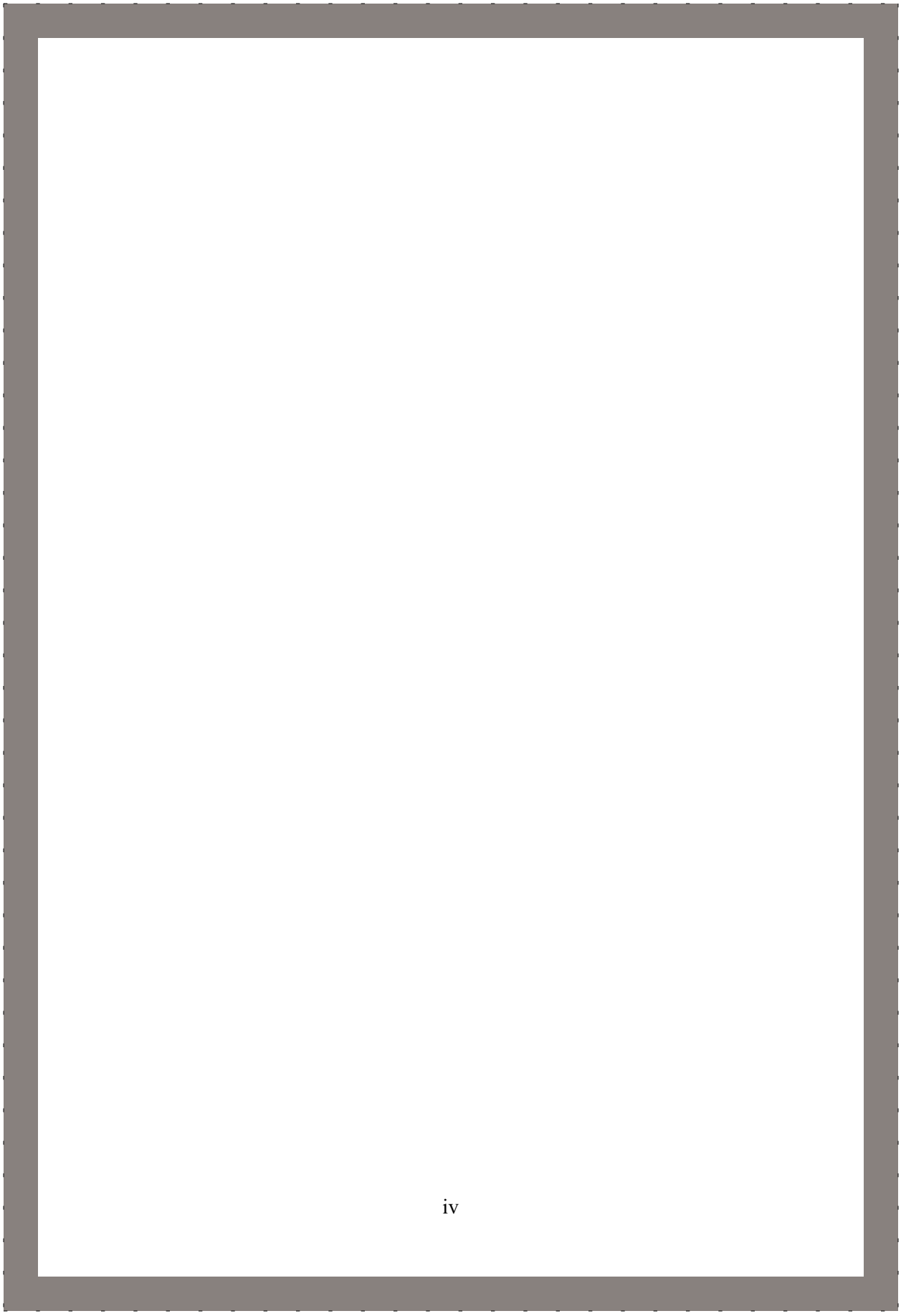
MEMBERS

JACK ABRAMS
MADISON BRESHEARS
ARMON GHAYOUMI
DILLON OSTLUND
HANNAH PETRUZZI
WALTER SMITH
DAVID WARD

44203-gme_17-1 Sheet No. 1 Side B 06/20/2022 12:22:58

CONTENTS

44203-gme_17-1 Sheet No. 2 Side A 06/20/2022 12:22:58



44203-gme_17-1 Sheet No. 2 Side B 06/20/2022 12:22:58

File: Bujakowski 17.1 To Publish_Edited Created on: 6/13/2022 5:47:00 PM Last Printed: 6/14/2022 9:27:00 PM

THE DECLINE AND RESURGENCE OF PEOPLE'S MEDIATION IN CHINA: AN EMPIRICAL ANALYSIS OF CHINESE PROVINCES

Douglas Bujakowski[1]

I. INTRODUCTION

A striking characteristic of China's legal system is the profound importance of people's mediation in the resolution of disputes. From 1995 to 2016, the number of claims for people's mediation exceeded the number of civil lawsuits filed with Chinese courts in all but four years.[2] Historically, people's mediation has played an even larger role, dwarfing the number of civil disputes resolved through any other means.[3]

People's mediation is a community-based form of dispute resolution with a rich tradition in China's history. The constitution of the People's Republic of China even mandates a people's mediation system, stating that, "residents' and villagers' committees [shall] establish committees for people's mediation, public security, public health and other matters in order to manage public affairs and social services in their areas, mediate civil disputes, help maintain public order and convey residents' opinions and demands and make suggestions to the people's government."[4] Committees tasked with the provision of mediation services are known as People's Mediation Committees (PMCs). As of 2016, 784,000 PMCs exist nationwide, employing 3,852,000 mediators.[5] Parties seeking people's mediation do so voluntarily, without charge, and often without legal representation.[6] If any

---

[1] Corresponding author. College of Business and Public Administration, Drake University, 2507 University Ave, Des Moines, IA 50311, 515-271-2871, doug.bujakowski@drake.edu.

[2] NAT'L BUREAU OF STAT. OF CHINA, http://www.stats.gov.cn/english/Statisticaldata/AnnualData/. The four years in which civil lawsuits exceeded people's mediation claims were 2008, 2009, 2015, and 2016.

[3] *See* Stanley Lubman, *Mao and Mediation: Politics and Dispute Resolution in Communist China*, 55 CAL. L. REV. 1284, 1286-87 (1967); Jerome A. Cohen, *Chinese Mediation on the Eve of Modernization*, 54 CAL. L. REV. 1201, 1201-02 (1966).

[4] CONST. OF CHINA. Dec. 4, 1982, § 5, art. 111.

[5] NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

[6] *See The People's Mediation Law of the People's Republic of China*, art. 2, 3 & 4 (promulgated by the Standing Comm. of the Nat'l People's Cong., Aug. 28, 2010, effective Jan. 1, 2011), https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/85806/96276/F1660942158/CHN85806 [hereinafter People's Mediation Law]; Bee Chen Goh, LAW WITHOUT LAWYERS, JUSTICE WITHOUT COURTS: ON TRADITIONAL CHINESE MEDIATION, 10-13 (2016).

party chooses not to participate (or to cease participation), the mediation process is terminated. If a mediated resolution is reached, an agreement stating the terms of the resolution may be drawn up and signed; however, until recently, the legal weight of such agreements has been questionable at best.[7]

The characteristics of people's mediation distinguish it from other types of mediation in China, such as judicial mediation and arbitral mediation. People's mediation is conducted by grassroots community mediators, whereas judicial and arbitral mediation are conducted by judges and arbitral administrative bodies, respectively. Additionally, judicial and arbitral mediation are often not a deliberate choice of parties to a dispute. Judicial mediation is often a necessary step in pursuing litigation, whether formalized in regulation or not, while arbitral mediation is often a contractually mandated form of recourse in specified situations.

Despite China's strong tradition of people's mediation, the utilization of this system has shifted dramatically over the past three decades. Utilization can be measured using mediation rates, defined as the number of people's mediation claims filed in a jurisdiction per 10,000 people. Figure 1 shows annual people's mediation rates in China from 1985 to 2016.[8] In the late 1980s, annual people's mediation rates fluctuated between 59.8 and 68 claims per 10,000 people.[9] However, through the 1990s and early 2000s, rates dropped by nearly 50 percent, reaching just 34 claims per 10,000 people in 2004.[10] Since the mid-2000s, people's mediation has experienced a tremendous resurgence, boasting over 65 claims per 10,000 people each year since 2011.[11]

We are by no means the first to observe these trends. Academics, lawyers, and policy experts have tracked mediation developments and sought explanations for their occurrence.[12] To date, a number of thoughtful, nuanced theories have emerged.[13] Though many are intuitively appealing and reveal a deep understanding of mediation processes in China, they lack empirical scrutiny and validation.[14] We aim to fill this gap.

---

7 *See* People's Mediation Law, *supra* note 5, art. 28, 29 & 31; Aaron Halegua, *Reforming the People's Mediation System in Urban China*, 35 HONG KONG L.J. 715, 721-22 (2005).

8 NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

9 *Id.*

10 *Id.*

11 *Id.*

12 *See*, *e.g.*, Yuning Wu, *People's Mediation Enters the 21st Century*, 10 J. COMP. L. 25, 27-29 (2015); Wenjia Zhuang & Feng Chen, *"Mediate First": The Revival of Mediation in Labour Dispute Resolution in China*, 222 CHINA Q. 380, 399-400 (2015); Hualing Fu & Richard Cullen, *From Mediatory to Adjudicatory Justice: The Limits of Civil Justice Reform in China*, *in* CHINESE JUST.: CIV. DISP. RESOL. IN CONTEMP. CHINA 25, 43 (Margaret Y. K. Woo & Mary E. Gallagher eds., 2011); Carl F. Minzner, *China's Turn Against Law*, 59 AM. J. COMP. L. 935, 940-47 (2011); John S Mo, *Understanding the Role of People's Mediation in the Age of Globalization*, 17 ASIA PAC. L. REV. 75, 82-85 (2009); Xiaohua Di & Yuning Wu, *The Developing Trend of the People's Mediation in China*, 42 SOCIO. FOCUS 228, 234-41 (2009); Halegua, *supra* note 6, at 718-24.

13 *See, e.g.*, Minzner, *supra* note 11, at 240; Halegua, *supra* note 6, at 719.

14 *See*, *e.g.*, Di & Wu, *supra* note 11, at 230.

In this paper, we employ 32 years (1985-2016) of province-level data from China to test a variety of factors thought to influence the development of people's mediation rates. Results suggest that legislative reforms, numbers of underlying disputes, economic development, and demographic shifts are related to people's mediation rates. Notably, we do not find evidence of a substitution effect of litigation for people's mediation – a commonly cited reason for the decline of people's mediation in the 1990s and early 2000s. To our knowledge, we are the first to conduct such empirical tests.

