized numerous seminars and public events, aiming to engage and educate the younger generation on liberal principles and policies.

Currently leading the FNF’s operations in Greece and Cyprus is Alona Tatarova, the first woman to head a Greek liberal think tank. Under her leadership, FNF (link) continues to champion pluralist democracy, the rule of law, and an open market economy. The foundation is dedicated to preserving and promoting freedom of action, with a specific focus on strengthening European values in Greece.

**Center for Liberal Studies (KEFiM, Greek: Κέντρο Φιλελεύθερων Μελετών)**

The Center for Liberal Studies (KEFiM) currently stands as the most prominent and active classical liberal think tank in Greece. The first two authors of this article have had a long-standing association with KEFiM, reflecting their deep
involvement and commitment to the organization’s mission and activities.

KEFiM was established in October 2011 by Gregory Vallianatos (President), Stratis Katakos (Vice-President) and Nikos Charalampous (General Secretary) as the official think tank of the Liberal Alliance (in Greek: Φιλελεύθερη Συμμαχία), a political party founded in April 2007. In its initial stages, KEFiM operated as a voluntary organization composed of individuals committed to classical liberal values. Its primary aim was to advocate for political and economic freedom in the Greek public sphere. Between 2011 and 2016, KEFiM engaged in various activities to promote these ideals. It published reports and translated influential books like Johan Norberg’s *The Klein Doctrine: The Rise of Disaster Polemics*, and *Why Liberty?* and *Peace, Love, Liberty*, both edited by Tom Palmer. Additionally, KEFiM organized public events on issues such as individual rights and economic freedom. Despite these efforts, KEFiM’s impact during this early stage was limited.

In 2017, KEFiM distanced itself from the Liberal Alliance and restructured as an independent, nonpartisan think tank with a focus on advocating classical liberal ideas in the political discourse and policy making. Under the leadership of Alexander Skouras as President and Nicos Rompapas as Executive Director, KEFiM gained greater prominence following its rebranding. This was further enhanced by its international connections with institutions like the Atlas Network and the Smith Family Foundation.

KEFiM promotes classical liberal ideas through various initiatives. Annually, it hosts a liberal academy focusing on the theory and practice of liberal concepts for young people. It also organizes the Greek Economics Olympiad, a competition for high school students designed to combat economic illiteracy. In 2021, KEFiM, with the support of the John Templeton Foundation, presented an educational program, *Greece 2021: Bicentennial of the Liberal Revolution*, which highlighted the influence of classical liberalism on the founding of modern Greece during the 1821 War of Independence.

In terms of policy impact, KEFiM published an initiative in 2019 titled *Greece 2021: Agenda for Freedom and Prosperity* (link). This publication proposed a set of reform policies deemed essential for steering the country toward prosperity and freedom following a decade of economic stagnation. The *Agenda* aimed “to serve

---

13. It should be noted that in 2008 KEFiM—Adamantios Korais was established as the think tank of the political party Liberal Alliance. Its three founders were Manolis Manoledakis, Stratis Katakos, and Dimitris Skalkos. The same people a couple of months before had also created their own party, the Liberal Alliance. The Center for Liberal Studies, or KEFiM—Markos Dragoumis (it was named for Markos Dragoumis, a prominent classical liberal intellectual, who was one of the most significant Greek champions of liberty of the 20th century), could be considered the successor of the Liberal Alliance. However, it was established in 2011 and is a separate legal entity.

14. For more regarding the impact of classical liberal ideas on the Greek Revolution, see Hatzis 2021.
as a roadmap for the implementation of those necessary reforms that will enhance prosperity and freedom in every area of public life in our country.”

KEFiM participates in international dialogue through the Atlas Network (U.S.A.) and the European Liberal Forum (Belgium). It also partners on various projects with mission-aligned organizations such as the Friedrich Naumann Foundation (Germany), Timbro (Sweden), Fraser Institute (Canada), Cato Institute (U.S.A), the Tax Foundation (U.S.A), the Foundation for Economic Education (U.S.A), the EPICENTER network (Belgium), and the Institute of Economic Affairs (UK).

Forum for Greece (Greek: Φόρουμ για την Ελλάδα)

Forum for Greece was a short-lived think tank created in 2010 by Dora Bakoyannis, who had been an ND MP and minister. The primary purpose of this think tank was to disseminate liberal ideas and to support the promotion of Bakoyannis’s new political party, the Democratic Alliance (in Greek: Δημοκρατική Συμμαχία).

The Forum began with 53 members; notable among them in the academic field were Giorgos Arsenos, Theodoros Diasakos, Manos Karagiannis, Olga Kolokitha, Vasiliki Lalagianni, Christos Balfoutis, Alexandra Tragaki, Kostas Ifantis, and Asteris Houliaras. While Bakoyannis initially served as the president of the Forum for Greece, she later passed the leadership role to Andreas Andrianopoulos, allowing her to concentrate on the Democratic Alliance. Following this transition, Dimitris Katsoudas, who was formerly the director of KPEE, assumed the position of director for the Forum for Greece after 2013.

The think tank was active for two years, but its momentum waned after Bakoyannis rejoined New Democracy in 2012, leading to a shortage of funding. Forum for Greece did collaborate with the FNF and the European Liberal Forum for several events in Greece, Brussels, and Egypt. Since 2014 it has been inactive.

Other short-lived think tanks

Several other liberal think tanks, mostly short-lived, emerged during the period from 1994 to 2007. The Center for the Development of Ideas for Greece in the 21st Century (Greek: Κέντρο Ανάπτυξης Ιδεών για την Ελλάδα του 21ου Αιώνα,

15. Bakoyannis was former mayor of Athens (2003–2006), former minister of foreign affairs of Greece (2006–2009), and MP with the center-right party New Democracy. She was expelled by the party in 2010 and re-joined it in 2012. Bakoyannis has remained an MP with New Democracy until the time of writing.
1994–1999), known as E21, was founded by Stephanos Manos. E21 organized a series of significant conferences covering various topics, including the economy, the pension system, and education. Theodoros Skylakakis, who has been serving as Greece’s minister of environment and energy since July 2023, was the director of E21 and its accompanying magazine. The editorial committee of this publication included Vassilis Kavvalos, Dimitris Bourantonis, Andreas Sideris, Miranda Herbertstein, and Michalis Psalidopoulos. Additionally, notable figures like Tryphon Kolintzas and Alexandra Nikolopoulou were actively involved in E21’s initiatives.

Citizen’s Defense (Greek: Άμυνα του Πολίτη) was a think tank established in 1997 by Stelios Argyros, former president of the Hellenic Federation of Enterprises (SEV) and then an ND member of the European Parliament, along with Kostis Hatzidakis (who has been the minister for national economy and finance since July 2023), Despina Patsavos, Michaela Kyriakopoulou, Tasos Avrantinis, and others. Stelios Argyros was the president of Citizen’s Defence until its dissolution in 2001, with Avrantinis serving as vice president. The think tank’s primary goal was to address violations of citizens’ individual property rights and issues related to bureaucracy.

Freedom Network (Greek: Δίκτυο Ελευθερίας) was founded in Athens in May 2002 and actively conducted regular meetings for approximately three years, primarily at the Zofilia Hotel on Alexandra Avenue. These meetings, generously facilitated by Iason Zafolias, were complemented by open events featuring speakers such as George Alogoskoufis, Kyriakos Mitsotakis, and Dora Bakoyannis. The founders of Freedom Network included a group of notable individuals: Kostas Christidis, who served as the president, Thanasis Papandropoulos as vice president, Tasos Avrantinis as general secretary, and members such as George Bitros, Dimitris Dimitrakos, Spyros Hatiras, Stavros Petrolekas, Katerina Prapopoulou, Dionysis Katranitzas, Nikos Dimou, Dimitris Skalkos, Lefteris Panagiotidis, Vangelis Psygas, and Telemachos Maratos. The primary aim of Freedom Network was to promote and disseminate the principles of political, social, cultural, and economic freedom in every lawful way. In its pursuit of these goals, the network also ventured into publishing. One notable publication was a small book by Murray

---

17. Professor of economics at the Athens University of Economics and Business since 1990 and member of the Hellenic Parliament from September 1996 till October 2009. He also served as Greece’s minister of economy and finance from March 2004 till January 2009.
19. Greek philosopher, currently professor emeritus of political philosophy in the Philosophy of Science Department of the University of Athens.
Rothbard titled *What Has Government Done to Our Money?* which was released by Elati Publications and featured a foreword by Iasonas Zafolias.

Several other groups and organizations emerged during the early years of the Greek crisis, specifically from 2011 to 2014. These included Aristides Hatzi’s John Stuart Mill research group, the Open Society Group (in Greek: Όμιλος Ανοιχτή Κοινωνία) led by Dimitris Dimitrakos; Students for Liberty under the leadership of Nikos Kostopoulos; Epekeina Hora (in Greek: Επέκεινα Χώρα) led by Nikos Yannis; Women for Liberty led by Yuli Foka Kavalieraki; and Greek Liberties Monitor led by Michael Iakovides. These entities played a role in organizing lectures, presentations, and events, often in collaboration with foreign think tanks.

There is rather a scarcity of information regarding the activities and impact of these think tanks. The available information is primarily limited to documents establishing their legal entities. Apart from these documents, there is a lack of other sources such as books or conference papers that could document their influence and significance. These groups and organizations were marked by an informal mode of operation, without the formal structure of an institution with paid staff. Their activities over the years were sporadic, primarily fueled by initiatives from a loosely connected collective of individuals.

The era of flourishing ideas 1974–1993: A new hope

After outlining the key liberal think tanks that have been active over the past 50 years, we now divide these five decades into three distinct periods, whimsically but aptly named after the original three Star Wars films.

The majority of the liberal think tanks were established and operated in the first two decades following the restoration of democracy in Greece in 1974. This development can be attributed to two primary factors: Firstly, for the first time, a major Greek political party—the New Democracy party, or ND—showed a willingness to experiment with new ideas and promote liberal ideas and policies in the country. Secondly, the global political climate was increasingly moving away from planned economies and rediscovering the value of free markets. The political agendas of leaders like Margaret Thatcher in the United Kingdom and Ronald Reagan in the United States significantly altered the international political landscape. ND, in its quest for political transformation, took cues from these global trends, aligning itself with the shift toward market-oriented policies. However, this shift toward liberal and market-friendly policies was not without its challenges.

In the mid-1970s, when Greece was at the early stages of democratic
restoration, the prevailing economic model in Western economies was largely Keynesian, characterized by elitism and corporatism. During the 1970s, Greece’s economy operated on a mixed model, heavily involving state intervention. Notably, from 1978 to 1991, about 80 percent of the banking sector in Greece was under the control of three state-owned banks (Halkos and Salamouris 2002, 8). This era was sometimes referred to as ‘socialmania’ (in Greek: σοσιαλμανία), a term that highlights the extensive state control over the economy, evident in the policies of both New Democracy’s governments (1974–1981) and PASOK’s governments (1981–1989).²⁰ It wasn’t until Konstantinos Mitsotakis became the leader of New Democracy in 1984 that a major Greek political party, then in opposition, openly advocated for a free-market economy. This shift was pivotal in attracting the majority of liberal think tanks in Greece toward an affiliation with ND, as they sought to move the party in a free-market orientation. In 1989, out of the four liberal think tanks that were active in Greece, three had ties to ND: KPEE was directly linked, EKOME and the Liberal Forum were indirectly associated, while FNF maintained an open dialogue with ND and smaller centrist parties.

The 1990 election victory of ND marked a turning point for the first wave of liberal think tanks in Greece. With this change in government, many staff members from KPEE were appointed to positions in the public administration and the cabinet. However, the influence of the remaining liberal think tanks was somewhat constrained, as their access was limited to certain cabinet members who did not represent the majority. This was partly because ND, despite some liberal leanings, remained a moderately conservative party, often cautious about free-market economic policies that could potentially diminish the state’s control over resources. Aris Trantidis (2014, 230) highlights that the base of ND was resistant to policies that might significantly reduce the size of the state, fearing that such measures would limit political intervention and patronage. Despite this, the ND government from 1990 to 1993 did employ some strategies aligned with free-market economics. Notably, there were cuts in real wages and pensions in the public sector, which contributed to decreasing the budget deficit, as reported by the OECD (1993, 80).

Theodore Pelagidis (1997, 79) critically views the government’s approach as a “wild monetarist” program, which he argues was not well-suited for Greece’s industrial base. While the government achieved a reduction in both the budget deficit and the inflation rate, its policy of maintaining a hard currency was believed

---
²⁰. The two oil crises of these periods and the subsequent economic recessions deteriorated the financial positions of private and public enterprises. As a result, in 1983, the Papandreou government introduced the Industrial Reconstruction Organization S.A. (IRO), a state organization that aimed to take control of the failed enterprises, trying to restore the financial viability of those companies via public funding (Halkos and Salamouris 2002, 14).
to have negatively affected production and failed to encourage investment.