Given the limited availability of Chinese data dating back to the 1980s, there are times when we simply cannot provide as much information as we would like. For instance, seven of China's 31 provinces, municipalities, and administrative regions do not publicly report annual mediation counts, and those that do report generally do not provide details about the types of disputes mediated. Nevertheless, much can be learned from the sample available to us. By empirically investigating the forces shaping people's mediation outcomes, we are able to evaluate proposed explanations for the decline and resurgence of people's mediation in China. Our findings support several of theories of mediation development, while calling others into question. These insights may be relevant to mediators and PMCs as they prepare for claims and to policymakers faced with the continued development and refinement of China's people's mediation system.

The paper proceeds as follows. In section two, we review prior studies to identify factors that might explain the decline and resurgence of people's mediation observed in recent decades. In section three, we discuss data and methods used to test those factors. In section four, we present our results. In section five, we conclude with a discussion of our findings and suggestions for future research.

II. BACKGROUND AND HYPOTHESIS DEVELOPMENT

The aim of this study is to identify and test factors thought to influence the development of people's mediation rates across provinces and time. In this section, we examine relevant studies to identify those factors and their potential interactions with people's mediation rates. To organize our discussion, we group factors into five broad categories: (1) reforms to the people's mediation system, (2) access to people's mediation services, (3) economic development, (4) demographic shifts, and (5) the role of litigation. Importantly, explanations within and across categories are not mutually exclusive. This categorization scheme simply serves to provide structure around the numerous and, at times, interdependent ideas put forth.

A. *REFORMS TO THE PEOPLE'S MEDIATION SYSTEM*

People's mediation in China dates back to China's imperialist era (221 BCE – 1911 CE). During this period, emperors over many centuries were vested with power in all arenas, including the legal sector. The law itself focused on the "law of punishment," or today's criminal law.[15] Civil harms were primarily addressed through social norms arising out of Confucianism. Emperors attempted to develop general rules for civil behavior by issuing edicts associated with certain types of disputes.[16] In practice however, these rules were often too specific in their application and failed to provide general principles concerning civil liability.[17] As a result, officials resorted to adjudicating individual cases as they arose.[18]

Faced with the large burden of handling individual cases, officials encouraged the local resolution of disputes.[19] Minor disagreements, such as those dealing with family matters or land, comprised the majority of cases.[20] These disputes were generally mediated by respected community leaders or village elders, who applied customary rules and concepts of morality to reach harmonious resolutions.[21] The existence of a strong, community-based mediation system is a tradition somewhat unique to China that is still widely employed as a modern mechanism of dispute resolution.[22]

Modern people's mediation practice began in 1954, under the rule of Mao Zedong, with the establishment of the first PMCs.[23] Maoists viewed people's mediation not only as a means to resolve disputes, but also as a political endeavor.[24] During this time, PMCs suppressed disputes that interfered with the Communist Party's goals and "organized meetings for adopting 'patriotic pacts' in which all present vowed to obey Chairman Mao, the Communist Party, labor discipline, policies, law, and Communist morality."[25]

People's mediation in the reform era (1976-1989) was far less politicized than during the rule of Mao. Nevertheless, mediation remained a point of criticism among individuals and government officials for its neglect of applicable laws in resolving disputes. According to Palmer, Mediators regularly sought compromises and concessions from parties, even when

---

15 *See* Ye Lin, *The Tort System in China*, 52 LAW & CONTEMP. PROBS. 143, 144-45 (1989).
16 *Id.*
17 *Id.*
18 *Id.*
19 *See, e.g.*, Cohen, *supra* note 2, at 1219-21.
20 *Id.*
21 *Id.*
22 *Id.* at 1211.
23 *Id.* n.4.
24 *Id.* at 1204.
25 Lubman, *supra* note 2, at 1316.

applicable laws strongly supported their cases.[26] In 1989, this criticism was formally addressed through the passage of the Regulations for the Organization of People's Mediation Committees (hereinafter 1989 Regulations).[27] Per these regulations, mediators are to perform their work "in conformity with the laws, statutes, rules, regulations and policies of the state," and are required to possess knowledge of the law.[28]

In addition to establishing the concept of "mediation based on law" and making mediators more professional, the 1989 Regulations also state that parties "should" carry out mediated agreements.[29] Despite this sentiment, most judges at the time held that people's mediation agreements could not be legally enforced.[30] According to Halegua, some judges required a trial to ensure the legality of people's mediation agreements, while most ignored the existence of people's mediation agreements altogether when hearing associated cases.[31] The wider population seemed to share this view. In a 2002 survey of 249 Beijing residents, Wang found that 80 percent of respondents perceived that people's mediation agreements did not carry any legal weight in court.[32] Perhaps the greatest critics of unenforceability were mediators. In interviews with Halegua, mediators expressed frustration that parties frequently violated agreements with impunity.[33]

In response to these concerns, reforms were again enacted in 2002, with the issuance of three documents in close succession (hereinafter "2002 Reforms"): (1) a "Judicial Interpretation on Hearing Civil Cases Involving People's Mediation Agreements" issued by the Supreme People's Court,[34] (2) "Some Provisions Concerning the Work of People's Mediation" issued by the Ministry of Justice,[35] and (3) "Opinions of the Supreme People's Court and the Ministry of Justice on Further Enhancing People's Mediation in the New

---

26 Michael Palmer, *The Revival of Mediation in the People's Republic of China: Extra-Judicial Mediation*, *in* YEARBOOK ON SOCIALIST LEGAL SYSTEMS 220, 233 (William E. Butler ed., 1988).

27 Renmin Tiaojie Weiyuanhui [*Regulations for the Organization of People's Mediation Committees*], (promulgated by the St. Council, June 17, 1989) art. 4, 6 (China) [hereinafter *1989 Regulations*].

28 *Id.* art. 10.

29 *Id.*

30 For comparison, mediation agreements in the U.S., U.K., and European Union constitute legal contracts that are enforceable by courts.

31 Halegua, *supra* note 6, at 725.

32 Ting Wang, EXTRAJUDICIAL MEDIATION OF CIVIL DISPUTES IN CHINA: THE DECLINE OF PEOPLE'S MEDIATION AND INSTITUTIONAL RESPONSE 38 (Mar. 2002) (Ph.D. Dissertation, Harvard University) (on file with the Harvard University Library system).

33 Halegua, *supra* note 6, at 722.

34 *Judicial Interpretation by the Supreme People's Court on Hearing Civil Cases Involving People's Mediation Agreement* (promulgated by the Supreme People's Court, Sep. 5, 2002) (China) [hereinafter *Hearing Civil Cases*].

35 Ren Min Tiao Jie Gong Zuo Ruo Gan Gui Ding [*Some Provisions Concerning the Work of People's Mediation*] (promulgated by Ministry of Justice, Sept. 26, 2002, effective Nov. 1, 2002) (China) [hereinafter *Provisions Concerning Mediation*].

Era" issued by the General Office of the Central Committee of the Chinese Communist Party and the General Office of the State Council.[36]

The 2002 Reforms clarify the legal status of people's mediation agreements, indicating that when such agreements are signed by both parties they become civil contracts.[37] As a result, signed agreements cannot be arbitrarily changed or absolved by either party.[38] The three documents also detail content that should be included in mediated agreements, provide instructions for mediators to check whether resolutions have been implemented, and outline procedures when implementation has not occurred.[39]

The 2002 Reforms also make people's mediation more attractive to disputants in other ways. They make procedures more standardized and based on law, increase privacy surrounding the people's mediation process, set time limits for the duration of mediated cases, increase the quality and professionalism of people's mediators, and explicitly outline disputants' rights, including the right to request a mediator or to withdraw from people's mediation.[40]

Despite the vast array of improvements ushered in by the 2002 Reforms, enforcement of mediated agreements by courts remained somewhat cumbersome, in that a violation of a people's mediation agreement constituted a breach of contract that would need to be litigated. This reality changed in 2011, when the People's Mediation Law of the People's Republic of China (hereinafter "PML"), adopted in 2010, became effective.[41]

The PML codifies into law the provisions of the 2002 Reforms.[42] Additionally, it provides a mechanism for direct enforcement of people's mediation agreements by courts.[43] According to the law, parties to a people's mediation agreement may apply to the People's Court for judicial confirmation.[44] If one party refuses to fulfill the terms of a judicially confirmed agreement, the other party may apply to the People's Court for mandatory execution.[45] This mechanism eliminates the need to litigate unfulfilled people's mediation agreements.[46]

---

36 *Opinions of the Supreme People's Court and the Ministry of Justice on Further Enhancing People's Mediation in the New Era* (promulgated by the General Office of the Central Committee of the Chinese Communist Party and the General Office of the State Council) (China) [hereinafter *Enhancing Mediation*].

37 *Hearing Civil Cases*, *supra* note 33; *Provisions Concerning Mediation*, *supra* note 34; *Enhancing Mediation*, *supra* note 35.

38 *Hearing Civil Cases*, *supra* note 33, art. 1.

39 *Provisions Concerning Mediation*, *supra* note 34, art. 35, 36 & 37.

40 *See id.* art. 6, 7, 29, 33, & 40.

41 *People's Mediation Law*, *supra* note 5.

42 *Id.*

43 *Id.*

44 *Id.*

45 *Id.*, art. 33.

46 *Id.*

The 1989 Regulations, 2002 Reforms, and 2011 PML have reshaped China's system of people's mediation, making the prospect of mediation more attractive to disputants. As such, these reforms are often cited as the primary explanation for the post-2004 resurgence of people's mediation in China.[47] To account for the potential influence of these reforms, we construct three binary variables that indicate whether or not each reform is in effect. Each variable takes the value 0 before its effective date and the value 1 after that date.[48]

B. *ACCESS TO MEDIATION SERVICES*

In China, people's mediation is provided free of charge.[49] As a result, costs associated with people's mediation primarily include time and effort needed to secure mediation services and undergo mediation processes. It stands to reason that when people's mediation services are more readily available, search costs will decline and people's mediation services will be used more frequently. Perhaps the most direct measure of the availability of people's mediation services is the number of people's mediators and the number of PMCs in a jurisdiction. Accordingly, we include in our analysis the number of people's mediators per 10,000 people and the number of PMCs per 10,000 people.

Despite the resurgence of mediation in recent years, official statistics report that both the number of people's mediators and the number of PMCs declined steadily from 1995 to 2016.[50] One possibility is that people's mediators and PMCs have become increasingly effective in resolving disputes, which has reduced the number of people's mediators and PMCs needed to handle cases. Indeed, several authors point to the emergence of "grand mediation" in the early 2000s as a key driver of increased effectiveness.[51] Grand mediation involves the coordination of mediation activities among PMCs, courts, and administrative bodies and the expansion of this mediation network into areas where no formal mediation organizations previously existed.[52]

---

47 *See*, *e.g.*, Halegua, *supra* note 6, at 724-29; Wu, *supra* note 11, at 28-29.

48 The 1989 Regulations and 2002 Reforms became effective mid-year. As such, one must decide whether to count 1989 and 2002 as effective years or noneffective years. We opt to count both as effective years and set 1989 Regulations and 2002 Reforms equal to 1 in those years, respectively. Counting these years as noneffective does not alter our results.

49 People's Mediation Law, *supra* note 5, art. 4; *See* Vai Io Lo, *Resolution of Civil Disputes in China*, 18 UCLA PAC. BASIN L.J. 117, 132 (2000) ("Mediation is free of charge").

50 NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

51 *See*, *e.g.*, Jieren Hu, *Grand Mediation in China*, 51 ASIAN SURV. 1065, 1072 (2011); Zhuang & Chen, *supra* note 11, at 395-99.

52 *Grand Mediation in China*, *supra* note 50, at 1077-79.

If people's mediators and PMCs are not growing in number, but are instead being relocated to more strategic areas, aggregate counts of people's mediators and PMCs may not fully capture one's ability to access mediation services. To add nuance to these measures, we consider whether individuals located in strategic placement areas are seeking people's mediation at higher rates than those in other areas.

One particular area of strategic focus has been the establishment of PMCs in the workplace. The 1989 Regulations specifically allowed for the establishment of PMCs by "enterprises and public institutions" and by 2001, roughly 100,000 factory, mine, and enterprise PMCs had been created.[53] The city of Dongwan, Guangdong has even experimented with the establishment of PMCs in foreign-owned companies.[54] If this strategic placement is effective in increasing utilization, we should expect employed individuals to seek people's mediation more often than unemployed individuals, all else equal. Yet, it may be that unemployment increases the number of disputes among parties, resulting in more claims for people's mediation. As a result, the association between unemployment and the use of people's mediation is an empirical question. To investigate this relationship, we include a measure of unemployment in our analysis, defined as the percentage of the labor force that is jobless.

C. *ECONOMIC DEVELOPMENT*

The Chinese economy has fundamentally transformed over the past three decades. From 1985 to 2016, real GDP per capita increased by nearly 1200 percent.[55] Additionally, output has shifted away from agriculture and manufacturing and toward service sector activities.[56] These changes may have implications for the number of disputes on which to mediate as well as the willingness of parties to mediate those disputes.

A number of authors contend that China's rapid economic growth has resulted in a rising number of economic and social disputes.[57] Economic activities, such as the production and sale of goods and services, the development of new technologies, and the formation of business relationships, carry the potential for accidents and harms. Furthermore, changing norms and social structures arising out of economic development may alter the incidence of commonly mediated disputes involving family and neighbor relations.[58]

---

53 *See 1989 Regulations*, *supra* note 26, art. 15; Halegua, *supra* note 6, at 741.

54 Halegua, *supra* note 6, at 741.

55 NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

56 *See* Douglas Bujakowski & Joan Schmit, *Economic Structural Transformation and Litigation: Evidence from Chinese Provinces,* DEPAUL BUS. & COM. L.J. (forthcoming 2021).

57 *See, e.g.*, Di & Wu, *supra* note 11, at 234; Minzner, *supra* note 11, at 942; Zhuang & Chen, *supra* note 11, at 385.

58 *See, e.g.*, Di & Wu, *supra* note 11, at 234-35.

Though underlying disputes are a potentially important driver of people's mediation rates, these events can be difficult to measure. Eisenberg notes that most disputes which are not resolved through established channels are never reported or documented.[59] Fortunately for this study, two common types of people's mediation claims in China – divorces and automobile accidents – are almost always reported to authorities, even when people's mediation is not pursued. As such, we can test whether variations in divorce and automobile accident disputes may partially explain the people's mediation rate fluctuations observed. Accordingly, we include in our analysis the number of newly divorced persons per 10,000 people and the number of automobile accidents per 10,000 people. For other types of people's mediation claims, we do not possess direct measures of underlying disputes, and so we employ real gross domestic product (GDP) per capita as a proxy of disputes arising out of economic activity.

A well-established body of literature suggests that economic development involves not only changes in output, but also changes in the composition of GDP across economic sectors.[60]

This has certainly been true of China, a nation that has undergone extensive structural change in its transition to a market economy. As output has shifted away from agriculture and manufacturing and toward service sector activities, disputes associated with service activities have become more prevalent. Compared with more traditional cases of marriage and land ownership, today's disputes involve breaches of contract, intellectual property, product liability, insurance coverage, rental agreements, and construction noise, among others.[61]

New types of disputes may be less conducive to people's mediation, especially when people's mediators are not particularly knowledgeable about laws governing such conflicts. In fact, for much of China's history, it was unclear whether PMCs could legally process these new types of cases. According to the 1989 Regulations, PMCs are to mediate "disputes among the people."[62] As a result, many people's mediators felt that they could not legally accept disputes involving legal persons or social organizations.[63] It was not until the enactment of the PML in 2011 that people's mediators were explicitly permitted to mediate such conflicts.[64]

Given that parties may be less inclined to mediate new, more complex cases and/or that people's mediators may be less inclined to accept such

---

59 Theodore Eisenberg, *The Need for a National Civil Justice Survey of Incidence and Claiming Behavior*, 37 FORDHAM URB. L.J. 17, 35 (2010).

60 *See, e.g.*, COLIN CLARK, THE CONDITIONS OF ECONOMIC PROGRESS (1940); Allan G. B. Fisher, *Production, Primary, Secondary and Tertiary*, 15 ECON. RECORD 24 (1939) (providing foundational studies); JEAN FOURASTIÉ, DIE GROßE HOFFNUNG DES ZWANZIGSTEN JAHRHUNDERTS [THE GREAT HOPE OF THE TWENTIETH CENTURY] (1954).

61 *See* Halegua, *supra* note 6, at 718.

62 *Id.* n.13.

63 *Id.*

64 *Id.* at 719.

cases, we anticipate a negative relationship between economic structural changes and mediation rates. However, this relationship is subject to change as PMCs become better equipped to handle more complicated cases. To test this hypothesis, we include in our analysis the percentage of real GDP generated from agricultural activities and the percentage generated from manufacturing activities to proxy the complexity of disputes. Service sector is the hold-out. Additionally, consider whether coefficients on these variables change over time. Positive coefficients on agriculture and manufacturing in early years and negative coefficients on those variables in later years would be consistent with the hypothesis.[65]

D. *DEMOGRAPHIC SHIFTS*

The past three decades have witnessed dramatic changes in Chinese society. The proportion of the population living in urban areas has grown and educational attainment has risen. Li et al. (2016) note that China's urban population increased from 22.77 percent in 1985 to 55.95 percent in 2014 and that proportion of the labor force with a college degree increased from 1.52 percent to 14.60 percent over the same period.[66] These changes may have implications for the willingness of parties to seek people's mediation.

A number of authors speculate that in the event of a dispute, rural residents may be more inclined to seek people's mediation than urban residents.[67] In highly connected rural communities, people's mediators are more likely to have personal ties to disputants and to be seen as respected authority figures. Furthermore, social norms may punish litigious behavior, prompting individuals to seek other methods of dispute resolution, such as people's mediation. To account for the potential influence of urbanization, we include in our analysis the percentage of the population living in urban areas.

Like urbanization, educational attainment may shape mediation preferences.[68] As Halegua explains, "More educated people may be less afraid of going to court as well as have less respect for the authority and abilities of an old-aged mediator with less education and perhaps poorer understanding of

---

65 Our dataset does not include information that could be used to assess the complexity of mediated cases. As such, we use measures of economic development to proxy the complexity of cases.

66 Haizheng Li, Junzi He, Qinyi Liu, Barbara Fraumeni & Xiang Zheng, *Regional Distribution and Dynamics of Human Capital in China 1985-2014: Education, Urbanization, and Aging of the Population*, NBER Working Paper Series No. 22906, available at https://www.nber.org/system/files/working_papers/w22906/w22906.pdf, 30 & 41 (2016).

67 *See*, *e.g.*, Kevin C. Clark, *The Philosophical Underpinning and General Workings of Chinese Mediation Systems: What Lessons Can American Mediators Learn*, 2 PEPP. DISP. RESOL. L.J. 117, 129-32 (2002); Ethan Michelson, *Justice from Above or Below? Popular Strategies for Resolving Grievances in Rural China*, 193 CHINA Q. 43, 61 (2008).

68 *See*, *e.g.*, Di & Wu, *supra* note 11, at 237.

law than they themselves have."[69] Though reforms to the people's mediation system mandate higher levels of mediator education and training than ever before, nearly 40 percent of people's mediators still do not have a high school degree and even fewer have a legal professional background.[70] As such, highly educated individuals are expected to derive fewer benefits from mediation. We measure educational attainment using the number of students currently enrolled in universities per capita.

E. *THE ROLE OF LITIGATION*

When individuals choose not to seek people's mediation, they may instead opt to litigate the terms of a dispute. This "substitution effect" is perhaps the most cited explanation for the decline in people's mediation through the 1990s and early 2000s.[71] Data from the National Bureau of Statistics of China show that from 1990 to 2004, the number of civil lawsuits rose by nearly 34 percent, while the number of people's mediation claims fell by over 40 percent.[72] Nevertheless, a wider time interval from 1985 to 2016 reveals that annual litigation and people's mediation claims moved in the same direction more often than in opposite directions.[73] It may be that some disputants use people's mediation as an initial attempt to resolve a dispute, with an intent to litigate if people's mediation fails. In this case, people's mediation is sequential to litigation, rather than a substitute for litigation. As such, the substitution hypothesis warrants greater scrutiny. To test this hypothesis, we include in our analysis the number of civil lawsuits involving a plaintiff attorney per 10,000 people.[74]

Ideally, one would employ an alternative measure of litigation – the number of civil lawsuits per 10,000 people, regardless of whether a plaintiff attorney was used. Unfortunately, this variable is only available for 109 of the 594 province-years in our study, and thus is generally unusable.[75]

---

69 Halegua, *supra* note 6, at 719-20.

70 Yi Huang, *A Study on the People's Mediation System in China: Compared with the Alternative Dispute Resolution (ADR) System in Japan*, 3 FRONTIERS OF LEGAL RSCH. 12, 14 (2015).

71 *See*, *e.g.*, Shahla Ali, *The Jurisprudence of Responsive Mediation: An Empirical Examination of Chinese Peoples Mediation in Action*, 45 J. LEG. PLUR. UNOFF. L. 227, 236 (2013); Di & Wu, *supra* note 11, at 229; Halegua, *supra* note 6, at 719; Minzner, *supra* note 11, at 943; Wu, *supra* note 11, at 27.