In terms of deregulation and liberalization, the 1990–1993 government took several steps, including privatizing profitable firms, liquidating non-viable ones, abolishing rent control, removing government interference in wage bargaining in the private sector, and liberalizing the financial sector and shopping hours (Trantidis 2014).

However, these free-market-oriented reforms were accompanied by an increase in indirect taxes. Pelagidis (1997, 76) notes that total tax revenue as a percentage of GDP rose from 23.8 percent in 1988 to 27.2 percent in 1992, largely with the goal of reducing the substantial budget deficit inherited from previous administrations. Despite these efforts, the policies implemented by the ND government were often considered insufficient in addressing the broader structural challenges facing the Greek economy (OECD 1993, 82).

The challenges of implementing classical liberal policies in Greece during the early 1990s were illustrated by the government’s decision to discreetly hire a few hundred ND members into the public sector just before the 1993 elections. This was an effort to secure political support, an effort that reflected the enduring tradition of clientelism in Greek politics (Trantidis 2014, 229). Andrianopoulos has stated that neoliberal principles such as reducing the public sector and lowering taxation were not effectively implemented in Greece during that period (ibid., 222).

Greek liberal think tanks did have an influence on the economic policies of the governing ND, but this influence was moderated by several significant obstacles. The first major obstacle was that the base of ND largely favored interventionist policies, with only a few senior leaders within the party being advocates of free-market principles. The second obstacle was the government of Konstantinos Mitsotakis, which came to power in 1984 with a narrow majority of just two seats in parliament and experienced internal opposition, with members within the party openly criticizing the government’s free-market policies (Trantidis 2014, 227). The third obstacle was that the Greek public in a large part felt they had benefited from the extensive government policies implemented during the previous PASOK administrations (1981–1989). Consequently, the new policies, perceived as ‘austerity measures,’ were not widely embraced.

In September 1993, the already slim parliamentary majority held by ND collapsed following the defection of two of its MPs. This loss of majority came at a

---

21. It is estimated that Mitsotakis’ government closed or privatized 85 to 111 enterprises during 1990–1993 (Trantidis 2014, 231).
22. The critique focused on the government’s “obsession” to reduce the deficit and its “hard drachma” policy (Trantidis 2014, 227).
23. The government’s policies limited the opportunities for clientelist exchange by the party’s prominent members, even though Mitsotakis was associated with patronage politics in his homeland, Crete.
time when Greece was experiencing a decline in industrial output, high unemployment rates, and generally poor economic performance outcomes that were attributed to monetary policies implemented by Konstantinos Mitsotakis’ government.

These economic challenges contributed to the Panhellenic Socialist Movement (PASOK) gaining power in the October 1993 elections. Subsequently, Miltiades Evert, a conservative member of ND, assumed leadership of the party. The change in leadership from Mitsotakis to Evert marked a significant shift in the party’s ideological direction, resulting in a diminished influence by liberal think tanks on ND’s policies and ideas.

This transition also led to the departure of prominent free-market advocates within the party, such as Andrianopoulos and Manos, who left ND in the following years. The party, now under more conservative leadership, struggled to regain its footing in the political arena, failing to secure a parliamentary majority in subsequent elections for 11 consecutive years.

**The anti-liberal empire strikes back: 1993–2007**

The appointment of much of the KPEE staff to public administration roles following New Democracy’s (ND) election to government in 1990 had a significant impact on the think tank. These appointments were made to assist the government in implementing its policies, but they left KPEE with a shortage of capable executives to continue its operations. Additionally, after 1993, funding for KPEE was substantially reduced, further hindering its activities.

ND, shifting its focus, was no longer inclined to allocate resources solely for the promotion of economic liberal ideas. This change was compounded by the withdrawal of international allies like FNF from Greece to focus on other regions in greater need of liberal efforts.

During this period, when classical liberal ideas were in retreat, Manos made a significant effort to revive economic liberalism in Greece. He founded a political party called the Liberals (Greek: οι Φιλελεύθεροι) in 1999. However, the party had difficulty gaining substantial political traction. In the 2000 elections, in an alliance with ND, the Liberals managed to elect only two MPs. By 2001, the party had ceased to exist.

The Third Way wave during the mid-1990s, exemplified by the political agendas of Tony Blair in the United Kingdom and Bill Clinton in the United States, had a notable influence on the political landscape in Greece. PASOK underwent an

---

24. According to World Bank indicators, Greek GDP slightly increased in 1992 by 0.7 percent and declined by 1.6 percent in 1993 (link).
ideological shift during this period, moving away from its earlier more frank leftism (Spourdalakis and Tassis 2006, 506). Under the leadership of Prime Minister Costas Simitis starting in 1996, PASOK demonstrated a shift toward more liberal economic policies. One of Simitis’ primary objectives was to ensure that Greece met the criteria necessary for entry into the European Monetary Union (EMU). His efforts were successful, and in 2001 Greece became a member of the Eurozone. Both Andrianopoulos and Manos (free-market-oriented ministers under the 1990–1993 Mitsotakis administration) were offered positions on PASOK’s electoral lists for the 2004 elections by the then-party leader George Papandreou. However, the term ‘liberalism’ was notably absent from PASOK’s official rhetoric.

As for ND, following Kostas Karamanlis’ election as party leader in 1997, succeeding the more conservative Evert, the party showed increased openness to moderate liberal economic policies. Under Karamanlis’ leadership, ND positioned itself as representing the ‘middle ground,’ adopting a centrist strategy to appeal to a broader electorate. This approach was based on the recognition that Greek voters tended to be pragmatic, generally lacking strong ideological biases, and exhibiting limited trust in the state (Pappas and Dinas 2006, 484–485). In line with this understanding, ND’s political program was intentionally de-ideologized. While the party refrained from explicitly aligning itself with liberalism or other specific ideological currents, it frequently used the slogan “the great liberal party.” This rhetorical strategy was indicative of ND’s efforts to maintain ideological flexibility to appeal to a wide range of voters.

The proliferation of small, short-lived liberal initiatives during this time reflects a broader trend: a lack of substantial public demand for classical liberal ideas, which in turn influenced the policies and orientations of political parties.

The re-emergence of classical liberalism
2007–2017: Return of the Jedi

Liberal Alliance, a small party established in 2007 on a platform rooted in both political and economic liberal principles, sought to occupy the political space perceived to exist between the center-left PASOK and the center-right ND.

In its founding declaration, the Liberal Alliance articulated a commitment to fundamental liberal tenets, recognizing life, liberty, private property, and the pursuit of happiness as its core principles. However, its electoral appeal has been nonexistent. Its highest electoral result was approximately 1.8 percent of the vote in

25. Liberal Alliance, Founding Declaration (link).
May 2012, achieved in collaboration with Drasi, a political party founded by Manos. This was considerably below the 3 percent threshold required for parliamentary representation in Greece. The Liberal Alliance’s lack of electoral success could partially be attributed to the two-party system. Despite ideological alignment with liberal principles to some extent, the electorate’s fragmented support and the predominance of major parties like New Democracy and PASOK hindered the Alliance’s traction. The electoral system, particularly the enhanced proportional representation law, further exacerbated this challenge. With a threshold of 3 percent of valid ballots required for representation, smaller parties struggle to secure seats. Moreover, the system’s provision of bonus seats disproportionately benefits leading parties, incentivizing voters to align with established entities rather than smaller liberal factions. The combination of voter behavior, party dominance, and electoral rules underscores the formidable barriers that confronted the Liberal Alliance in achieving significant electoral gains. And the limited demand for liberal policies, coupled with dominant state interventionism, impedes the electoral success of any liberal party in Greece.

However, the Liberal Alliance’s contribution to the Greek political landscape extended beyond electoral politics. It was instrumental in founding KEFiM–Adamantios Korais (later renamed KEFiM–Markos Dragoumis), the first classical liberal think tank in Greece since 1989. KEFiM aimed to create a platform for liberals from various political backgrounds to engage and collaborate, transcending party affiliations (link). In November 2008, KEFiM–Adamantios Korais became a member of the European Liberal Forum.

In 2009, two years after the establishment of Liberal Alliance, the political party named Drasi (Greek: Δράση) was founded in Greece by the experienced politician, former MP, and former minister Stephanos Manos. He called for a broad coalition of progressive, center-left, center-right, social democrat, liberal, modernizer, and pragmatist forces (link). However, like the Liberal Alliance, Drasi struggled to achieve significant electoral success and did not manage to elect any Members of Parliament or Members of the European Parliament. In 2019, Drasi suspended its operations and expressed support for ND.

Meanwhile, Dora Bakoyannis, the former MP and ND minister, established the Democratic Alliance party and its affiliated think tank, Forum for Greece, in 2010. This led to some high-level members of the Liberal Alliance joining the Democratic Alliance and staffing the Forum for Greece. However, the Democratic Alliance was short-lived and ceased to exist in 2012 after Bakoyannis rejoined ND.

The post-2010 period in Greece, characterized by the sovereign debt crisis and subsequent bailout programs, created a crucial need for reform-oriented ideas. However, the highly polarized political environment of the time did not favor the emergence of liberal political parties as a moderate force between the two dominant...
parties. Instead, the political landscape saw the entry of radical left and anti-liberal right parties into the Greek parliament.

However, amidst this backdrop, various liberal initiatives like Liberal Alliance, Drasi, Democratic Alliance, KEFiM, and Forum for Greece, despite their fragmented nature, succeeded in promoting classical liberal ideas to specific audiences. The groundwork laid by these organizations contributed to the rebranding and emergence of KEFiM–Markos Dragoumis in early 2017 as the most active classical liberal, non-partisan, and independent think tank in Greece.

Currently, KEFiM’s executive and academic board is composed of members who were part of previous liberal efforts, including Panagiotis Evangelopoulos, associate professor of economics (Adam Smith Club, Liberal Forum); Dimitrios Katsoudas, former Secretary General for European Affairs (KPEE, Forum for Greece); Tasos Avrantinis, lawyer (Adam Smith Club, Liberal Forum, Drasi); Nikos Charalampous, architect (Liberal Alliance, KEFiM–Adamantios Korais); Miranda Xafa, former member of the board of the IMF (Drasi); Harry Papasotiriou, professor of international relations (EKOME).

Since 2012, KEFiM and the FNF have been the only two active classical liberal think tanks in Greece. Figure 1 highlights the timeline of liberal think tanks in Greece from 1974 through 2023: the initial proliferation of these institutions in the first period after 1974, their decline starting from 1993, and their resurgence from 2007 onwards.

**Figure 1.** Liberal think tanks in Greece, 1974–2023

---

26. We mention only the ones that were documented as professionally organized or operating on a regular basis. Therefore, we omitted the short-lived think tanks with limited resources that we mentioned above.
Concluding remarks: The situation of classical liberalism in Greece

Since 2015, classical liberalism in Greece has experienced a remarkable increase in popularity. Survey data illustrate this trend: In 2009, only about 15 percent of the population self-identified as liberals. This figure slightly decreased to 12 percent in 2012 (Public Issue 2012). By 2016, the proportion of the population identifying as liberals rose to 18 percent. This upward trend continued, with surveys in 2019 and 2023 indicating that approximately 19–20 percent of the population self-identified as liberal (diaNeosis 2022; Eteron/Aboutpeople 2023).

The increasing percentage of the Greek population identifying as liberals, as indicated by recent surveys, does not necessarily equate to a comprehensive or accurate understanding, or even active support of specific liberal ideas and policies.

Figure 2. Self-identified ideological description of people in Greece, 2009–2023

Questions:
Indicate your ideological self-placement (2009 and 2012)
Which of the following ideological terms would you say fits you most (2015–2023)

27. Nationwide survey in November 2009, with a sample of 1331 individuals. The standard error of the survey was ±3.2 percent. Question: “Ideological self-placement” (link).
29. Nationwide survey in April 2022, with a sample of 1255 individuals. The standard error of the survey was ±2.8 percent. Question: “Which of the following ideological terms would you say fits you most?” (link).
Characteristically, Greek governments have faced challenges in effectively implementing liberal reforms, even though there have been some efforts, such as tax cuts before the COVID-19 pandemic. The overall tax burden remains significant. As measured by Tax Freedom Day, which represents the point in the year when citizens have theoretically earned enough income to cover their tax obligations, the burden is substantial. In 2022, Tax Freedom Day was expected to reach 181 days in 2022, only 2 days lower than in 2021 (Saravakos and Moutsatsou 2022). If state borrowing due to budget deficits is also accounted for, Tax Freedom Day extends to the 200th day of the year.