72 NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

73 *Id.*

74 Complaints/petitions (xinfang) may also serve as a complement or substitute to people's mediation; however, we do not possess data on these occurrences. *See*, *e.g.*, Carl F. Minzner, *Xinfang: An Alternative to Formal Chinese Legal Institutions*, 42 STAN. J. INT'L L. 103, 162 (2006).

75 The availability of litigation involving a plaintiff attorney but not total litigation stems from the fact that province-level litigation data are derived from two sources: courts and law firms. Court data are used to construct unconditional litigation rates, while law firm data are used to construct litigation rates involving a plaintiff attorney. Court data are available for 109 of the 594 province-years in our study,

National data from 1985 to 2016 reveal that the correlation between litigation involving a plaintiff attorney and total litigation is 98.4 percent.[76] Additionally, provincial data in which both data sources are available reveal a correlation of 90.0 percent.[77] Given that the two litigation measures exhibit highly similar patterns across provinces and time, we anticipate that litigation involving a plaintiff attorney serves as an adequate proxy of total litigation.

while law firm data are available for all 594 province-years. Thus, use of the former would dramatically reduce our sample size.

[76] This correlation compares counts of total lawsuits with counts of lawsuits involving a plaintiff attorney at the national level. Both counts are measured from 1985 to 2016. Thus, we assess this correlation using 32 values for each variable.

[77] This correlation compares counts of total lawsuits with counts of lawsuits involving a plaintiff attorney at the province level. There are 109 province-years for which both counts are available. Thus, we assess this correlation using 109 values for each variable.

F. *HYPOTHESES*

We use our five categories of drivers to define our main hypotheses. Namely,

H1. People's mediation rates are positively related to mediation system reforms.

H2. People's mediation rates are positively related to access to people's mediation services; however, the relationship between people's mediation and unemployment is ambiguous.

H3. People's mediation rates are positively related to numbers of underlying disputes and negatively related to the complexity of those disputes in early years.

H4. People's mediation rates are negatively related to general demographic trends, including increases in urbanization and education.

H5. People's mediation rates are related to litigation rates, however, the direction of this relationship is ambiguous.

III. DATA AND METHODOLOGY

A. *DATA*

To test our hypotheses, we employ 32 years of province-level data (1985-2016) from 24 of China's 31 provinces, municipalities, and autonomous regions (collectively referred to as provinces in this paper).[78] These data are taken from the China Statistical Yearbooks Database (CSYD), which includes provincial statistical yearbooks published annually by the National Bureau of Statistics of China.[79] The CSYD does not contain mediation data

[78] The Constitution of the People's Republic of China places municipalities (Beijing, Chongqing, Shanghai, and Tianjin) and autonomous regions (Guangxi, Inner Mongolia, Ningxia, Tibet, and Xinjiang) on the same administrative level as provinces. Although these jurisdictions are unique from provinces and from each other in various ways, we account for that variation within our analysis.

[79] NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

for the seven provinces not included in our analysis.[80] Subsequent searches for Chinese mediation data suggest that data for the seven missing provinces are not publicly available.

Mediation data in the CSYD reflect the number of disputes for which parties sought people's mediation services.[81] These counts include both successful and unsuccessful instances of people's mediation and exclude judicial mediation, arbitral mediation, and other matters handled by courts.[82] The CSYD also contains economic, demographic, and legal information, which we use to construct our other variables.[83] Table 1 provides definitions for these variables. All variables are measured at the end of each year and at the province level. Mediation rate is our dependent variable. All other variables are explanatory variables.

Figure 2 shows people's mediation rates for select provinces from 1985 to 2016. For clarity, the figure includes only eight of the 24 provinces for which we have data. The eight provinces shown were selected based on two criteria: (1) they include those provinces with the lowest and highest average mediation rate over the timespan and (2) they are generally representative of the 16 provinces not shown in that excluded provinces display mediation patterns similar to one or more provinces in the graph. This figure represents the large degree to which people's mediation rates vary across Chinese provinces. Chongqing has the highest average people's mediation rate of all provinces in our study with over 104 cases per 10,000 people, while Guangdong has the lowest average mediation rate of those provinces with fewer than 26 cases per 10,000 people.[84] We account for this variation in our analysis and leverage it in testing our hypotheses.

Figure 2 also reveals a general pattern in the evolution of people's mediation rates over the study period, one that has already been observed on a national scale. People's mediation rates appear to decline throughout the 1990s and early 2000s and then increase beginning in the mid-2000s. In fact, for all provinces in our study except Shaanxi, the lowest people's mediation rate for that province occurs between 2001 and 2008.[85] Yet, there is substantial variation in mediation rate resurgence following low points in the mid-2000s. During the period of regrowth, Heilongjiang and Anhui boast the greatest relative gains in people's mediation rates of 350 percent and 334

---

80 The seven missing provinces include Gansu, Guangxi, Hebei, Inner Mongolia, Shandong, Tibet, and Yunnan.

81 NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

82 If parties pursue other means of dispute resolution after unsuccessful people's mediation, the instance of attempted mediation is included within our counts.

83 NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

84 Chongqing's people's mediation rates begin in 1997, rather than 1985, because the province was created in that year. It is the newest of all of China's provinces.

85 Shaanxi's lowest people's mediation rate of 33.1 cases per 10,000 people occurs in 2015. This rate is slightly below Shaanxi's 2006, 2007, and 2008 rates of 35.5, 34.0, and 34.2 cases per 10,000 people, respectively.

percent, while Shaanxi witnesses the lowest relative gain of less than 2 percent.

Summary statistics are reported in Table 2. To account for skewness, all variables except those that are percentages or indices are expressed in log form. Most variables have small standard deviations relative to their mean, indicating that these variables are rather symmetric. The average people's mediation rate is 51.7 cases per 10,000 people, the average number of people's mediators is 53.3 per 10,000 people, and the average number of PMCs is 7.2 per 10,000 people. These numbers are comparable to those observed on a national scale, where average people's mediation rates, numbers of people's mediators, and numbers of PMCs are 52.1, 54.2, and 7.4 per 10,000 people, respectively. Thus, while mediation data are only available for 24 of China's 31 provinces, our sample appears to be representative of China's mediation experience generally.

Another important item to note is that for many of the variables, variation between provinces is close to or greater than variation within provinces. This result illustrates the vast disparities that exist within Chinese society, which our study design exploits. Consider that our analysis spans over three decades in which China underwent significant economic and demographic change, and yet, variation between provinces is comparable to that across time.

B. *ECONOMETRIC MODEL*

Given that we are working with a longitudinal dataset, we can exploit both cross sectional and time series variation to test relationships between people's mediation rates and the other variables in our analysis. In purely cross-sectional studies, variation in people's mediation rates may arise from sources beyond the particular variables studied, such as differences in cultural norms and attitudes toward mediation across jurisdictions. A longitudinal design addresses this limitation by examining dynamics across time, holding constant jurisdiction-specific attributes.

Like many longitudinal studies, we begin by specifying a traditional fixed effects model that accounts for province-level heterogeneity. This model is shown in Equation (1):

$$y_{it} = \alpha_i + \boldsymbol{x}_{\boldsymbol{it}}'\boldsymbol{\beta} + \delta_1 \mathrm{Year}_t + \delta_2 \mathrm{Year}_t^2 + \varepsilon_{it} \quad (1)$$

where the natural log of the people's mediation rate in province i in year t (y_{it}) is a function of a fixed country-specific intercept (α_i), a vector of explanatory variables ($\boldsymbol{x}_{\boldsymbol{it}}'$), a quadratic time trend (defined by Year_t and Year_t^2), and an error term (ε_{it}).

Using logs, rather than levels, of the dependent variable benefits our regression model in several ways. First, logging allows for a general, non-

linear relationship among our variables. This non-linear relation is apparent when transforming back to un-logged mediation rates: $Mediation\ Rate_{it} = e^{\alpha_i} e^{\mathbf{x}'_{it}\boldsymbol{\beta}} e^{\delta_1 Year_t} e^{\delta_2 Year_t^2} e^{\varepsilon_{it}}$. Second, logging makes residuals more normally distributed and reduces heteroskedasticity. Third, logging addresses the fact that mediation rate is a limited dependent variable, in that it is bounded from below at zero. After taking logs, the result is unbounded, avoiding the need for limited dependent variable methods.

The inclusion of province fixed effects controls for unobserved province characteristics that are time-invariant.[86] Figure 2 reveals a high degree of variation in people's mediation rates across provinces. This variation may partially arise from factors beyond those tested in our study. As such, it is important that we hold these unobserved characteristics constant in our analysis through the use of province fixed effects. In doing so, we isolate the impact of independent variables on people's mediation rates using variation within provinces over time.

In addition to province fixed effects, we include a quadratic time trend in our analysis.[87] Figures 1 and 2 reveal that people's mediation rates declined through the 1990s and early 2000s and then increased beginning in the mid-2000s. If this temporal trend is not accounted for and regressors exhibit similar (or opposite) trends, we may observe strong correlations between mediation rates and regressors, even when no meaningful relationship exists.

A final consideration surrounding our regression framework relates to the potential endogeneity of certain explanatory variables. Specifically, Mediators, PMCs, and Litigation Rates may each influence mediation rates, yet the reverse may also be true. For example, when people's mediation rates are high, more people's mediators and PMCs may be necessary to meet demand and disputes may be resolved without the need for litigation.

We anticipate the potential for reverse causality to be somewhat mitigated by the short-run inelasticity of people's mediators and PMCs to changes in mediation rates.[88] Nevertheless, endogeneity concerns may persist, especially with respect to litigation rates. To reduce these concerns, we use one-year lagged versions of Mediators, PMCs, and Litigation Rate in our analysis. Lagging ensures that mediator, PMC, and litigation rate values precede mediation rate values in time, lessening the potential for reverse

[86] The Hausman specification test indicates that province fixed effects are preferred to random effects.

[87] Subsequent analyses indicate that a quadratic relationship is sufficient to account for the influence of time. Higher order polynomials do not improve model fit.

[88] Mediators serve three-year terms and PMCs are most commonly established by village committees or resident committees through a bureaucratic process.

causality.[89] We investigate higher order lags and find that results are similar when lagging by two or more years.

IV. RESULTS

Table 3 shows results from the estimation of Equation 1. Model 1 includes all years for which we have data (1985-2016), while Models 2 and 3 partition our study period into two segments: the years of national-level mediation decline (1985-2004) and the years of resurgence (2005-2016).[90] Partitioning the study period allows explanatory variable coefficients to differ across time periods. This flexibility may be important in assessing factors related to the decline and resurgence of people's mediation. In Model 1, the regressors explain over 46 percent of the variation in people's mediation rates. In Models 2 and 3, the regressors explain over 67 and 61 percent of that variation, respectively.

The inclusion of a quadratic time trend appears to be warranted when examining all sample years. In Model 1, the coefficient on Year is negative and significant, while the coefficient on $Year^2$ is positive and significant. This result is consistent with the U-shaped pattern in people's mediation rates over our study period. In Models 2 and 3, coefficients on Year are not significantly different from zero and coefficients on $Year^2$ are significant at only the 10 percent level. Thus, a quadratic time trend does not appear to be necessary when separately examining the periods of mediation decline and resurgence.

Given that mediation rates are expressed in log form, interpretation of explanatory variable coefficients requires transforming to un-logged values. For explanatory variables that are expressed in logs (Mediators, PMCs, Divorces, Auto Accidents, GDP, Education, and Litigation Rate), a one-percent increase in that variable is associated with a $\hat{\beta}$ percent increase in mediation rates, where $\hat{\beta}$ is the variable's estimated coefficient. For explanatory variables that are not expressed in logs (1989 Regulations, 2002 Reforms, 2011 PML, Unemployment, Agriculture, and Manufacturing), a one-unit increase in that variable is associated with a $100(e^{\hat{\beta}} - 1)$ percent increase in mediation rates.

In examining estimates on the mediation reform variables, we find strong evidence that people's mediation rates increased following the 2011 PML, lending support to Hypothesis 1. In Models 1 and 3, coefficients on

[89] Our dataset does not include variables that could serve as valid instruments for mediators, PMCs, or litigation rates; hence, we cannot employ an instrumental variables approach to address possible reverse causation.

[90] Subsequent analyses reveal that altering the partition break point by plus or minus two years has virtually no effect on our results. Similarly, restricting the period of decline to years 1990-2004 and the period of resurgence to years 2005-2013 does not meaningfully alter our results.

the 2011 PML are positive and statistically different from zero at the one percent significance level.[91] We do not find evidence that the 1989 Regulations or the 2002 Reforms altered subsequent mediation rates. Coefficients on these variables are not statistically different from zero in any model.[92] Though we can only speculate as to why the 2011 reforms appear more influential than those in 1989 and 2002, a distinction in their provisions is apparent: the 2011 PML[93] codifies into law the enforceability of mediated resolutions, while the 1989 Regulations[94] and 2002 Reforms[95] do not. If enforceability is particularly important to those considering people's mediation, this preference could explain the results we observe.

Direct measures of underlying disputes – Divorces and Auto Accidents – only appear to be associated with people's mediation rates from 1985-2004. Here, the coefficient on Divorces is positive and significant at the five percent level and the coefficient on Auto Accidents is positive and significant at the one percent level. Coefficients on these variables are not statistically different from zero in the full study period or in the 2005-2016 subset.

We also find that measures of economic development and educational attainment have differential impacts on mediation rates in the periods of decline and resurgence. From 1985-2004, province-years with lower GDP per capita, lower levels of economic structural change (i.e. larger agricultural and manufacturing sectors relative to the service sector), and lower levels of educational attainment tended to have higher mediation rates, all else equal. Yet, from 2005-2016, this tendency largely reversed, with GDP per capita and economic structural change becoming positively related to mediation rates and educational attainment displaying no statistical relationship with mediation rates. On balance, these effects appear to somewhat negate each other. Over the full study period, coefficients on GDP and Education are not statistically different from zero and the coefficients on Agriculture and Manufacturing, while positive and significant, are smaller in magnitude than those in the 1985-2004 subset.

Coefficients on Divorces, Auto Accidents, GDP, Agriculture, Manufacturing, and Education reveal a striking shift between 1985-2004 and 2005-2016. The direction and/or statistical significance of all of these coefficients changed over our study period. One possible explanation for this change stems from changes in the nature of people's mediation practice. In the early and mid 2000s, China began to standardize people's mediation procedures,

---

91 These effects are measured after controlling for a quadratic time trend and thus, are not simply reflective of an underlying temporal process.

92 The 2011 PML is not included in Model 2 because that model only includes years 1985-2004, in which the 2011 PML had not yet been enacted. Similarly, the 1989 Regulations and 2002 Reforms are not included in Model 3 because that model only includes years 2005-2016, in which the 1989 Regulations and 2002 Reforms had not yet been enacted.

93 People's Mediation Law, *supra* note 5.

94 *1989 Regulations*, *supra* note 26.

95 *Hearing Civil Cases*, *supra* note 33; *Provisions Concerning Mediation*, *supra* note 34; *Enhancing Mediation*, *supra* note 35.

increase the legal weight of mediated agreements, and integrate people's mediation with courts and administrative bodies. These changes may have had the effect of equipping PMCs to handle more complex disputes arising out of economic and social development. In this case, we might expect more traditional types of cases, such as divorces and auto accidents, to play less of a role in the subsequent period and economic and social development to have a positive rather than negative influence on numbers of mediation claims. This is indeed what we observe.

Rising levels of urbanization appear to be associated with fewer people's mediation claims across all time periods. Coefficients on Urbanization are negative and statistically significant in all three models. Urban residents may be less likely to have personal ties with people's mediators or other parties to a dispute and may face lower social costs to litigating. As such, they are expected to derive fewer benefits from people's mediation.

Perhaps even more interesting than what our analysis uncovers is what it does not find.

Coefficients on mediators, PMCs, and unemployment are not significantly different from zero in any model. Recall that these variables are intended to represent the ability of parties to access people's mediation services. As such, we do not find evidence that mediation access influences people's mediation rates. One possibility is that China's people's mediation infrastructure is so expansive relative to the number of people's mediation claims it processes that utilization is quite insensitive to changes in the supply of people's mediation services. Indeed, over the entire 32-year timespan of this study, the ratio of people's mediation claims to people's mediators never exceeded 2.4 to one and the ratio of people's mediation claims to PMCs never exceeded 117 to 1.[96]

Interestingly, coefficients on Litigation Rate are also not statistically different from zero in any of the three models. This finding runs counter to the widely held belief that substitution of litigation for people's mediation produced the dramatic decline in mediation rates observed from 1990-2004. A substitution effect would imply a negative coefficient on Litigation Rate. Similarly, we do not find support for the notion that people's mediation is sequential to litigation, potentially serving as an initial attempt to resolve a dispute, with an intent to litigate if people's mediation fails. A complementary relationship between mediation and litigation would imply a positive coefficient on Litigation Rate.

In summary, we find strong support for hypotheses one, three, and four – that people's mediation rates are positively related to mediation system reforms and numbers of underlying disputes and negatively related to the complexity of disputes in early years and to general demographic trends, including increases in urbanization and education. We do not find support for hypothesis two, that people's mediation rates are related to access to mediation

96 NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

services. Similarly, we do not find support for hypothesis five, that people's mediation rates are related to litigation rates, either as a complement or as a substitute.

To investigate the robustness of our results, we re-estimate all models holding out each possible regressor. Doing so results in 14 alternative specifications for each model, corresponding with the 14 regressors included in our main analysis. We then report the smallest and largest coefficients from the 14 specifications along with their statistical significance. The results of this analysis are provided in Tables 4 and illustrate the stability of our regressors. Specifically, we find that even when variables are excluded, coefficients on 2011 PML, Divorces, Auto Accidents, GDP, Urbanization, and Education retain their direction and statistical significance across all models. Agriculture and Manufacturing retain their direction and statistical significance in Models 1 and 2, but in Model 3, certain specifications result in a loss of statistical significance.

V. CONCLUDING REMARKS

The decline and resurgence of people's mediation in China is well documented, yet until now, proposed explanations for people's mediation trends have not been tested. In the current analysis, we review theories of mediation development and utilize 32 years of province-level people's mediation rate data to test those theories.

Perhaps the most cited explanation for the decline of people's mediation rates from 1990 to the mid-2000s is China's turn toward litigation. From 1990 to 2004, the number of civil lawsuits rose by nearly 34 percent, while the number of people's mediation claims fell by over 40 percent.[97] Nevertheless, in examining a wider time interval from 1985 to 2016, we find that annual litigation and mediation claims moved in the same direction more often than in opposite directions.

In an empirical analysis of people's mediation rates, we do not find evidence of a substitution effect of litigation for mediation. This result holds regardless of whether one examines the period of mediation decline or our entire study period. Instead, we observe that declining people's mediation rates are associated with rising levels of educational attainment and urbanization as well as changes in economic output and composition across economic sectors. Highly educated, urban individuals likely derive fewer benefits from people's mediation, while structural changes in the Chinese economy may yield more complicated disputes that are less suitable to people's mediation.

Since the bottom of the decline in 2004, the annual number of people's mediation claims has more than doubled. To explain this resurgence, scholars generally point to legislative and procedural reforms that have made

[97] NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

people's mediation more attractive to disputants and to increasing numbers of economic and social disputes arising out of economic development. Our results support these hypotheses, suggesting that the 2011 People's Mediation Law, GDP per capita, and economic structural changes are positively associated with people's mediation rates in the period of mediation resurgence.

The findings of this study may be useful in setting public policy and in preparing for people's mediation claims. Specifically, information about factors associated with people's mediation rates could be used to better understand and more accurately predict mediation shifts. While our analysis provides evidence of a connection between several variables and people's mediation rates, further tests are necessary to understand the mechanisms which underlie those relationships. For example, one might investigate whether increases in the enforceability of mediated resolutions are responsible for the positive correlation between legislative changes and mediation rates by surveying people's mediation participants. Unfortunately, our dataset does not provide a sufficient level of granularity to conduct such tests. More fully identifying underlying mechanisms remains a promising area for future research.

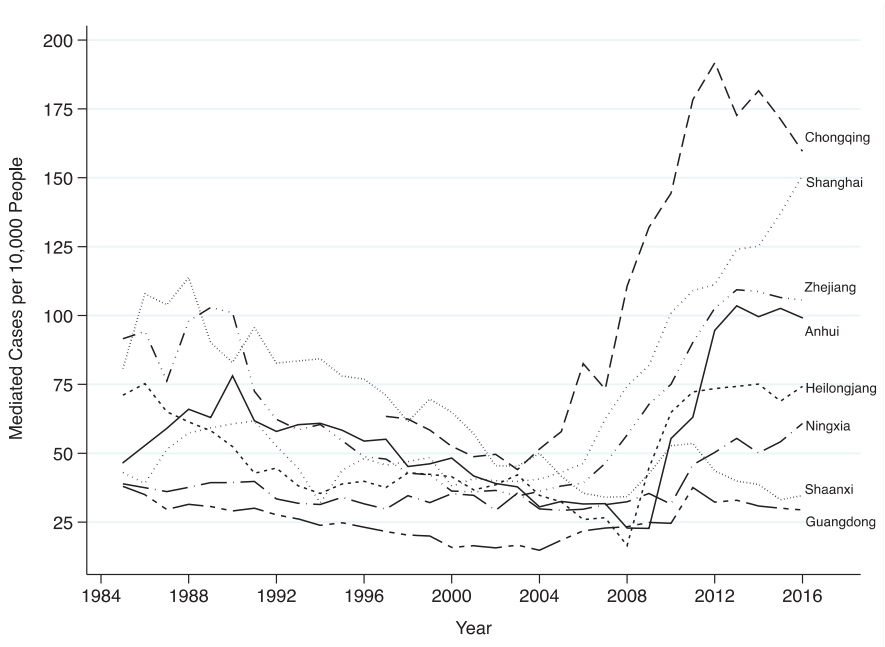
The current analysis is limited somewhat by the availability of Chinese data. Ideally, we would incorporate people's mediation rate information from all of China's 31 provinces, municipalities, and autonomous regions for all 32 years of our study. Additionally, more detailed information regarding the people's mediation process, including the types of disputes mediated, mediation times, attributes of mediators, the probability of successful mediation, the nature of mediated agreements, and alternatives to mediation could provide additional insight into the development of people's mediation rates. Gathering such data would allow for a richer understanding of the factors associated with people's mediation in China.