Moreover, in terms of economic freedom, Greece ranks relatively low on a global scale, occupying the 85th place among 165 economies all over the world in the 2022 Economic Freedom of the World Index (Gwartney et al. 2022). Notably, Greece ranks last among EU countries and is positioned at 153rd in the “Size of Government” component of the index.

Greece has shown signs of backsliding in terms of the institutional checks and balances that are fundamental to liberal democracy. There has been a trend toward a hyper-concentration of powers in the executive branch, as noted by Evie Papada et al. (2023).

Educating the public, clarifying misconceptions, and promoting a deeper understanding of what classical liberalism entails are major tasks facing today’s liberal think tanks in Greece. They are actively working, often collaboratively, to promote classical liberal ideas and influence policymaking in the country. Their projects and initiatives cover a range of areas that are central to classical liberalism. KEFiM focuses on issues such as economic literacy, regulatory quality, tax reduction, and the expansion of the liberal youth movement, while FNF’s projects in Greece concentrate on empowering the liberal youth movement, on civil and individual rights, and on economic freedom.

Together, KEFiM and FNF–Greece play a significant role in the Greek political and social landscape, striving to broaden the understanding and implementation of classical liberal ideas and policies in Greece in a time of ongoing economic and political challenges.

References


Andrianopoulos, Andreas. 1988. Αυτός είναι ο Φιλελευθερισμός [This is Liberalism]. Athens:


Kappa Research. 2020. Ο φιλελευθερισμός στην Ελλάδα, σήμερα [Liberalism in Greece, Today]. October. Friedrich Naumann Foundation (Athens). Link


About the Authors

Constantinos Saravakos is a Ph.D. candidate in International and European Studies at the University of Macedonia and the Research Coordinator at the Center for Liberal Studies (KEFIM). He holds an M.A. in Political Science and Sociology (Hons) and an M.Sc. in Applied Economics and Administration from Panteion University. He also holds a B.Sc. in Philosophy and History of Science from the University of Athens. He is a V-Dem Institute (University of Gothenburg) expert for Greece and Cyprus and a Fellow in the European Policy Information Center (EPICENTER). His email address is ksaravakos@uom.edu.gr.

Georgios Archontas is adjunct professor at the European Communication Institute and head of Educational Programs at the Center for Liberal Studies (KEFiM). He holds a Ph.D. in political philosophy and economic theory from Panteion University, Athens. He is a graduate of the Department of Communication and Mass Media of Panteion and of the postgraduate program in Political Science and Sociology, University of Athens. During the period 2016–2017, he was a visiting researcher at the Department of Political Science and International Relations of the University of Peloponnese. From 2016 until today, he has taught at the University of Peloponnese (Department of Political Science and International Relations) and the National and Kapodistrian University of Athens (Department of History and Philosophy of Science). He was the coordinator of the comprehensive reform proposal Greece 2021: Agenda for Freedom and Prosperity. His email address is georgios.archontas@kefim.org.

Chris Loukas is a high school student. He has been a member of the Greek team in the International Economics Olympiad for 2022 and 2023, winning a bronze and a silver medal in each year respectively. His email address is chrisloukas648@gmail.com.
Both Trygve Hoff and Ragnar Frisch were born in 1895, both were Norwegians, and both were economists. The exchange of letters published here is remarkable for several reasons. Both men were forceful personalities with clear, uncompromising views.

Hoff graduated in economics from the University of Oslo in 1916, spent a few years abroad and became an economic commentator for Norwegian newspapers in 1920. In 1935 he bought the business magazine *Farmand* (which in Norwegian means ‘traveling salesman’). Hoff was a classical liberal, and in 1939, before a packed assembly hall at the University of Oslo, he defended a doctoral dissertation (Hoff 1938) on economic calculation under socialism, restating and refining the arguments of Ludwig von Mises (1922) and Friedrich A. Hayek (1935) against the feasibility of central economic planning. He was a firm opponent of totalitarianism, both national socialism and communism. His magazine was banned during the Nazi occupation of Norway.

For a while Hoff was himself imprisoned. In gaol, he worked on a book he was writing on classical liberalism (Hoff 1945). After the war he became somewhat of a voice in the wilderness in Norway, albeit with strong support in the business community. Hoff was a founding member of the Mont Pelerin Society in 1947, and in his magazine he published many articles by other members, including Mises, Hayek, and Milton Friedman, and tirelessly criticized the restrictions on economic
freedom implemented by the dominant Labour Party. It was quite an event when in January 1952 he debated Labour Leader Einar Gerhardsen at the Students’ Association in Oslo, using arguments that he used in the November–December 1941 correspondence with Frisch.\(^2\) Hoff died in 1982.

Frisch completed his doctorate on mathematical statistics at the University of Oslo in 1926 and taught for some years in the United States, but returned in 1931 to Norway where he was appointed Professor of Economics and Statistics at the University of Oslo. A year later, he become the Director of the Institute of Economics at the University, originally financed by the Rockefeller Foundation. Like Hoff, during the War he was for a while imprisoned by the Nazis. A staunch socialist, Frisch became influential in Norway in the next few decades, filling the posts at his Institute with likeminded people and training many economists who became high officials, academics, or politicians, mostly in the Labour Party. His closest associate at the Institute, Leif Johansen, was a member of the Norwegian Communist Party, which was unservingly loyal to the Soviet Union. Together, Frisch and Johansen formed what became known as the ‘Oslo School’ (Eriksen et al. 2007; Eid 2007). Over the years, Frisch seemed to become more dogmatic in his views. Nonetheless, he maintained an influential presence in academia, government, and public discourse, and he was, with Dutch economist Jan Tinbergen, a first recipient of the Nobel Prize in Economics in 1969. Frisch had by then abandoned the Labour Party and become a vocal supporter of the Socialist People’s Party which had been formed in the early 1960s in opposition to Norway’s pro-Western foreign policy. Frisch died in 1973.

The two main questions discussed by Hoff and Frisch were: Was economic planning likely to be better in the long run than the free play of market forces? And might it threaten individual freedom?

Frisch did not necessarily support Soviet-style central economic planning. In the correspondence with Hoff he was vague about what he wanted himself except that government, aided by experts (like himself), had to correct what he found to be obvious market failures. Nevertheless, Frisch was impressed by the Soviet model. Later, he saw no reason to change his view. In 1961, for example, he wrote: “The blinkers will fall once and for all at the end of the 1960s (perhaps before). At this time the Soviets will have surpassed the US in industrial production. But then it will be too late for the West to see the truth” (Sæther et al. 2014, 63). He was far off the mark. In 1990, the last year of the Soviet Union, her GDP per capita was estimated to be $6,894, whereas in the United States it was $23,201 the same year (link).

It is true that for decades figures from the Soviet Union showed rapid economic growth. But those figures were questionable, not only because they

---

2. The debate was widely reported, for example in *Nordisk Tidende* 1952.
might have been falsified, but also because they did not always reflect reality: for example, capital accumulation in the munitions industry did not improve the living standards of ordinary people; neither did abandoned factories, even if their construction meant a nominal addition to GDP. In his letters to Frisch, Hoff gave several examples of a third reason: the large amount of defective goods in a country where production was not governed by consumer choice. Moreover, rapid economic growth can be achieved without any economic planning. The United States was by no means a perfectly free society, but rapid economic growth took place there without terror or famines as in the Soviet Union. In the United States in 1876, GDP per capita was estimated to be $2,570, increasing in the next twelve years, till 1888, to $3,282, or by $712. In 1928, GDP per capita in the Soviet Union was estimated to be $1,370, increasing in the next twelve years, till 1940, to $2,144, or by $774, about the same as in the comparable period in American history (link).

Frisch’s theories were however tested not in the Soviet Union but in Norway after the War where the Social Democrats, largely under his influence, pursued much more restrictive policies than their counterparts in the other Scandinavian countries. They reluctantly had to abolish some of their detailed foreign trade regulations when Norway accepted Marshall aid and joined the Organisation of European Economic Cooperation, OEEC (later the Organisation of Economic Cooperation and Development, OECD), and the General Agreement on Tariffs and Trade, GATT. This, according to Arild Sæther and Ib Eriksen (2014, 59), “saved the country from the worst excesses of a regime bent on economic planning.”

Nevertheless, extensive economic controls were retained much longer than in other Western countries. The Price Directorate under Wilhelm Thagaard had wide-ranging powers to direct consumption and investment into channels deemed to be in the public interest. In 1953, the Labour government, under fierce criticism from the opposition parties and the business community, abandoned plans to reinforce economic controls. When Norway joined the European Free Trade Association, EFTA, in 1960, she had to go further in reducing foreign trade regulations, but the planners made up for it by continuing strictly to regulate money and credit. Instead of rationing commodities, they rationed credit by keeping interest rates artificially low and directing savings to favored sectors of the economy.

The record of the Oslo School is however far from impressive. The relevant period of reference must be from 1950 until the mid-1970s, before the Norwegian oil boom. In 1950–1960, average annual economic growth in Norway was 2.6 percent, somewhat lower than the OECD average of 3.3 percent, but similar to that of the other Scandinavian countries. In 1960–1973, economic growth in Norway was 3.7 percent, whereas the OECD average was 4.0 percent. It was less than that of Finland (4.3 percent), on par with that of Iceland (3.7 percent), but slightly more
than that of Denmark (3.5 percent) and Sweden (3.1 percent). This cannot be seen as anything spectacular, despite claims to the contrary (Gerhardsen 1972, 155). Moreover, from 1950 to 1975 investment ratios in Norway were much higher than in most other OECD countries. In 1950–1959 the ratio was on average almost 32 percent, compared to 17 percent in Denmark and 21 percent in Sweden. In 1960–1969, the ratio in Norway was on average 29 percent, compared to 21 percent in Denmark and 23 percent in Sweden. “Year after year Norwegians sacrificed better living (consumption) to pay extra for only average growth rates” (Sæther and Eriksen 2014, 67). It was no coincidence that in the 1980s Norwegian politicians, officials, and economists largely discarded Frisch’s ideas. In the long run, Hoff had been proven right on the economic issue.

The second important question debated by Hoff and Frisch in 1941 was whether a concentration of power in the hands of planners might threaten individual freedom, or in Hayek’s words (1944), whether socialism was “the road to serfdom.” Frisch insisted that he was in favor of freedom of thought, because without it science would just wither away. But he was “unable to accept that this ideal of freedom necessarily presupposes a right of the individual to attend to all his business affairs at will.” In other words, he wanted freedom for scientists to discover and innovate, not for businessmen whom he dismissed as “the unenlightened plutocracy” (Sæther et al. 2014, 49).

Hoff, in parallel, makes arguments about the likely mentality of would-be government planners: that those who sought power most energetically were the people least likely to restrain themselves in using it; that if they were convinced they knew better than others what was in the public interest, they would be reluctant to yield to others; and that if they would not tolerate individual choice in the economy, they most likely would not tolerate individual choice elsewhere. Frisch conceded that there was something in these arguments but effectively chose to ignore them.

Hoff presented the ‘road to serfdom’ thesis three years before Hayek published his famous book. In Italy, the economist Luigi Einaudi (1957/1931) had also made some of the same points in a notable exchange with the philosopher Benedetto Croce (1957/1927), and in Sweden, Eli F. Heckscher (1934) and Gustav Cassel (1934) had warned in no uncertain terms that central planning might lead to tyranny. All of these 20th-century diagnoses and warnings may be seen as the continuation of a great many earlier warnings about the perils of centralization.

Frisch’s challenge to Hoff was revealing: “Would two persons in a boat not find it practical to row in the same rhythm, even if one is a Muslim and the other one a Catholic?” Frisch gives this analogy to suggest that government intervention is analogous to people finding a common rhythm in rowing. But the crucial difference is whether the rhythm is brought about by voluntary mutual adjustments of economic agents or by conforming to commands from above. Hoff’s main
point in his letters to Frisch was that the free market is a forum for separate and independent boating, where each person finds his own rhythm, whereas central planning requires rowing to the beat of the same drummer.

However, can the postwar record of Norway and the other Scandinavian countries not be invoked, as by Jeffery Sachs (2006), against the ‘road to serfdom’ thesis? These countries still seem to be democracies that have maintained civil liberties. First, it should be mentioned that these were not paradises without serpents. In the second place, appearances can deceive. Preferences may be changed not by direct coercion but by raising the costs of defying consensus. Wills can be bent rather than broken. While censorship was abolished in Norway a long time ago, self-censorship may often have been practiced in a country ruled by the same political party for decades. Indeed, a well-known Norwegian historian once labelled the post-war Norwegian state a one-party state (Seip 1963). “I have always believed that this sort of servitude, regulated, mild and peaceful,” Alexis de Tocqueville observed, “could be combined better than we imagine with some of the external forms of liberty” (Tocqueville 2010/1840, IV:1252).