Figure 1: Annual people's mediation rates in China from 1985 to 2016[98]

The graph shows annual people's mediation rates in China from 1985 to 2016, defined as the number of people's mediation claims per 10,000 people. Both people's mediation claims and population counts are measured at the end of each year.

98 NAT'L BUREAU OF STAT. OF CHINA, *supra* note 1.

Figure 2: People's mediation rates for select provinces



The graph shows people's mediation rates from 1985 to 2016 for select provinces. For clarity, the figure includes only eight of the 24 provinces for which we have data. The eight provinces shown were selected based on two criteria: (1) they include those provinces with the lowest and highest average people's mediation rate over the timespan and (2) they are generally representative of the 16 provinces not shown in that excluded provinces display mediation patterns similar to one or more provinces in the graph. Missing values for Chongqing before 1997 reflect the fact that Chongqing was created in 1997.

Table 1: Variable Definitions

The table provides definitions for the variables included in our analysis. The sample period is 1985-2016. All variables are measured at the end of each year and at the province-level.

Variable	Definition
Mediation Rate	Natural log of the number of people's mediation cases per 10,000 people
1989 Regulations	Indicator for the enactment of the 1989 Regulations (1 in years 1989-2016; 0 otherwise)
2002 Reforms	Indicator for the enactment of the 2002 Reforms (1 in years 2002-2016; 0 otherwise)
2011 PML	Indicator for the enactment of the 2011 People's Mediation Law (1 in years 2011-2016; 0 otherwise)
Mediators	Natural log of the number of registered people's mediators per 10,000 people
PMCs	Natural log of the number of people's mediation committees per 10,000 people
Unemployment	Percentage of the labor force that is jobless
Divorces	Natural log of the number of newly divorced persons per 10,000 people
Auto Accidents	Natural log of the number of automobile accidents per 10,000 people
GDP	Natural log of real gross domestic product per capita where the base year is 2016
Agriculture	Percentage of GDP associated with agricultural activities
Manufacturing	Percentage of GDP associated with manufacturing activities
Urbanization	Percentage of the population living in urban areas
Education	Natural log of the number of students currently enrolled in universities per 10,000 people
Litigation Rate	Natural log of the number of civil lawsuits involving a plaintiff attorney per 10,000 people

Table 2: Summary Statistics

			Standard Deviation				
Variable	Obs.	Mean	Overall	Between	Within	Min.	Max.
Mediation Rate	594	3.945	0.431	0.296	0.332	2.416	5.256
1989 Regulations	594	0.949	0.219	0.046	0.214	0.000	1.000
2002 Reforms	594	0.549	0.498	0.156	0.483	0.000	1.000
2011 PML	594	0.217	0.413	0.139	0.402	0.000	1.000
Mediators	594	3.976	0.515	0.263	0.447	2.424	5.349
PMCs	594	1.974	0.388	0.296	0.264	0.970	2.762
Unemployment	594	3.295	1.063	0.733	0.809	0.300	7.400
Divorces	594	2.605	0.659	0.458	0.479	1.105	3.908
Auto Accidents	594	0.937	0.772	0.489	0.580	-1.227	3.456
GDP	594	9.649	0.927	0.535	0.780	7.844	11.665
Agriculture	594	15.681	9.261	6.934	6.475	0.388	41.561
Manufacturing	594	44.789	7.437	6.206	5.417	21.306	68.460
Urbanization	594	44.604	17.736	15.188	9.895	6.842	89.607
Education	594	4.256	1.011	0.577	0.862	2.048	5.876
Litigation Rate	594	1.801	0.845	0.494	0.683	-0.785	3.937

The table shows summary statistics for all variables included in our analysis. The sample period is 1985-2016. All variables are measured at the end of each year and at the province-level. See Table 1 for variable definitions. Mediation Rate, Mediators, PMCs, Divorces, Auto Accidents, GDP, Education, and Litigation Rate are expressed in natural logs. Mediators, PMCs, and Litigation Rate are lagged by one year. "Standard deviation between" reflects variation across provinces, while "standard deviation within" reflects variation within provinces over time.

Table 3: Regression Results

	Model 1			Model 2			Model 3		
	Years 1985-2016			Years 1985-2004			Years 2005-2016		
	Est.	*S.E.*		*Est.*	*S.E.*		*Est.*	*S.E.*	
1989 Reg.	0.061	0.077		0.076	0.059				
2002 Reforms	0.013	0.064		0.057	0.045				
2011 PML	0.274	0.061	***				0.295	0.061	***
Mediators	0.005	0.043		-0.039	0.035		0.141	0.107	
PMCs	-0.135	0.096		0.151	0.095		-0.114	0.299	
Unemployment	-0.028	0.018		-0.013	0.013		-0.009	0.064	
Divorces	0.124	0.084		0.163	0.078	**	0.251	0.172	
Auto Accidents	-0.031	0.026		0.135	0.025	***	-0.048	0.067	
GDP	-0.069	0.139		-0.265	0.125	**	1.171	0.401	***
Agriculture	0.016	0.006	***	0.019	0.006	***	-0.028	0.014	**
Manufacturing	0.006	0.003	*	0.013	0.004	***	-0.016	0.008	**
Urbanization	-0.011	0.004	***	-0.015	0.004	***	-0.061	0.013	***
Education	-0.110	0.074		-0.262	0.082	***	-0.150	0.230	
Litigation Rate	0.033	0.034		0.024	0.033		-0.095	0.073	
Year	-0.044	0.018	**	-0.022	0.022		0.162	0.128	
$Year^2$	0.002	0.000	***	0.002	0.001	*	-0.004	0.002	*
Provinces		24			24			24	
Observations		594			334			260	
Province FE		Yes			Yes			Yes	
R^2		0.46			0.67			0.61	

The table shows estimates from three regressions: (1) a model that includes all years for which we have data (1985-2016), (2) a model that primarily includes the years of national-level mediation decline (1985-2004), and (3) a model that primarily includes the years of national-level mediation resurgence (2005-2016). In all models, the dependent variable is Mediation Rate, defined as the natural log of the number of people's mediation cases per 10,000 people. All models include a constant term, province fixed-effects, and economic and demographic variables. See Table 1 for variable definitions. Mediators, PMCs, and Litigation Rate are lagged by one year. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively, using heteroskedasticity robust standard errors.

Table 4: Robustness Tests

	Model 1: Years 1985-2016						Model 2: Years 1985-2004						Model 3: Years 2005-2016					
	Min. Estimates			Max. Estimates			Min. Estimates			Max. Estimates			Min. Estimates			Max. Estimates		
	Est.	*S.E.*		*Est.*	*S.E.*		*Est.*	*S.E.*		*Est.*	*S.E.*		*Est.*	*S.E.*		*Est.*	*S.E.*	
1989 Reg.	0.049	0.076		0.101	0.072		0.058	0.059		0.108	0.056	*						
2002 Reforms	-0.042	0.053		0.034	0.062		0.012	0.046		0.075	0.043	*						
2011 PML	0.264	0.060	***	0.298	0.058	***							0.256	0.063	***	0.304	0.060	***
Mediators	-0.018	0.040		0.018	0.042		-0.049	0.035		-0.010	0.034		0.083	0.109		0.212	0.111	*
PMCs	-0.136	0.092		-0.111	0.097		0.091	0.094		0.176	0.093	*	-0.177	0.299		-0.101	0.300	
Unemployment	-0.033	0.019	*	-0.019	0.018		-0.020	0.013		0.000	0.014		-0.057	0.066		0.006	0.063	
Divorces	0.087	0.080		0.148	0.079	*	0.128	0.071	*	0.188	0.078	**	0.201	0.180		0.323	0.173	*
Auto Accidents	-0.040	0.024		-0.015	0.026		0.109	0.024	***	0.138	0.025	***	-0.073	0.070		0.025	0.066	
GDP	-0.136	0.138		0.032	0.122		-0.317	0.118	***	-0.265	0.125	**	0.908	0.399	**	1.279	0.403	***
Agriculture	0.012	0.006	**	0.019	0.006	***	0.011	0.005	**	0.022	0.006	***	-0.029	0.015	**	-0.006	0.014	
Manufacturing	0.003	0.003		0.007	0.003	**	0.008	0.004	**	0.014	0.004	***	-0.017	0.008	**	-0.006	0.007	
Urbanization	-0.012	0.004	***	-0.009	0.004	***	-0.017	0.004	***	-0.014	0.004	***	-0.066	0.012	***	-0.047	0.012	***
Education	-0.198	0.073	***	-0.067	0.071		-0.306	0.083	***	-0.154	0.084	**	-0.374	0.236		0.090	0.217	
Litigation Rate	0.031	0.033		0.061	0.034	*	0.015	0.033		0.050	0.034		-0.107	0.072		-0.072	0.070	
Year	-0.065	0.016	***	-0.036	0.015	**	-0.044	0.023	**	0.004	0.021		0.138	0.128		0.387	0.116	***
Year^2	0.002	0.000	***	0.003	0.000	***	0.000	0.001		0.002	0.001	**	-0.007	0.002	***	-0.003	0.002	
Provinces		24			24			24			24			24			24	
Observations		594			594			334			334			260			260	
Province FE		Yes			Yes			Yes			Yes			Yes			Yes	
R^2		0.44			0.46			0.64			0.67			0.56			0.61	

The table shows results from an analysis in which each regressor is individually omitted to determine how other estimates are affected. We produce 14 alternative specifications for each model, corresponding with the 14 regressors included in our main analysis, and report the smallest and largest estimates for each variable, along with their statistical significance. In all models, the dependent variable is Mediation Rate, defined as the natural log of the number of people's mediation cases per 10,000 people. All models include a constant term, province fixed-effects, and economic and demographic variables. See Table 1 for variable definitions. Mediators, PMCs, and Litigation Rate are lagged by one year.