Thirdly, Hoff presented a warning, not a prediction, and it was a warning directed against the idea of central economic planning rather than against the network of petty economic restrictions which was developed in postwar Norway. To some extent, Social Democrats in the Nordic countries heeded the warnings presented by Hoff, Hayek, and other economic liberals. The economies of the Nordic countries remained relatively free. Fourthly, the Nordic conservative-liberal tradition of liberty under the law was strong enough not only to withstand the assault of kings during the Middle Ages, but also of socialists in the twentieth century. The difference was that the kings based their claim to absolute power on the grace of God and the socialists on the grace of the People. The three crucial factors enabling the Nordic nations to maintain their liberties were the rule of law, an open economy, and social cohesion.

About the text: The four letters translated here appear in Trygve J. B. Hoff, *Tanker og Ideer* (Oslo: Aschehoug, 1975), pp. 49–65; they are the only letters between the two men in that volume and, to my knowledge, that passed between them. The letters are provided in full. All of the footnotes are mine. The two minor in-text insertions by me are in square brackets, giving original-language names of two periodicals referred to.

---

3. For example, after the war the Norwegians passed retroactive laws against collaborators despite a clear constitutional stipulation that no law could be given retroactive effect. They sent their most famous novelist to a lunatic asylum, although he was obviously sane as he subsequently demonstrated with a book about this experience (Hamsun 1949). They treated Norwegian girlfriends of German soldiers with great cruelty (Aarnes 2018). During the Cold War, the ruling Labour Party used the secret service against its enemies, mostly communists (Lund 1996).
References


Hoff, Trygve J. B. 1938. Økonomisk kalkulasjon i sosialistiske samfunn. Oslo: Aschehoug.


Hannes H. Gissurarson, born in 1953, holds a B.A. in philosophy and history and a cand. mag. [M.A.] in history from the University of Iceland, and a D.Phil. in politics from the University of Oxford. The author of more than fifteen books on politics, history and current affairs, in English, Swedish, and Icelandic, he is professor emeritus at the University of Iceland. He has been a Visiting Scholar at the Hoover Institution, Stanford University, at UCLA, at LUISS in Rome, and at several other universities. In 1998–2004 he served on the board of the Mont Pelerin Society and in 2001–2009 on the board of overseers of the Central Bank of Iceland. He is director of academic studies at an Icelandic think tank, RNH (link), Rannsóknarsetur um nyskopun og hagvoxt (Research Centre on Innovation and Economic Growth). His email address is hannesgi@hi.is.
1. Hoff to Frisch

Oslo, 4 November 1941.

Professor Ragnar Frisch
Ris, Vinderen.

Dear Frisch,

You said some time ago that your goal was ‘a planned economy and a free culture.’ For indisputable reasons you have a lot of influence, and since you might have an impact on the people determining the future of our county, I would like to make a few comments.

I also have a free culture as a main goal. But I am convinced that we will not achieve this under a planned economy. We have to choose.

I indeed recognize that you ride a big and powerful wave when you wish for a planned economy. Planning is not only the objective of the totalitarian states, but businessmen in democratic societies also push for it at full speed.

Partly, businessmen want to eliminate a troublesome factor, competition (even if it means their disappearance as independent businessmen). And partly, they are influenced by the war economies and totalitarian propaganda. Your proposal of a planned economy will therefore not meet much resistance from them.

The last time we spoke you expressed some qualifications about the extent of planning. This is a crucial point. When you nevertheless insist on a planned economy, I must assume that you are not only proposing anti-trust laws and traffic controls, but some more extensive planning, an economy directed according to a plan.

1. The four letters in the original Norwegian are published in Trygve J. B. Hoff, Tanker og Ideer (Oslo: Aschehoug, 1975), 49–65.
2. University of Iceland, 101 Reykjavik, Iceland.
You know my own point of view so I shall be quite brief. It is based on the psychological fact that human beings have difficulties in being tolerant on one sphere (such as culture) when they have at the same time to be intolerant and issue orders on another sphere (such as the economy).

Moreover, I am in general skeptical about the concentration of power, as psychologically power seems to create the need for ever more power. And planning the economy in our modern complex world requires concentrated power.

Finally, the people doing the planning will find it easy to identify their own interests as the public interest and to regard any criticism of them and of planning as socially harmful. Then it gets going. First, they will try and control the press and then logically turn to the arts and sciences. I would not defend the assertion that there is only a difference of degree between currency controls and concentration camps, but I would not deny, either, that there is quite a lot of truth in it.

This is not mere speculation. The tendency to control culture once the economy has been brought under control is as discernible in our modern planners as it was in the ancient despots. We had a typical example in 1935 when Ole Colbjørnsen, Member of Parliament, in The Workers’ Paper [Arbeiderbladet] threatened your colleague Professor Wilhelm Keilhau with dismissal for having dared to object to central planning.3

In the case that this is not sufficient to you, and because I regard this as being of the utmost importance, I am going to cite some people whom you would accept as serious commentators.

Professor Bernard Lavergne writes:

Central planning leads directly to political dictatorship in the same way as it presupposes economic dictatorship. Because of its internal momentum the regime will not survive unless it can acquire both political and administrative authority. This has been the case in all totalitarian states and also in the countries which have just embarked on this course. Every planned economy tends to be more and more complex and commandeering. To implement the plan inevitably the state has to be dictatorial. … The political and the economic dictatorship require and condition each other. Have the supporters of democracy and central planning thought of this necessary connection?4

---

3. Ole Colbjørnsen (1897–1973) was a Norwegian economist, journalist, and vocal advocate of central economic planning in Norway. Wilhelm Keilhau (1888–1954) was a Norwegian historian and economist, Professor of Economics at the University of Oslo and, like Hoff, a strong opponent of central economic planning.

I suppose that you are familiar with the conditions in totalitarian states—we are becoming acquainted with them in our own country as well—but just to make sure I cite a report on the organization of the German economy: 'The national socialist movement identifies with the state and seeks in the public interest totally to control society in all its sectors and to the utmost extent.'

The most comprehensive documentation of the interference by totalitarian states in scientific and cultural affairs is to be found in Annals of the American Academy. It covers both the situation in Germany and the Soviet Union and is quite valuable.

'It is not a mere coincidence that the cult of a directed civilization should be accompanied by a general foreboding that modern civilization is doomed,' Walter Lippmann writes in The Good Society. Frank H. Knight comments that 'a regime of collectivism will inevitably be a dictatorship, a tyranny over the whole cultural and personal life of man as well as over economic activity, however defined.' And: 'It is reasonable to suppose that the actual human beings in charge of the system would have to adopt this course whether they wanted to or not; and will anyone argue that those who would gravitate into positions of such power would be persons who would abhor the exercise of power beyond the carefully considered necessities of the situation?'

Knight also writes: 'It seems to me certain: (a) that the governing personnel in a socialistic state would be in a position to perpetuate themselves in power if they wished to do so; (b) that they would be compelled to assume permanence of tenure and freedom from the necessity of seeking frequent re-election as a condition of administering the economic life of a modern nation, even if they did not wish to do so; and (c) that they would wish to do so—that we cannot reasonably imagine

5. Svensk Utrikeshandel, 2, 1941.
8. Frank H. Knight, Lippmann’s The Good Society, Journal of Political Economy, 46:6 (1938), 865. In his letter, Hoff quoted from memory some of the passages in Knight’s 1938 articles. Here the references are to the original.
political power on the scale involved falling into the hands of persons of whom this would not be true.\textsuperscript{10}

In another passage, Knight writes that ‘the probability of the people in power being individuals who would dislike the possession and exercise of power is on a level with the probability that an extremely tender-hearted person would get the job of whipping-master on a slave plantation.’\textsuperscript{11}

Walter Sulzbach writes in an article on ‘Tolerance and the Economic System’:

The very essence of the conception of planning is that a design can be adopted to which the people will thereafter conform. That is equivalent to saying that a democratic people cannot have a planned economy, and that so far as they desire a planned economy they must suspend responsible government. A socialist government can solve this difficulty only by forbidding free discussion and by insisting on the acceptance of the belief in the superior wisdom of the government.\textsuperscript{12}

Professor J. M. Keynes wrote in 1933 an article called ‘National Self-Sufficiency’ where he recommended autarchy, in other words a soft version of planning. Nonetheless Keynes realizes that a possible consequence might be intolerance and suppression of criticism. Keynes says that if this would be the result, he would start again preaching the ideals of the 19th century.\textsuperscript{13} But how would he have the opportunity if the access to criticism has been blocked? You can find more comments in Professor Louis Rougier’s book, \textit{Les Mystiques Economiques}, and in articles by Professors Röpke and Rüstow in \textit{Die Friedens-Warte}.\textsuperscript{14}

Generally speaking, it is hopeless to persuade people to change their minds. The reason why I make the effort to write to you—besides the fact that I consider this matter to be of utmost importance—is that I know you are ready to be convinced by valid arguments and solid facts. Both do exist in this case.

\textsuperscript{10} Frank H. Knight, Socialism: The Nature of the Problem, \textit{Ethics}, 5:3 (1940), 264.
\textsuperscript{12} Walter Sulzbach, Tolerance and the Economic System, \textit{Ethics}, 50:3 (1940), 295. The first two sentences are actually a quotation from Walter Lippmann’s book, \textit{The Good Society}. Sulzbach (1889–1969) was a German sociologist and banker who moved to the United States in the 1930s and taught at Claremont College.
The conclusion is in my opinion inevitable: If you want to regulate the economy, you will soon have to regulate the world of ideas. This prepares the way to the totalitarian state. If you want to fight for the freedom of expression, you also have to fight for a free economy. This is the same front.

Greetings,

Trygve J. B. Hoff.

PS. As mentioned, I take the liberty to send copies of this letter to a few common acquaintances.

PPS. And finally, a reminder: ‘Political liberty can survive only within an effectively competitive economic system,’ Professor Henry C. Simons writes.15

2. Frisch to Hoff

Oslo, 10 November 1941.

Director Trygve Hoff
Oslo

Dear Hoff,

‘Economic regulation and freedom of expression’

Thank you for your letter from 4 November referring to our discussions some time ago.

In the interest of clarity, I will also permit myself to express my opinions in a letter. I support—just like you—as much freedom and individuality as possible in the world of ideas. The experience of centuries has for example shown us that without it science will just wither away. Science cannot survive without continually vigorous and independent critical thought and new questions. Therefore, I am willing to fight for my neighbor’s right to say what he says, even when he says he is disagreeing with me. The same applies to the other highest forms of human flourishing. Altruism and self-sacrifice are for example only possible under freedom.

I am however unable to accept that this ideal of freedom necessarily presupposes a right of the individual to attend to all his business affairs at will. Would two persons in a boat not find it practical to row in the same rhythm, even if one is a Muslim and the other one a Catholic?

You quote a lot of authors who hold that freedom of thought cannot be separated from economic freedom. There is of course something to what those authors mention, but permit me to say that these 'proofs' do not in general have any impact on me. Because I remember comparable proofs in a different cause. In the years before the present war many books were written—including books with many quotations from other authors—with 'proofs' that the Russian type of central planning was inefficient. Whatever one may think about spiritual life under the Soviet system—on which I shall not comment in this connection—there were those who 'proved' that this system was not only technically and economically inferior to the competitive system, but that it was also such a failure that it would sooner or later fall like a house of cards. I can for example mention that in a doctoral dissertation a while ago at our university it was seriously asserted that the utilization of capital goods under central planning would be so chaotic that molybdenum might be used to make toy swords instead of ordinary swords. A war for five months has demonstrated how much those 'proofs' are worth. The Soviet system has both technically and institutionally shown a strength which has surprised the world. There is reason to believe that despite everything the molybdenum found its way to the real swords, and in this connection, it is not far off to compare Russia's technical and institutional performance now with her performance in the previous war.

This is not at all said in order to glorify the Soviet system. There is much good and bad to be said about it. This is just said in order to characterize a certain kind of 'proofs' and to explain why I must be skeptical about this although I try hard to keep an open mind.

Personally I believe that we are entering a period where more developed forms of economic regulation will force themselves upon us. They are inevitable—and I believe appropriate—reactions against the defects which had emerged. The monstrous distortions of the Great Depression in the 1930s: the deliberate destruction of goods, the persistent unemployment and machines standing still in a world of distress, they could in my opinion to a large extent be traced to some 'individualistic' features of our economic system.