***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively, using heteroskedasticity robust standard errors.

File: Grinapell 17.1 To Publish_Edited Created on: 6/14/2022 8:40:00 PM Last Printed: 6/14/2022 9:25:00 PM

WHAT DRIVES THE USE OF DUAL-CLASS STRUCTURES IN TECHNOLOGY IPOS?

Adi Grinapell[1]

INTRODUCTION

In 2004, Google surprised the market by going public with a dual-class stock structure, in which Class B shares, held by management and existing shareholders, had ten times the voting power of Class A shares offered to the public.[2] Though this stock structure was legally permissible, it had been widely viewed by the media, academia, and the institutional investor community as suitable only for certain industries, like the media industry for instance, where it allows firms to concentrate on their core, long-term interest in serious news coverage, despite fluctuations in quarterly results. Under a single-class structure, where outside shareholders hold a majority of votes, concerns about fluctuations in quarterly results might force editors of media firms to settle for the quality of news coverage in order to cater to shareholders demands. For a technology firm like Google however, dual-class structures have been considered an unnecessary defensive measure rather than a core business need, isolating holders of high-voting-power shares from market discipline, while potentially increasing agency costs. And yet, today, dual-class structured technology-based IPOs account for an increasingly significant portion of all recent IPOs— a trend that should be of growing interest to the debate over the desirability of dual-class stock structure more broadly. This paper contributes to this debate by providing important new evidence on the factors influencing technology-based firms in their decision to adopt this structure.

Despite various efforts made by Google's founders—including an extraordinary letter to potential new shareholders justifying their decision[3]—

---

[1] JSD Graduate, the University of Chicago Law School. I am gratefully indebted to my advisors, Dhammika Dharmapala and Anthony Casey, for numerous conversations and invaluable advice. For helpful comments and discussions, I thank Omri Ben Shahar, Lisa Bernstein, Zohar Goshen, William Hubbard, Adi Leibovitch, Kobi Kastiel, Ariel Porat, Guy Rolnik, Ziv Schwartz and Hanock Spitzer. I would also like to thank Martin Kenney and Donald Patton for generously sharing their database on emerging growth initial offerings from 1990 though 2015, and their database on management and directors' roles in initial public offerings between 1990 and 2010. Any errors are my own.

[2] Google Inc., Registration Statement (Amend. No. 9 to Form S-1) (Aug 18, 2004).

[3] *See*, for instance, Google Inc., *supra* note 1,where Google founders, Larry Page & Sergey Brin, acknowledge that their registration for dual-stock structure required future public shareholders to place a

the institutional investor community continued to oppose Google's move, insisting on a substantial price discount to dual-class shares.[4] The lower value assigned to dual-class shares would presumably reflect the fact that dual-class stock structure facilitates extraction of private benefits of control, solidifying and empowering management at the expense of shareholders. If institutional investors indeed place a lower value on technology stocks with dual-class structure, a typical "market for lemons" should theoretically exist: investors cannot distinguish between firms that wish to maximize value ("good" firms) and firms that plan to extract private benefits at the expense of public shareholders ("bad" firms). Thus, investors should discount all dual-class firms, driving "good" value-enhancing firms out of the dual-class market. In practice though, despite investors' outcry, and contrary to theoretical predictions, many prominent technology-based firms, including Facebook, LinkedIn, Snap, and more recently Dropbox, and Lyft, have followed Google in making the same dual-class choice.[5] Thus, the dual-class structure seen as so unusual in 2004, has come to be regarded as a kind of industry standard.[6] This paper explores whether investors can actually observe firms' intentions and distinguish between "good" firms, seeking to employ dual-

---

potentially risky, long-term bet on Google that would enable the company to retain many of the positive aspects of being private so that that everyone —new *public* shareholders included— would benefit.

[4] *See*, Simon London, U.S. Fund Criticizes Google's IPO Structure, FIN. TIMES (May 4, 2004), http://www.nbcnews.com/id/4900355/ns/business-financial_times/t/us-fund-criti-%20cizes-googles-ipostructure/%20#.Xctr9m5FxPZ%20[https:/%20/%20perma.cc/%207RHS-7SQP] (quoting Peter Chapman, senior vice-president at the Teachers Insurance Annuity Association of America ("TIAA") remark that Google's shares should be priced at "a substantial discount to reflect the fact that the founders will retain control of the firm following its IPO," as "this structure effectively disenfranchises outside shareholders").

[5] An especially extreme case is Snap Inc., which listed nonvoting stock to the public in 2017 (2017). See Snap Inc., Prospectus, (Form 424B) (Mar. 1, 2017), http://www.sec.gov/Archives/edgar/data/1564408/000119312517068848/d270216d424b4.htm.

[6] *See* Jeffrey Green & Ari Levi, *Zuckerberg Grip Becomes New Normal in Silicon Valley*, BLOOMBERG BUSINESS (May 7, 2012), https://www.bloomberg.com/news/articles/2012-05-07/zuckerberg-stock-gripbecomes-new-normal-in-silicon-valley-tech (quoting Lise Buyer, principal at Class V Group in California: "When Google did it, there was tremendous pushback from the banks…. Today the bankers are often the ones suggesting it"); Steven Davidoff Solomon, Shareholders Vote with Their Dollars to Have Less of a Say, N.Y. Times (Nov. 4, 2015), https://www.nytimes.com/2015/11/05/business/dealbook/shareholders-votewith-their-dollars-to-have-less-of-a-say.html (noting that more than 13.5 percent of the 133 companies listing shares on United States exchanges in 2015 have set up a dual-class structure, according to the data provider Dealogic. About half the companies choosing the structure were in the technology industry. That compares with 12 percent in 2014 and just 1 percent in 2005); Alice Gomstyn, Supervoters, Stocks, and Silicon Valley: What Investors Should Know About Dual-Class Voting Structures: Tech companies are joining the trend of giving more voting rights to selected shareholders, The Alert Investor (Dec. 5, 2015), https://www.foxbusiness.com/markets/supervoters-stocks-and-silicon-valley-what-investors-should-knowabout-dual-class-voting-structur (mentioning that a growing number of U.S. firms have adopted the dualclass structure: Between 2013 and late 2015, 98 companies newly listed on U.S. ex changes had dual-class IPOs, compared to 59 between 2010 and 2012, according to data from information provider Dealogic).

File: Grinapell 17.1 To Publish_Edited Created on: 6/14/2022 8:40:00 PM Last Printed: 6/14/2022 9:25:00 PM

class structure in order to pursue idiosyncratic vision, and "bad" firms, aiming to use this structure to extract private benefits of control. If investors are able to observe firms' intentions pre-IPO and adjust the discount appropriately, then it would at least partly solve the adverse selection problem of "good" firms leaving the market, as the discount on a dual-class stock would be lower for "good" firms and higher for "bad" firms.

To determine this distinction, this paper analyses how the unique needs of technology firms—including their emphasis on idiosyncratic vision—might incline them towards a dual-class structure. Specifically, this paper explores how the protections from capital market pressure afforded by the use of dual-class structure might be crucial for fostering innovation and pursuing idiosyncratic vision. Indeed, my findings suggest that the conditions necessary for innovation, which are so vital to technology-based firms, are vulnerable to increased involvement by shareholders in three key ways. First, since technological innovation requires constant investment in new ideas with returns that may only exist in the long term, technology-based firms are at greater risk for quarter-to-quarter volatility, disrupting market's ability to evaluate long-term investments and potentially reducing the value of technology-based firms. Thus, under a single-class, one-share, one-vote structure, where technology-based firms are pressured to keep their earnings in line with forecasts, firms may counterproductively accept smaller, predictable earnings rather than larger and less predictable returns.[7] Second, specialized knowledge associated with technology investment and limitations on the ability to share details on innovative projects in progress, create asymmetric information between founders and outside shareholders that can generate uncertainty and divergent views concerning projects' economic potential.[8] More broadly, the process of transforming new ideas into tangible products is an ongoing and unexpected process.[9] Finally, within the context of the innovation process, a founder's value may lie not only in the specialized knowledge required to accurately evaluate technological innovation, but in

[7] For a general discussion of short versus long term value incentives see, for example, Leo E. Strine, *Towards a True Corporate Republic: A Traditionalist Response to Bebchuk's Solution for Improving Corporate America*, 54 HARV. L. REV. 1759 (2006) (responding to Professor Bebchuk's proposal to empower shareholders to amend corporate charters, by arguing it might undermine managerial flexibility and lead to a counterproductive short-term perspective); *see also* Martijn K.J. Cremers et al., Staggered Boards and Long Term Firm Value, Revisited, 126 J. of Fin. Econ. 422 (2017) (demonstrating that firms with lower firm value tend to adopt staggered boards, because of the potential to promote long-term value creation as a credible commitment device by the shareholders). The adoption of antitakeover defenses has a stronger positive association with firm value, where longer-term commitment plays a greater role in the firm's operation, as is the case with innovative firms).

[8] *See* Zohar Goshen & Assaf Hamdani, Corporate Control and Idiosyncratic Vision, 125 YALE L. J. 560, 565-566 (2016).

[9] *See* Ronald J. Gilson et al., *Braiding: The Interaction of Formal and Informal Contracting in Theory, Practice, and Doctrine*, 110 COLUM. L. REV. 1377, 1422-1444 (2010); Goshen & Hamdani, *supra* note 8, at 565-566 (explaining how where unexpected difficulties and redefinitions of a product can lead to delays in delivery, and delays can lead to disagreements between founders and outside shareholders about the product's future).

the ability to see the true value of the innovation where others do not, and to execute it according to her idiosyncratic vision. Thus, any limitation on founder's discretion can lead to early termination of potentially successful inventions. This paper tests the centrality of founders' idiosyncratic vision to the adoption of dual-class structure among technology-based firms.

Taking these unique features of the technology industry into account, this paper explores whether investors can observe firms seeking to pursue their founders' idiosyncratic vision and reflect it in the discount assigned to firm's value. If founders' control is important to firm value, and investors can observe it and distinguish the firm from firms wishing to extract private benefits, then a lower discount would allow an easier adoption of dual-class structure. Certain characteristics of firms employing dual-class structures have already been well-documented in the literature; media firms, firms including a person's name, and family or founder owned firms are all more likely to use a dual-class structure. Dual-class firms also tend to be larger, older, with higher profits relative to single-class firms. This paper contributes to the literature by developing a new, key correlation between firms and their adoption of dual-class structure that has so far been overlooked; media coverage on a firm's founders prior to firm's IPO. By quantifying media mentions of firm founders, and applying them to a hand-collected dataset of technology IPO from 2000 to 2015, this paper has constructed a novel empirical measure of founder's idiosyncratic vision. This measure also uncovered important differences in the patterns of media coverage across dual-class and single-class firms.