Now this has to be corrected, but it does not imply that the period of regulation has to last forever. Perhaps it will only last until in a natural process businessmen become sufficiently enlightened and public spirited that a return to economic liberalism 'on a higher level' will take place.

I am not dogmatic and I reserve the right to judge each of the new regulatory

---

17. Frisch is of course referring to the war on the Eastern Front which began when Hitler invaded Russia on 22 June 1941.
measures from a practical point of view. I do not desire regulation for its own sake. I would prefer that a sufficient coordination would be achieved without interference, but I would not either shy away from regulation where it turns out to be necessary. In principle, I would say that regulations to the greatest extent possible should be in the form of ‘rules of the game,’ in other words general prescriptions about price calculations, methods of exchange, use of bank accounts and so on, and not in the form of ‘guardianship,’ in other words not authorities which interfere with the daily operations of enterprises. When regulations are in the form of ‘rules of the game,’ there is least danger that they will threaten freedom of thought. This is like improving the machine instead of watching over the machinist. I believe that there are many opportunities for such improvements. But it presupposes quite much greater skill in managing the social and economic machinery than our businessmen and politicians have had. Therefore I believe that the time ahead will demand much of political economy, and especially of what we are now calling econometrics and national accounting.

With regards, Ragnar Frisch.

PS: I send you a few copies of this letter if you want to forward it to some of those to whom you sent copies of your own letter. I have myself sent a few copies to some common acquaintances. R. F.

3. Hoff to Frisch

Oslo, 22 November 1941.

H/JR

Personal

Professor Ragnar Frisch
Oslo

Dear Frisch:

On the Subject of the Discussion.

You have to be so kind as to explain what you mean by economic regulation if we are to get any result out of the main question (to what extent freedom of thought is compatible with economic regulation).

You say that you would like ‘general prescriptions about price calculations, methods of exchange, use of bank accounts and so on.’ What do those prescriptions mean? And what does ‘so on’ mean? Words are of little use if their meaning
is not clear. In Germany, private enterprise is supposed to be maintained whereas a Swedish observer (sympathetic) writes this year that ‘the government directives often seem in practice to be commands.’

If you only mean by ‘rules of the game’ decisions that one should for example determine prices according to costs rather than according to repurchase prices, and that factories should be obliged (or not obliged) to use wholesale dealers, then this will have economic and material consequences, but not necessarily consequences for freedom of thought.

If on the other hand your ‘rules of the game’ mean that businessmen may not use their bank accounts for purchases when they think that there are opportunities, or to buy from whomever they choose or to sell to whomever they choose, and that they may not charge the prices which offers and demand and risk determine, then we find ourselves in a coercive society which will, as both theoretical considerations and the empirical evidence tell us, present very limited possibilities for your ‘as much freedom and individuality as possible in the world of ideas.’

On your aside comments.

I am happy with your aside comments. They raise interesting questions which deserve a better analysis.

On Freedom of Thought.

You embrace Voltaire’s noble exclamation: ‘I disagree strongly with what you say but I would die defending your right to say it.’ This is also my motto, but would Voltaire also have defended this right against opponents who wanted to destroy freedom of thought? Against opponents who would not allow Voltaire to express his opinions, who would have burned his books, and locked him up, or made any discussion of ideas impossible? And would you?

You say that you are willing to fight for your neighbor’s right to say what he says, even when he says he is disagreeing with you. If I had known this, then I would have asked for your help in influencing your associate in political economy, Ole Colbjørnsen, who consistently refuses to publish factual corrections in the Workers’ Daily [Arbeiderbladet] and who even refuses to print your colleague Professor Keilhau’s responses to attacks.

Hopefully you agree that in a society where a large segment of the population becomes insular and dull by only reading one party organ, freedom of thought becomes illusory when the party organs refuse to publish divergent opinions.

One might ask whether the law on the freedom of the press and of assembly should not be complemented by a stipulation that would oblige editors and

organizers to provide opportunity for people with divergent opinions to express
them.

*On the Soviet System versus the Competitive System.*

As an explanation of your unwillingness to accept my arguments and of those
whom I quote, you refer to Soviet Russia’s effort during the war.

As will be seen below, I am quite happy that you refer to Soviet Russia, but
the example is irrelevant.

You contrast the competitive system with the Soviet system of central
planning during a war. The comparison should of course be made between the two
systems in peace, in the case when, if I may quote you, ‘the total value of output in
the communist system would be used to achieve the greatest possible welfare for
the total population.’

Economic planning becomes quite different and much easier during a war,
because it then has just one goal, to win the war, because all the available instru-
ments have to be directed towards that goal, and because the military experts can
then create preference functions to achieve the optimal utilization of the resources
available.

To quote myself: ‘If, for example, the society takes as its only aim the
development of a war machine of a certain size ..., a military central authority will,
presumably, be able better than any other to determine the relative value of existing
resources in the light of *this particular aim.* The task of the military authorities will be
considerably easier, if they can disregard questions of satisfying the wishes of the
individuals.’

There are economic experts (for example Professor Rüstow) who consider
free competition to be superior and indispensable also in wartime because varying
prices will reveal to the military command which raw materials are scarce, but I am
not going to discuss that.

I conclude that central economic planning using coercion is most effective
in war (and when the task is to eliminate economic variables for econometricians),
but I repeat that the premise of war is not acceptable when central planning is being
compared to free competition.

Even if it had been acceptable, your reference would not have had any force
because of the unrealistic simplification you are making.

The power of the Soviet military operations is also caused by several other
factors of which I mention the following: that Soviet Russia was with Germany

---

20. Hoff, *Economic Calculation*, 183. The italics of the Norwegian originals are missing in the English
translation.
the most militarized state in the world, that the youth are brought up as warriors and that the soldiers constitute a privileged (well-fed and well-equipped) class. In addition, one could mention psychological factors such as the fanatical patriotism of the Russians, their Messianic belief that they will save the world (through communism) and their strange and barbaric destructive urge which has made possible the complete and extraordinary destruction of cities, factories and resources (a very important factor). It is also well known that conditions of climate and nature have played a role.

But the munitions? Yes, they have been both good and not so good. I have personally spoken with people who watched the Russian troops in the Baltic countries and who are very critical of their equipment. In his book, *Red Flood Rising*, the representative of the Red Cross in Lithuania, Ignas J.–Scheynius, writes:21

We sum up our observations about the ability of the red transport vehicles and war machines to reach their destinations. The total number of vehicles on the roads is difficult to estimate but it may be as much as seven thousand. The vehicles pausing for repairs are according to my calculations about 61 percent. Wooderson [another Red Cross employee] is more cautious, his estimate is only 59.5 percent. . . . After a week I can compare the estimates of Wooderson and me to those of a Lithuanian engineer working in munitions. He has made observations in many places and on many occasions, concluding that on average 60 percent of the Soviet military vehicles come to a self-induced standstill on the roads. He considers the causes to be the following: second-rate material, defective manufacturing, inadequate care and poor organization of the repairs.

Even if most of the Russian weapons would be first class (as I believe), this would not say much about the Soviet system, because the aeroplanes and the weapons are partly imported and partly manufactured on the base of foreign patents and with foreign machinery.

In times when the Soviets were regarded as allies, they were quite adroit in obtaining expert advice and patents. For example, they got the ‘Hispano-Suiza’ patents for aeroplane motors from France and Curtiss-Wright patents from the United States. Presumably you would agree that weapons and aeroplanes obtained on this basis tell us more about the cleverness and foresight of Soviet diplomats than anything else.

21. Ignas Jurkunas (1889–1959) was a Lithuanian diplomat and author who wrote under the pen name Ignas Seinius. He changed the name to Scheynius when he fled to Sweden in 1940, having served for a while for the Red Cross in Lithuania. His book, *Den röda floden stiger* (Stockholm: Bonniers, 1940), has been republished under the name *Stalin ockuperar Lätten 1940* (Hallstavik: Svenskt militärhistoriskt bibliotek, 2008).
You mention people who have ‘proved’ that the Soviet system is inefficient and so inferior to the system of free competition that ‘it must sooner or later collapse like a house of cards.’ Are you not attacking a strawman which you have yourself invented? I know nobody who has asserted this.

Since you are obviously referring to me—although you are polite enough not to mention my name—I may mention that I have written almost the opposite, namely that ‘a socialist society with workers, material and engineering ability can build factories, workshops, and power stations, and produce consumer goods. Nor is there anything to prevent the managers, whether impelled by force or reward, from conducting them with technical efficiency. … The question, however, is not whether factories can be built and efficiently conducted, but whether the factors of production could have been put to a more advantageous use by employing them elsewhere.’

Here we return to another of your points, the assertion which seems to have troubled you so much: that Soviet central planning might lead to the use of molybdenum for toy swords. There is no reason however to despair because the doctoral candidate in question was right (as anyone who makes the effort to study the actual situation can easily confirm).

Leonard Hubbard who lived for many years in Russia writes: ‘For example a factory may use large quantities of rather scarce raw materials which would be needed in another enterprise producing more useful goods.”

The French economist Dr. Robert Mossé who is sympathetic to socialism, in 1939 published a book based on studies in Soviet Russia (which incidentally confirmed the correctness of the doctoral candidate mentioned above). Mossé writes: ‘The Soviet system does not take into account natural resources in their raw state and this makes impossible the rational allocation of resources to alternative uses, with enormous waste as a result.” Is it possible to get closer to the molybdenum toy swords?

And here you can get a taste of the efficiency of the system:

Pravda reports (8 August 1936) that of 9,992 automobiles examined, 1,958 turned out to be defective.

Pravda reports (23 September 1936) that 1,300 out of 2,345 chairs produced were unusable.

Pravda reports (18 November 1936) that according to plan the main gramophone records factory in Nojevik was supposed to deliver four million records.

22. Hoff, Economic Calculation, 190.
in 1935. It delivered 1,992,000 of which 309,800 were unusable. The number of rejected gramophone records in 1936 was on the increase. It was 156,200 in the first quarter, 259,400 in the second quarter, and 614,000 in the third quarter. In October 1936, 607,000 records were rejected.

In Pravda (23 September 1936) Professor Burdenko complains about the inferior quality of surgical equipment. 25

Pravda reports (4 November 1936) that 99 percent of the copybooks produced in the factory ‘Heroes of Labour’ were unusable.

Izvestia reports (12 December 1936) that 8,000,000 copybooks had to be thrown away in Rostov.

Do you believe this could happen in the system of free competition?

The inefficiency is not limited to the year 1936. Readers of Russian newspapers confirm that similar denunciations are made every year. Here is an extract of a long article the Soviet Commissar of Finance, Comrade Zverev, 26 writes in Pravda on 15 May 1939:

The Noginsky gramophone records factory has used too much of the material which can be used to produce gramophone records. … Instead of 230 g a record it uses on average 280 g. As a result, the factory produced 10,233,000 less than planned. … In several enterprises belonging to the People’s Commissariat for Machine Production the loss from rejected products is just as big in 1938 as it was in 1937. In the factory Kommunar (agricultural equipment) the loss from unusable products in 1938 amounted to 98 percent of the cost.

The newspaper Komsomolskaya Pravda reports on 17 May 1939:

The foundry guild was commissioned to cast eight cylinders for a servo motor. The deadline has long since been passed, but now only 5 of them have been cast, out of which 4 have been rejected. Of 47 shovels 23 have been rejected.

The Water Transport writes in May 1939:

There is no fight against delays and breakdowns. The fleet is not ready. The work on the largest Volga wharf has come to a halt. The lack of discipline aboard the ships is terrible. Carelessness and economic ignorance are common.

25. Nikolay N. Burdenko (1876–1946) was Professor of Neurosurgery at Moscow State University.
Do you believe that this could take place in a system of free competition? And these are not single instances. There are thousands of them. And remember that here we are talking about inefficiency in a sphere where you yourself have indicated a natural goal: to produce the greatest fulfilment of human needs possible for the whole population.

The results are even sadder when spiritual life is concerned although you say that you are not going to comment on it. I will therefore not dwell on it, but should we not discuss the effects of central economic planning on spiritual life?

**Economic Regulation and Monstrous Distortions.**

As an argument for more regulation of economic life, you refer to ‘the monstrous distortions of the Great Depression in the 1930s: the deliberate destruction of goods, the persistent unemployment and machines standing still,’ because you believe that they could ‘to a large extent be traced to some “individualistic” features of our economic system.’

It is here that I get troubled. I have heard and seen this assertion hundreds of times from economists supporting central planning and from would-be economists. A long time ago I realized that the case against the efficiency of the system of free competition is made by people who do whatever they can to obstruct the free economy (with the result that it has not stood a fair chance in Europe in the last twenty years). But I had not expected to hear this assertion from you.

I am sorry that I have to say that the truth is quite the opposite of what you assert. The monstrous distortions are caused by not allowing individual adjustments and free competition to operate. Minimum prices have been set—often imposed from above—which have affected both demand and supply, so that the pileup of goods—and consequently their destruction—has been inevitable.

Do you not know that it was with the consent of the Brazilian government that a lot of the coffee supply was destroyed? Do you not know that stipulations in the ‘Agricultural Adjustment Act’ in America led to wheat being burnt and six million piglets being slaughtered? This is indeed the case, and if you are in doubt, I can refer you to the existing literature on the topic, including reports from the Economic Committee of the League of Nations.

Your letter suggests that you believe these ‘monstrous distortions’ are derived from the Great Depression. This is not so. Price interferences (with pileups as results) have taken place in various countries since 1906, and the American stipulations which led to the pileup of goods, burning of wheat and slaughtering of piglets were introduced in 1929, long before the swing downwards.

If you however want to discuss the problems of depression and unemployment, then I am willing to do so, but then we must dig deeper and confirm the goals and premises (for example whether or not there should be a free choice between
trades and workplaces, and whether or not society should encourage economic growth, and so on).

Speaking of unemployment, are you aware of the enormous progress in the agro-biological field? The doubling of plant chromosomes is considered to be as momentous as the invention of the steam engine. Do you know that to cultivate a bushel of wheat today only requires two-fifths of the working hours required in 1914–1918? 27

If government wants to maintain the number of people employed in agriculture by regulation and minimum prices, then it is obvious that the result will be unemployment (unless the wheat will be used in industrial production—or as fuel).

Liberalism.

You suggest that ‘a return to economic liberalism “on a higher level” will take place’ when ‘in a natural process businessmen become sufficiently enlightened and public spirited.’

I have also hoped for the return to a new liberalism, but probably it is a different one from that which you wish for. I have hoped for a liberalism with regulation to ensure free competition and the price mechanism, and thus to frustrate the attempts of businessmen, with or without government support, to form monopolies, cartels and price arrangements in order to exploit the consumers.

I do not know what you mean by businessmen being ‘public spirited.’ If you are thinking about the awareness by industrialists and merchants of their group interests, I can assure you that this awareness is certainly there. After receiving help from the Thagaards of various countries, 28 they will be even keener on eliminating free competition than in Adam Smith’s times.

If you use ‘public interest’ in a political sense, something akin to ‘Gemeinnutz,’ 29 then we have some experience to guide us if you will make the effort to study it.

Conclusion.

I have enjoyed our exchange of ideas as a contribution to a real discussion. From a scholarly point of view I am happy to have alerted our only professor of political economy to circumstances of which one must know in order to discuss

---

27. Farm Economics, 124 (April 1941), 3115. Published by New York State College of Agriculture at Cornell University.
28. Wilhelm Thagaard (1890–1970) was a Norwegian lawyer and public servant. He was Director of the Price Directorate from 1940 to 1960. A law from 1945 was often named after him, Lex Thagaard, which gave extensive powers to the Price Directorate. It was abolished in 1953, when the Labour Party had to give in after a bitter confrontation with the business community and the Conservative Party.
29. ‘Gemeinnutz’ was the slogan of the German Nazi Party, ‘Gemeinnutz geht vor Eigennutz.’
these important and urgent questions.

‘Sub specie futurae,’ from the point of view of the future, this exchange of ideas has however been quite dispiriting because it has made me realize that national socialism will triumph—quite independently of the military outcome of the war.

For reasons which it would take too long to explain here, I still maintain a small hope about neo-liberalism, but the odds are against it given how the extraordinary lessons of the interwar period have been passed by.

It is not to be expected that the many economists supporting central planning because they seek power, will change their opinions. There is little reason, also, to rely on the business community, although I had not anticipated that it would so quickly and so strongly be influenced by the authoritarian mindset and autarchic arguments.

But I had put some hopes into political economists of an independent mind who were not seeking any personal gain. When I nevertheless see what an outstanding representative of this group can write without knowledge of the facts, I must agree with Professor Simons when he says: ‘The real enemies of liberty are the naïve advocates of managed economy or national planning.’

Greetings,

Trygve J. B. Hoff.

4. Frisch to Hoff

Oslo, 18 December 1941.

Dear Hoff,

Thank you for your long letter (produced as a duplicate) of 22 November. I apologize that I did not reply earlier.

In my response to your first message (on 4 November) I did not, contrary to what you asserted, discuss the Soviet system in war. I referred to what we can conclude about the Soviet system in the years preceding the war, precisely in the period under discussion in your doctoral dissertation for which your many quotations are relevant. The evidence shows that ‘proofs by quotations’ about Soviet Russia in this period are in vain. You must forgive me that this makes me in general skeptical about ‘proofs by quotations’.

Incidentally, I have no occasion to continue this discussion although it has been entertaining.

Greetings,

Ragnar Frisch.

PS. A short paper is attached. Sent by delivery.

**About the Authors**

**Trygve Hoff** (1895–1982) graduated in economics from the University of Oslo in 1916, became an economic commentator for Norwegian newspapers, and in 1938 completed a doctoral dissertation on economic calculation under socialism. After the Second World War he became somewhat of a classical liberal voice in the wilderness in Norway. Hoff was a founding member of the Mont Pelerin Society in 1947. In his magazine he published many articles and criticized restrictions implemented by the dominant Labour Party.

**Ragnar Frisch** (1895–1973) completed his doctorate on mathematical statistics at the University of Oslo in 1926 and later became Professor of Economics and Statistics at the University of Oslo and Director of the Institute of Economics at the University. A democratic socialist, Frisch became influential in Norway in the next few decades, filling the posts at his Institute with likeminded people and training many economists who became high officials, academics, or politicians, mostly in the Labour Party. He was, with Dutch economist Jan Tinbergen, a first recipient of the Nobel Prize in Economics in 1969.
Christianity Changes the Conditions of Government

LINK TO ABSTRACT

Foreword

Daniel B. Klein¹

After a lot of ribbing about assuming a can-opener, empiricist rituals, and so-called experimental evidence, 20th-century economists gradually came to a number of realizations:

• that to better understand economic affairs they needed to better understand something they called ‘institutions,’
• and then culture,
• and then human nature—
• and, throughout the preceding progression, history.

Economists’ thought and sentiment have been enlarged.

The explananda (that is, the things to be explained) have gotten bigger. How to explain the Great Enrichment? Even bigger, WEIRD? Yet still, the remarkable historical arc, since Homer, say, of the western peninsula of the Eurasian landmass.

Big ideas, necessarily speculative, lurk above and beyond narrow findings and theories, which may be helpful as details in the contemplation of broad speculations about the course of human societies and about human ontology itself—that is, about the constitution of the human being. Max U may be helpful, but it is not sufficient.

For example, there are broad speculations about human nature being rooted

¹. George Mason University, Fairfax, VA 22030.
in part in our existence as pack animals or members of small bands, instinctively responsive to manifest social nudges of our social environment. This speculation has led to interpretations of modern collectivist politics as a tapping into primeval bents and mentalities—that is, as an atavism, since collectivism is not apt for the modern world.

One speculation picks up the story in the ancient world, and then highlights a startling development—Christianity. This interpretation may be associated with our author here, the Frenchman Numa Denis Fustel de Coulanges (1830–1889), and his 1864 book *La Cité Antique: Étude sur le culte, le droit, les institutions de la Grèce et de Rome*. Fustel interprets the ancient city rather differently than it had been interpreted by thinkers during the so-called Renaissance and Enlightenment. Fustel’s ancient city is, as I see things, rather like an excrescence of band existence. The ancient world of the Greeks and Romans was an array or set of nested collections of human congelations, each firm in its manifest signals now elaborated as rites and rituals for its particular initiates.


Christianity began in the 1st century but liberalism emerges only much later—its arc cannot be said to have begun until sometime after Gutenberg’s printing press. So, why the lag? If Christianity made liberalism possible, why did it take some 16 or 17 centuries for liberalism to find elaborate expression?

I read Siedentop’s book shortly after it was published in 2014. From the dust jacket and opening pages, I gleaned the book’s contention, and thought to myself: Hmm, why’d it take so long? I would read a chapter, contemplate the developments explained in the chapter, and, indeed, found myself thinking: That would take long. And after the next chapter: That, too, would take long. And so on, chapter after chapter. Some of the developments are to political and jural organization within Western Christendom, developments that liberalism may be said to be predicated upon. To hear why it took so long to translate the moral intuitions of Christianity into social practice, then, you simply have to read the chapters of Siedentop’s book.

The story begins well before Christ, in Fustel’s antiquity. Fustel’s book is devoted to the ancient city. By entering into his understanding of the ancient city we better appreciate how transformative Christianity was. Today, we swim in the water of a world that is downstream of, as Siedentop (2014, 51) puts it, the world’s being “turned upside-down” by Christianity during its first several centuries. Only by projecting ourselves back into Fustel’s ancient city can we understand what social existence had once been, and the transformation that society underwent. The sweeping speculation gets us to rethink what we, after all, are, for the primeval and
ancient social forms are more native to us.

Fustel’s final chapter, reproduced here, is entitled “Christianity Changes the Conditions of Government.” It speaks of the transformative force of Christianity and opens the way to Siedentop’s great telling. Fustel writes in his final chapter:

Christ…separates religion from government. … It is the first time that God and the state are so clearly distinguished. … Christ…proclaims that religion is no longer the state, and that to obey Caesar is no longer the same thing as to obey God. … [T]his new principle was the source whence individual liberty flowed. (Fustel 1956/1864, 393–394)

In *The Theory of Moral Sentiments*, Adam Smith described the apotheotic delusions of Alexander the Great and then noted: “The religion and manners of modern times give our great men little encouragement to fancy themselves either Gods or even Prophets” (1790, 251).

Also reproduced here are two brief chapters from near the beginning of *The Ancient City*, “The Domestic Religion” and “Religion Was the Constituent Principle of the Ancient Family.” Fustel writes: “[R]eligion dwelt not in temples, but in the house; each house had its gods; each god protected one family only, and was a god only in one house” (1956/1864, 38).

Hanging over everything is the belief—of Fustel and of Siedentop—that man is, in his essence, a spiritual creature. Economists who pretend to a better understanding of economic affairs must know institutions, culture, history, to know the spirits of the creatures they would teach us about.

Siedentop says that many thinkers and leaders of the Renaissance and thereafter—thinkers and leaders who, Siedentop says, did not recognize that many of their own presuppositions about man and society had been the result of Christianization—misunderstood the ancient world as quite secular, in the sense of church-state separation (as opposed to the irreligiosity sense of secular). They knew little of household-specific gods and rituals. They thought that the rituals and legends of the Greek and Roman pantheons of famed gods were the main gods of ancient religion. As they saw devotions to such gods as rather secondary to the business of ancient existence, utilized chiefly only ceremonially and opportunistically, their image of ancient social and political life was one in which religion was nonessential. Siedentop says:

The trouble with this account is that it looks in the wrong place for religion… As Fustel de Coulanges demonstrated in *The Ancient City*, the religion of the Greek and Roman pre-history…spoke to and through the family. And it is to the family that we have to look to find religion and priesthood. The ancient family was itself a religious cult, with the father as its high priest tending the
family alter and its ‘sacred flame’, the flame that made his ancestors visible. Ancient religion thus consisted in worship of divine ancestors through the paterfamilias, a radical inequality of roles within the family and a series of elaborate ritual requirements. The family was, at least originally, a self-contained moral universe. It did not seek or welcome any deep or ‘moral’ connection with humans outside. (Siedentop 2014, 351, italics added)

Arguing that Christianity made liberalism possible, Siedentop explores the nature of Christianity in depth and at length, far beyond Fustel’s final chapter. Siedentop’s contention—which I, for one, embrace—ought not to be judged merely on the basis of reading Fustel’s final chapter.

The chapters of Fustel reproduced here are superbly written and provide a concise statement of the large speculation one may associate with Fustel and Siedentop, among others. That speculation carries with it many large suggestions, for example, that the animus, from say 1500, against the institutional manifestations of Christianity—the Church—was perhaps often a reaction against blobbish accretions not essential to the spirit of Christianity, and that that animus often threw the baby out with the bathwater—and, again, while sustaining presuppositions afforded by long centuries of Christianization.

A brief biography of Fustel is provided after the selection below.

About the text and the present reproduction: *La Cité Antique: Étude sur le culte, le droit, les institutions de la Grèce et de Rome* by Fustel de Coulanges was originally published in 1864 in Paris (Durand, Rue des Grèses). The English translation is by William Small, first published as *The Ancient City: A Study of the Religion, Laws, and Institutions of Greece and Rome* (Boston and New York: Lee and Shepard, 1874), the title page stating: “Translated from the last French edition.” To collect the text used here, I lifted the text from an online pdf of the book (here) and then made sure that the text matched my own physical copy of *The Ancient City* (Garden City, New York: Doubleday Anchor Books, 1956). In the reproduction below, Fustel’s footnotes have been eliminated, except for two placed in square brackets and indicated as having been footnote material.

**References**


Christianity Changes the Conditions of Government

Numa Denis Fustel de Coulanges

The Domestic Religion

We are not to suppose that this ancient religion resembled those founded when men became more enlightened. For a great number of centuries the human race has admitted no religious doctrine except on two conditions: first, that it proclaimed but one god; and, second, that it was addressed to all men, and was accessible to all, systematically rejecting no class or race. But this primitive religion fulfilled neither of these conditions. Not only did it not offer one only god to the adoration of men, but its gods did not accept the adoration of all men. They did not offer themselves as the gods of the human race. They did not even resemble Brahma, who was at least the god of one whole great caste, nor the Panhellenian Zeus, who was the god of an entire nation. In this primitive religion each god could be adored only by one family. Religion was purely domestic.

We must illustrate this important point; otherwise the intimate relation that existed between this ancient religion and the constitution of the Greek and Roman family may not be fully understood.

The worship of the dead in no way resembled the Christian worship of the saints. One of the first rules of this worship was, that it could be offered by each family only to those deceased persons who belonged to it by blood. The funeral obsequies could be religiously performed only by the nearest relative. As to the funeral meal, which was renewed at stated seasons, the family alone had a right to take part in it, and every stranger was strictly excluded. They believed that the dead ancestor accepted no offerings save from his own family; he desired no worship save from his own descendants. The presence of one who was not of the family disturbed the rest of the manes. The law, therefore, forbade a stranger to approach a tomb. To touch a tomb with the foot, even by chance, was an impious act, after which the guilty one was expected to pacify the dead and purify himself. The word by which the ancients designated the worship of the dead is significant; the Greeks said πατριάζειν, the Romans said parentare. The reason of this was because the prayer and offering were addressed by each one only to his fathers. The worship of the dead was nothing more than the worship of ancestors. [Footnote: In the
beginning at least; for later the cities had their local and national heroes, as we shall
see.] Lucian, while ridiculing common beliefs, explains them clearly to us when
he says the man who has died without leaving a son, receives no offerings, and is
exposed to perpetual hunger.

In India, as in Greece, an offering could be made to a dead person only
by one who had descended from him. The law of the Hindus, like Athenian law,
forbade a stranger, even if he were a friend, to be invited to the funeral banquet. It
was so necessary that these banquets should be offered by the descendants of the
dead, and not by others, that the manes, in their resting-place, were supposed often
to pronounce this wish: “May there be successively born of our line sons who, in all
coming time, may offer us rice, boiled in milk, honey, and clarified butter.”

Hence it was, that, in Greece and Rome, as in India, it was the son’s duty
to make the libations and the sacrifices to the manes of his father and of all his
ancestors. To fail in this duty was to commit the grossest act of impiety possible,
since the interruption of this worship caused the dead to fall from their happy state.
This negligence was nothing less than the crime of parricide, multiplied as many
times as there were ancestors in the family.

If, on the contrary, the sacrifices were always accomplished according to the
rites, if the provisions were carried to the tomb on the appointed days, then the
ancestor became a protecting god. Hostile to all who had not descended from him,
driving them from his tomb, inflicting diseases upon them if they approached, he
was good and provident to his own family.

There was a perpetual interchange of good offices between the living and the
dead of each family. The ancestor received from his descendants a series of funeral
banquets, that is to say, the only enjoyment that was left to him in his second life.
The descendant received from the ancestor the aid and strength of which he had
need in this. The living could not do without the dead, nor the dead without the
living. Thus a powerful bond was established among all the generations of the same
family, which made of it a body forever inseparable.

Every family had its tomb, where its dead went to repose, one after another,
always together. This tomb was generally near the house, nor far from the door,
“in order,” says one of the ancients, “that the sons, in entering and leaving their
dwelling, might always meet their fathers, and might always address them an invo-
cation.” Thus the ancestor remained in the midst of his relatives; invisible, but
always present, he continued to make a part of the family, and to be its father.
Immortal, happy, divine, he was still interested in all of him whom he had left
upon the earth. He knew their needs, and sustained their feebleness; and he who
still lived, who labored, who, according to the ancient expression, had not yet
discharged the debt of existence, he had near him his guides and his supports—his
forefathers. In the midst of difficulties, he invoked their ancient wisdom; in grief,
he asked consolation of them; in danger, he asked their support, and after a fault, their pardon.

Certainly we cannot easily comprehend how a man could adore his father or his ancestor. To make of man a god appears to us the reverse of religion. It is almost as difficult for us to comprehend the ancient creeds of these men as it would have been for them to understand ours. But, if we reflect that the ancients had no idea of creation, we shall see that the mystery of generation was for them what the mystery of creation is for us. The generator appeared to them to be a divine being; and they adored their ancestor. This sentiment must have been very natural and very strong, for it appears as a principle of religion in the origin of almost all human societies. We find it among the Chinese as well as among the ancient Getae and Scythians, among the tribes of Africa as well as among those of the new world.

The sacred fire, which was so intimately associated with the worship of the dead, belonged, in its essential character, properly to each family. It represented the ancestors; it was the providence of a family, and had nothing in common with the fire of a neighboring family, which was another providence. Every fire protected its own and repulsed the stranger. The whole of this religion was enclosed within the walls of each house. The worship was not public. All the ceremonies, on the contrary, were kept strictly secret. Performed in the midst of the family alone, they were concealed from every stranger. The hearth was never placed either outside the house or even near the outer door, where it would have been too easy to see. The Greeks always placed it in an enclosure, which protected it from the contact, or even the gaze, of the profane. The Romans concealed it in the interior of the house. All these gods, the sacred fire, the Lares, and the Manes, were called the consecrated gods, or gods of the interior. To all the acts of this religion secrecy was necessary. If a ceremony was looked upon by a stranger, it was disturbed, defiled, made unfortunate simply by this look.

There were neither uniform rules nor a common ritual for this domestic religion. Each family was most completely independent. No external power had the right to regulate either the ceremony or the creed. There was no other priest than the father: as a priest, he knew no hierarchy. The pontifex of Rome, or the archon of Athens, might, indeed, ascertain if the father of a family performed all his religious ceremonies; but he had no right to order the least modification of them. Suo quisque rite sacrificia faciat—such was the absolute rule. Every family had its ceremonies, which were peculiar to itself, its particular celebrations, its formulas of prayer, its hymns. The father, sole interpreter and sole priest of his religion, alone had the right to teach it, and could teach it only to his son. The rites, the forms of prayer, the chants, which formed an essential part of this domestic religion, were a patrimony, a sacred property, which the family shared with no one, and which they were even forbidden to reveal to strangers. It was the same in India. “I am strong
against my enemies,” says the Brahmin, “from the songs which I receive from my family, and which my father has transmitted to me.”

Thus religion dwelt not in temples, but in the house; each house had its gods; each god protected one family only, and was a god only in one house. We cannot reasonably suppose that a religion of this character was revealed to man by the powerful imagination of one among them, or that it was taught to them by a priestly caste. It grew up spontaneously in the human mind; its cradle was the family; each family created its own gods.

This religion could be propagated only by generation. The father, in giving life to his son, gave him at the same time his creed, his worship, the right to continue the sacred fire, to offer the funeral meal, to pronounce the formulas of prayer. Generation established a mysterious bond between the infant, who was born to life, and all the gods of the family. Indeed, these gods were his family—θεοὶ ἐγγενεῖς; they were of his blood—θεοὶ σύναιμοι. The child, therefore, received at his birth the right to adore them, and to offer them sacrifices; and later, when death should have deified him, he also would be counted, in his turn, among these gods of the family.

But we must notice this peculiarity—that the domestic religion was transmitted only from male to male.

This was owing, no doubt, to the idea that generation was due entirely to the males. The belief of primitive ages, as we find it in the Vedas, and as we find vestiges of it in all Greek and Roman law, was that the reproductive power resided exclusively in the father. The father alone possessed the mysterious principle of existence, and transmitted the spark of life. From this old notion it followed that the domestic worship always passed from male to male; that a woman participated in it only through her father or her husband; and, finally, that after death women had not the same part as men in the worship and the ceremonies of the funeral meal. Still other important consequences in private law and in the constitution of the family resulted from this: we shall see them as we proceed.

**Religion Was the Constituent Principle of the Ancient Family**

If we transport ourselves in thought to those ancient generations of men, we find in each house an altar, and around this altar the family assembled. The family meets every morning to address its first prayers to the sacred fire, and in the evening to invoke it for a last time. In the course of the day the members are once more assembled near the fire for the meal, of which they partake piously after prayer and libation. In all these religious acts, hymns, which their fathers have handed down,
are sung in common by the family.

Outside the house, near at hand, in a neighboring field, there is a tomb—the second home of this family. There several generations of ancestors repose together; death has not separated them. They remain grouped in this second existence, and continue to form an indissoluble family.

Between the living part and the dead part of the family there is only this distance of a few steps which separates the house from the tomb. On certain days, which are determined for each one by his domestic religion, the living assemble near their ancestors; they offer them the funeral meal, pour out milk and wine to them, lay out cakes and fruits, or burn the flesh of a victim to them. In exchange for these offerings they ask protection; they call these ancestors their gods, and ask them to render the fields fertile, the house prosperous, and their hearts virtuous.

Generation alone was not the foundation of the ancient family. What proves this is, that the sister did not bear the same relation to the family as the brother; that the emancipated son and the married daughter ceased completely to form a part of the family; and, in fine, several other important provisions of the Greek and Roman laws, that we shall have occasion to examine farther along.

Nor is the family principle natural affection. For Greek and Roman law makes no account of this sentiment. The sentiment may exist in the heart, but it is not in the law. The father may have affection for his daughter, but he cannot will her his property. The laws of succession—that is to say, those laws which most faithfully reflect the ideas that men had of the family—are in open contradiction both with the order of birth and with natural affection. [Footnote: It must be understood that we here speak of the most ancient law. We shall soon see that, at a later date, these early laws were modified.]

The historians of Roman laws, having very justly remarked that neither birth nor affection was the foundation of the Roman family, have concluded that this foundation must be found in the power of the father or husband. They make a sort of primordial institution of this power; but they do not explain how this power was established, unless it was by the superiority of strength of the husband over the wife, and of the father over the children. Now, we deceive ourselves sadly when we thus place force as the origin of law. We shall see farther on that the authority of the father or husband, far from having been a first cause, was itself an effect; it was derived from religion, and was established by religion. Superior strength, therefore, was not the principle that established the family.

The members of the ancient family were united by something more powerful than birth, affection, or physical strength; this was the religion of the sacred fire, and of dead ancestors. This caused the family to form a single body, both in this life and in the next. The ancient family was a religious rather than a natural association; and we shall see presently that the wife was counted in the family only after the
sacred ceremony of marriage had initiated her into the worship; that the son was no longer counted in it when he had renounced the worship, or had been emancipated; that, on the other hand, an adopted son was counted a real son, because, though he had not the ties of blood, he had something better—a community of worship; that the heir who refused to adopt the worship of this family had no right to the succession; and, finally, that relationship and the right of inheritance were governed not by birth, but by the rights of participation in the worship, such as religion had established them. Religion, it is true, did not create the family; but certainly it gave the family its rules; and hence it comes that the constitution of the ancient family was so different from what it would have been if it had owed its foundation to natural affection.

The ancient Greek language has a very significant word to designate a family. It is ἐπίστιον, a word which signifies, literally, *that which is near a hearth*. A family was a group of persons whom religion permitted to invoke the same sacred fire, and to offer the funeral repast to the same ancestors.

**Christianity Changes the Conditions of Government**

*The final chapter of the book*

The victory of Christianity marks the end of ancient society. With the new religion this social transformation, which we saw begun six or seven centuries earlier, was completed.

To understand how much the principles and the essential rules of politics were then changed, we need only recollect that ancient society had been established by an old religion whose principal dogma was that every god protected exclusively a single family or a single city, and existed only for that. This was the time of the domestic gods and the city-protecting divinities. This religion had produced laws; the relations among men—property, inheritance, legal proceedings—all were regulated, not by the principles of natural equity, but by the dogmas of this religion, and with a view to the requirements of its worship. It was this religion that had established a government among men; that of the father in the family; that of the king or magistrate in the city. All had come from religion,—that is to say, from the opinion that man had entertained of the divinity. Religion, law, and government were confounded, and had been but a single thing under three different aspects.

We have sought to place in a clear light this social system of the ancients, where religion was absolute master, both in public and private life; where the state was a religious community, the king a pontiff, the magistrate a priest, and the law
a sacred formula; where patriotism was piety, and exile excommunication; where individual liberty was unknown; where man was enslaved to the state through his soul, his body, and his property; where the notions of law and of duty, of justice and of affection, were bounded within the limits of the city; where human association was necessarily confined within a certain circumference around a prytaneum; and where men saw no possibility of founding larger societies. Such were the characteristic traits of the Greek and Italian cities during the first period of their history.

But little by little, as we have seen, society became modified. Changes took place in government and in laws at the same time as in religious ideas. Already in the fifth century which preceded Christianity, the alliance was no longer so close between religion on the one hand and law and politics on the other. The efforts of the oppressed classes, the overthrow of the sacerdotal class, the labors of philosophers, the progress of thought, had unsettled the ancient principles of human association. Men had made incessant efforts to free themselves from the thraldom of this old religion, in which they could no longer believe; law and politics, as well as morals, in the course of time were freed from its fetters.

But this species of divorce came from the disappearance of the ancient religion; if law and politics began to be a little more independent, it was because men ceased to have religious beliefs. If society was no longer governed by religion, it was especially because this religion no longer had any power. But there came a day when the religious sentiment recovered life and vigor, and when, under the Christian form, belief regained its empire over the soul. Were men not then destined to see the reappearance of the ancient confusion of government and the priesthood, of faith and the law?

With Christianity not only was the religious sentiment revived, but it assumed a higher and less material expression. Whilst previously men had made for themselves gods of the human soul, or of the great forces of nature, they now began to look upon God as really foreign by his essence, from human nature on the one hand, and from the world on the other. The divine Being was placed outside and above physical nature. Whilst previously every man had made a god for himself, and there were as many of them as there were families and cities, God now appeared as a unique, immense, universal being, alone animating the worlds, alone able to supply the need of adoration that is in man. Religion, instead of being, as formerly among the nations of Greece and Italy, little more than an assemblage of practices, a series of rites which men repeated without having any idea of them, a succession of formulas which often were no longer understood because the language had grown old, a tradition which had been transmitted from age to age, and which owed its sacred character to its antiquity alone,—was now a collection of doctrines, and a great object proposed to faith. It was no longer
exterior; it took up its abode especially in the thoughts of man. It was no longer matter; it became spirit. Christianity changed the nature and the form of adoration. Man no longer offered God food and drink. Prayer was no longer a form of incantation; it was an act of faith and a humble petition. The soul sustained another relation with the divinity; the fear of the gods was replaced by the love of God.

Christianity introduced other new ideas. It was not the domestic religion of any family, the national religion of any city, or of any race. It belonged neither to a caste nor to a corporation. From its first appearance it called to itself the whole human race. Christ said to his disciples, “Go ye into all the world, and preach the gospel to every creature.”

This principle was so extraordinary, and so unexpected, that the first disciples hesitated for a moment; we may see in the Acts of the Apostles that several of them refused at first to propagate the new doctrine outside the nation with which it had originated. These disciples thought, like the ancient Jews, that the God of the Jews would not accept adoration from foreigners; like the Romans and the Greeks of ancient times, they believed that every race had its god, that to propagate the name and worship of this god was to give up one’s own good and special protector, and that such a work was contrary at the same time to duty and to interest. But Peter replied to these disciples, “God gave the gentiles the like gift as he did unto us.” St. Paul loved to repeat this grand principle on all occasions, and in every kind of form. “God had opened the door of faith unto the gentiles.” “Is he the God of the Jews, only? Is he not also of the gentiles?” “We are all baptized into one body, whether we be Jews or gentiles.”

In all this there was something quite new. For, everywhere, in the first ages of humanity, the divinity had been imagined as attaching himself especially to one race. The Jews had believed in the God of the Jews; the Athenians in the Athenian Pallas; the Romans in Jupiter Capitolinus. The right to practice a worship had been a privilege. The foreigner had been repulsed from the temple; one not a Jew could not enter the temple of the Jews; the Lacedaemonian had not the right to invoke the Athenian Pallas. It is just to say, that, in the five centuries which preceded Christianity, all who thought were struggling against these narrow rules. Philosophy had often taught, since Anaxagoras, that the god of the universe received the homage of all men, without distinction. The religion of Eleusis had admitted the initiated from all cities. The religion of Cybele, of Serapis, and some others, had accepted, without distinction, worshippers from all nations. The Jews had begun to admit the foreigner to their religion; the Greeks and the Romans had admitted him into their cities. Christianity, coming after all this progress in thought and institutions, presented to the adoration of all men a single God, a universal God, a God who belonged to all, who had no chosen people, and who made no distinction.
in races, families, or states.

For this God there were no longer strangers. The stranger no longer profaned the temple, no longer tainted the sacrifice by his presence. The temple was open to all who believed in God. The priesthood ceased to be hereditary, because religion was no longer a patrimony. The worship was no longer kept secret; the rites, the prayers, the dogmas were no longer concealed. On the contrary, there was thenceforth religious instruction, which was not only given, but which was offered, which was carried to those who were the farthest away, and which sought out the most indifferent. The spirit of propagandism replaced the law of exclusion.

From this great consequences flowed, as well for the relations between nations as for the government of states.

Between nations religion no longer commanded hatred; it no longer made it the citizen’s duty to detest the foreigner; its very essence, on the contrary, was to teach him that towards the stranger, towards the enemy, he owed the duties of justice, and even of benevolence. The barriers between nations or races were thus thrown down; the pomerium disappeared. “Christ,” says the apostle, “hath broken down the middle wall of partition between us.” “But now are they many members,” he also says, “yet but one body.” “There is neither Greek nor Jew, circumcision nor uncircumcision, Barbarian, Scythian, bond nor free: but Christ is all, and in all.”

The people were also taught that they were all descended from the same common father. With the unity of God, the unity of the human race also appeared to men’s minds; and it was thenceforth a religious necessity to forbid men to hate each other.

As to the government of the state, we cannot say that Christianity essentially altered that, precisely because it did not occupy itself with the state. In the ancient ages, religion and the state made but one; every people adored its own god, and every god governed his own people; the same code regulated the relations among men, and their duties towards the gods of the city. Religion then governed the state, and designated its chiefs by the voice of the lot, or by that of the auspices. The state, in its turn, interfered with the domain of the conscience, and punished every infraction of the rites and the worship of the city. Instead of this, Christ teaches that his kingdom is not of this world. He separates religion from government. Religion, being no longer of the earth, now interferes the least possible in terrestrial affairs. Christ adds, “Render to Caesar the things that are Caesar’s, and to God the things that are God’s.” It is the first time that God and the state are so clearly distinguished. For Caesar at that period was still the pontifex maximus, the chief and the principal organ of the Roman religion; he was the guardian and the interpreter of beliefs. He held the worship and the dogmas in his hands. Even his person was sacred and divine, for it was a peculiarity of the policy of the emperors that, wishing to recover the attributes of ancient royalty, they were careful not to forget
the divine character which antiquity had attached to the king-pontiffs and to the priest-founders. But now Christ breaks the alliance which paganism and the empire wished to renew. He proclaims that religion is no longer the state, and that to obey Caesar is no longer the same thing as to obey God.

Christianity completes the overthrow of the local worship; it extinguishes the prytanea, and completely destroys the city-protecting divinities. It does more; it refuses to assume the empire which these worships had exercised over civil society. It professes that between the state and itself there is nothing in common. It separates what all antiquity had confounded. We may remark, moreover, that during three centuries the new religion lived entirely beyond the action of the state; it knew how to dispense with state protection, and even to struggle against it. These three centuries established an abyss between the domain of the government and the domain of religion; and, as the recollection of this period could not be effaced, it followed that this distinction became a plain and incontestable truth, which the efforts even of a part of the clergy could not eradicate.

This principle was fertile in great results. On one hand, politics became definitively freed from the strict rules which the ancient religion had traced, and could govern men without having to bend to sacred usages, without consulting the auspices or the oracles, without conforming all acts to the beliefs and requirements of a worship. Political action was freer; no other authority than that of the moral law now impeded it. On the other hand, if the state was more completely master in certain things, its action was also more limited. A complete half of man had been freed from its control. Christianity taught that only a part of man belonged to society; that he was bound to it by his body and by his material interests; that when subject to a tyrant, it was his duty to submit; that as a citizen of a republic, he ought to give his life for it, but that, in what related to his soul, he was free, and was bound only to God.

Stoicism had already marked this separation; it had restored man to himself, and had founded liberty of conscience. But that which was merely the effort of the energy of a courageous sect, Christianity made a universal and unchangeable rule for succeeding generations; what was only the consolation of a few, it made the common good of humanity.

If, now, we recollect what has been said above on the omnipotence of the states among the ancients,—if we bear in mind how far the city, in the name of its sacred character and of religion, which was inherent in it, exercised an absolute empire,—we shall see that this new principle was the source whence individual liberty flowed.

The mind once freed, the greatest difficulty was overcome, and liberty was compatible with social order.

Sentiments and manners, as well as politics, were then changed. The idea
which men had of the duties of the citizen were modified. The first duty no longer consisted in giving one’s time, one’s strength, one’s life to the state. Politics and war were no longer the whole of man; all the virtues were no longer comprised in patriotism, for the soul no longer had a country. Man felt that he had other obligations besides that of living and dying for the city. Christianity distinguished the private from the public virtues. By giving less honor to the latter, it elevated the former; it placed God, the family, the human individual above country, the neighbor above the city.

Law was also changed in its nature. Among all ancient nations law had been subject to, and had received all its rules from, religion. Among the Persians, the Hindus, the Jews, the Greeks, the Italians, and the Gauls, the law had been contained in the sacred books or in religious traditions, and thus every religion had made laws after its own image. Christianity is the first religion that did not claim to be the source of law. It occupied itself with the duties of men, not with their interests. Men saw it regulate neither the laws of property, nor the order of succession, nor obligations, nor legal proceedings. It placed itself outside the law, and outside all things purely terrestrial. Law was independent; it could draw its rules from nature, from the human conscience, from the powerful idea of the just that is in men’s minds. It could develop in complete liberty; could be reformed and improved without obstacle; could follow the progress of morals, and could conform itself to the interests and social needs of every generation.

The happy influence of the new idea is easily seen in the history of Roman law. During several centuries preceding the triumph of Christianity, Roman law had already been striving to disengage itself from religion, and to approach natural equity; but it proceeded only by shifts and devices, which enervated and enfeebled its moral authority. The work of regenerating legislation, announced by the Stoic philosophers, pursued by the noble efforts of Roman jurisconsults, outlined by the artifices and expedients of the pretor, could not completely succeed except by favor of the independence which the new religion allowed to the law. We can see, as Christianity gained ground, that the Roman codes admitted new rules no longer by subterfuges, but openly and without hesitation. The domestic penates having been overthrown, and the sacred fires extinguished, the ancient constitution of the family disappeared forever, and with it the rules that had flowed from this source. The father had lost the absolute authority which his priesthood had formerly given him, and preserved only that which nature itself had conferred upon him for the good of the child. The wife, whom the old religion placed in a position inferior to the husband, became morally his equal. The laws of property were essentially altered; the sacred landmarks disappeared from the fields; the right of property no longer flowed from religion, but from labor; its acquisition became easier, and the formalities of the ancient law were definitively abolished.
Thus, by the single fact that the family no longer had its domestic religion, its constitution and its laws were transformed; so, too, from the single fact that the state no longer had its official religion, the rules for the government of men were forever changed.

Our study must end at this limit, which separates ancient from modern polities. We have written the history of a belief. It was established, and human society was constituted. It was modified, and society underwent a series of revolutions. It disappeared, and society changed its character. Such was the law of ancient times.

About the Author

Numa Denis Fustel de Coulanges (1830–1889) was a French historian and became an eminent professor in the chair of history at Strasbourg and subsequently the École Normale Supérieure and the Sorbonne. His most famous work is *The Ancient City* (1864; English translation 1874), which emphasized that in ancient Greece and Rome, the person’s most important religion was located in the family, its ancestors, and its domestic house. He subsequently wrote a multivolume work on the political institutions of ancient France. Another work by Fustel available in English is *The Origin of Property in Land*, a translation of an article he published in French in 1889. In that work Fustel questions the historiographical basis for claims to the effect that agricultural land had been communally owned before it was privately owned.