Unlike other possible proxies of a founder's idiosyncratic vision, media coverage is an observable measurement created by external media agencies rather than by founders themselves. The intrinsic motivation of founders captured by the media coverage variable, shows that a founder's idiosyncratic vision plays a crucial role in the decision to employ a dual-class structure. In a regression framework that controls for firm characteristics, the association of media coverage and the probability of having a dual-class status is both statistically significant and meaningfully substantial: a one standard deviation increase in media coverage increases the probability of having a dual-class structure from 9 percent to 11.5 percent. The data also demonstrates that the effect of media coverage differs significantly by a firm's state of incorporation, such that the effect of high media coverage on the probability of having a dual-class structure is more than 14 times higher for firms incorporated outside of Delaware than for firms incorporated in Delaware. The strength of corporate law, as well as the various network effects associated with incorporation in different states, might provide possible explanations for this striking finding.

This paper challenges the contention that technology-based firms are inclined to use dual-class structure based on the extraction of private benefits of control alone. Instead, its finding complements the private benefits theory by demonstrating that, at least in the technology sector, the desire to protect

founders' idiosyncratic vision is a key factor in choosing a dual-class structure. The data analyzed in this paper should be taken into account by institutional investors in their vigorous attempts to discourage the use of dualclass structure. If the incentive of technology-based firms in adopting dual-class structure is to implement their founder's idiosyncratic vision in a manner that maximizes firm value for all shareholders, then imposing penalties on these IPOs, or restricting or banning them more generally, is counterproductive.

Before continuing further, it is important to note that while the findings presented in this paper and the implications derived from them have been produced from technology-based firms, these IPOs account for a significant portion of all recent dual-class IPOs. About 22 percent of all dual-class IPOs from 2000 to 2015 were in the technology sector, with an increasing trend of dual-class IPOs.[10] Dual-class firms had an aggregated market value between $4 trillion and $5 trillion, which means that technology-based firms account for between $880 billion to $1.1 trillion.[11] In 2018, roughly 19 percent of the IPOs listed on U.S. exchanges had a dual-class stock structure, compared with 12 percent of firms in 2014, and just 1 percent in 2005.[12] As the prominence of dual-class stock firms steadily increases, this paper's analysis and its implications are increasingly crucial in the broader debate on the use of dual-class structure.

The remainder of this paper is organized as follows: Section I reviews the debate over the use of dual-class structure and the empirical evidence that supports each view. Section II describes the data used in the empirical analysis that follows. Section III analyzes the determinants of the dual-class structure in the technology sector. The last section concludes.

---

[10] *See* Bernard Sharfman, *A Private Ordering Defense of a Compay's Right to Use Dual-Class Shares in IPOs*, 63 VILL. L. REV. 1, 3 (2018) (estimating that the aggregated market value is close to $4 trillion as of June 2018); Robert J. Jackson, Jr., SEC Comm'r, Perpetual Dual-Class: The Case Against Corporate Royalty (Feb. 15, 2018), https://www.sec.gov/news/speech/perpetual-dual-class-stock-case-against-corporateroyalty (mentioning that public firms using dual-class worth more than $5 trillion as of Feb. 2018); *See* also SEC, Investor Advisory Committee, Discussion Draft: Dual Class and Other Entrenching Governance Structures in Public Companies (Dec. 17, 2017), https://www.sec.gov/spotlight/investor-advisorycommittee-2012/discussion-draft-dual-class-recommendation-iac-120717.pdf; Charles M. Elon & Craig K. Ferrere, *Unequal Voting and the Business Judgement Rule*, HARV. L. SCH. F. ON CORP. GOV. AND FIN. REG. (April 2, 2018), https://corpgov.law.harvard.edu/2018/04/07/unequal-voting-and-the-business-judgmentrule/.

[11] Stephen M. Bainbridge, *Should We Worry About Tech Sector Dual-Class Stock? In Short, No*, PROFESSORBAINBRIDGE.COM, April 4, 2017), http://professorbainbridge.com/professorbainbridgecom/2017/04/should-we-worry-about-techsector-dualclass-stock-in-short-no.html (describing data compiled by University of Florida finance professor Jay Ritter and explains that approximately 15 percent of all technology-based firms which went public between 2012 and 2016 used a dual-class structure, versus only 7 percent of all technology-based firms between 2007 and 2011).

[12] Council of Institutional Investors, investors Petition NYSE, NASDAQ To Curb Listings of IPO Dual-Class Share Companies (Oct. 24, 2018), https://www.prnewswire.com/news-releases/investors-petition-nysenasdaq-to-curb-listings-of-ipo-dual-class-share-companies-300737019.html ;SEC, Investor Advisory Committee, *supra* note 11; Elon & Ferrere, *supra* note 11.

I. *THE DUAL-CLASS DEBATE*

The use of dual-class structure has long been the subject of heated debate. Specifically, critics have been especially concerned with the "wedge" it creates between voting rights and cash-flow rights, allowing founders to misbehave, while public shareholders who bear the consequences have limited options to influence the firm. When Google went public with a dual-class structure in 2004, the debate gained steam; it broke unprecedented ground with Snap's IPO without voting rights in 2017. The growing trend of technology-based firms adopting a dual-class structure has heightened the discussion over the key drivers to retain unconditional control over the firm's decision making.

The prevailing critical view represented by the media, academia, and the institutional investor community points to agency costs and entrenchment problems, such as the availability of private benefits of control, and the ability to isolate from market pressure.[13] On this view, founders of technology-based firms adopt a dual-class stock structure in order to extract private benefit of control, such as empire building and tunneling resources, or make managerial decisions with limited accountability. They are able to "get away with it" due to their dominant position in the market. When Facebook went public in 2012, Institutional Shareholder Services, a proxy advisory firm, articulated this stance precisely, arguing that "this IPO event itself presents a Hobson's choice: accept governance structures which diminish shareholder rights and board accountability, or miss out on what appears to be one of the hottest business models of the internet age".[14] Consistent with this view, empirical studies demonstrate that dual-class stock structure is associated with lower shareholder value. Smart et al., found that the discount assessed to dual-class structure at the IPO date appears to be rational in the sense that post-IPO dual-class abnormal returns are zero.[15] Gompers et al., showed that firm value is negatively associated with the "wedge" between voting rights and cash-flow rights increases.[16] Furthermore, Masulis et al., reported that the wider the divergence between insider voting and economic interests, the less corporate cash reserves were worth to public shareholders, the more CEO compensation tended to be excessive, and the more managers were

---

[13] Lucian A. Bebchuk, A Rent Protection Theory of Corporate Ownership and *Control*, at 24 (Harvard Law and Econ. Discussion Paper No. 260, 1999).

[14] *The Tragedy of the Dual Class Commons*, Institutional Shareholder Services, Feb. 13, 2012, at 1, http://online.wsj.com/public/resources/documents/facebook0214.pdf.

[15] Scott B. Smart et al., *What's in a Vote? The Short- and Long- Run Impact of Dual-Class Equity on IPO Firm Values*, 16 J. ACCT. & ECON. 94 (2008).

[16] Paul Gompers et al., *Extreme Governance: An Analysis of Dual-Class Shares in the United States*, 23 REV. OF FIN. STUD. 1051, 1084 (2010).

File: Grinapell 17.1 To Publish_Edited Created on: 6/14/2022 8:40:00 PM Last Printed: 6/14/2022 9:25:00 PM

likely to carry out value-destroying acquisitions or were to make capital expenditures that yielded lower returns.[17]

Proponents of dual-class stock structure argue that it facilitates a focus on long term value.[18] Such benefit should be of a greater significance as innovation has become increasingly vital in the marketplace: to meet the growing pace of innovative development and user's demand for more frequent upgrades, technology-based firms have found they must invest in innovation and the R&D that generates it. This requires exploring new ideas, pursuing risky long-term plans and keeping information private for longer periods than ever before. The unique features of this value-creation process have led technology-based firms to seek more control on their innovative activities to insulate themselves from short-term financial pressures and to reach their full potential.[19] In addition, founders' vision, talent and skills are of greater importance in technology based firms, where a specialized knowledge is required and outside shareholders are less informed. Limiting market pressure and allowing founders to invest significant firm specific human capital in the firm could generate materially better results than under other, non-founder leadership.[20] Empirical support for this view is mainly found in studies on dual-class recapitalizations,[21] but several studies on dual-

---

17 Masulis et al., *Agency Problems at Dual-Class Companies*, 64 J. FIN. 1697-1727 (2009); *See also* Claessens et al., *Disentangling the Incentive and Entrenchment Effect of Large Shareholdings*, 57 J. FIN. 2741-2771 (2002) (showing that in East Asian firms, separation of cash flow and control decreases firm value. Note that the sample was dominated by pyramid ownership in Asian business groups, such that the authors could not clearly attributed the decrease in firm value to the dual-class structure); Karl V. Lins, *Equity Ownership and Firm Value in Emerging Markets*, 38 JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS 159 (2003) (indicating that in emerging markets, the wedge between voting and economic interests is negatively correlated with firm value).

18 *See, e.g.*, Jeremy C. Stein, *Takeover Threats and Managerial Myopia*, 96 J. POL. ECON. 61 (1988); Jeremy C. Stein, Efficient Capital Markets, Inefficient Firms: A Model of Myopic Corporate Behavior, 104 Q. J. ECON. 655 (1989).

19 *See* Letter from Larry Page, CEO and Co-Founder, Google, & Sergey Brin, Co-Founder, Google, to Google Shareholders (Apr. 2012), (noting that "technology products often require significant investment over many years to fulfill their potential. For example, it took us over three years to ship our first Android handset, and then another three years on top of that before the operating system truly reached critical mass. These kinds of investments are not for the faint-hearted").

20 Armen A. Alchian & Harold Demsetz, *Production, Information Costs, and Economic Organization*, 62 AM. ECON. REV. 777 (1972); Harry DeAngelo & Linda DeAngelo, *Managerial Ownership of Voting Rights: A Study of Public Corporations with Dual Classes of Common Stock*, 14 J. FIN ECON.. 33, 35 (1985); Strine, *supra* note 7, at 1763; Microsoft and Apple are two examples of successful firms which had struggled without Bill Gates and Steve Jobs, their respective founders.

21 Kenneth Lehn et al., *Consolidating Corporate Control: Dual-Class Recapitalizations Versus Leveraged Buyouts*, 27 J. FIN ECON. 557, 559 (1990) (showing that dual-class recapitalization firms grow faster and engage in secondary equity offerings (SEOs) more frequently following recapitalizations); Valentin Dimitrov & Prem Jain, *Recapitalization of One Class of Common Stock into Dual-Class: Growth and Long Run Stock Return*, 12 J. CORP. FIN. 342 (2006) (demonstrating that growth is beneficial to the shareholders, as they earn significant positive abnormal returns in the years to follow the announcement); Scott W. Bauguess et al., *Large Shareholder Diversification, Corporate Risk Taking, and the Benefits of*

class IPOs reinforce it as well: Bohmer et al. reported normal to above-normal market and operating performance of dual-class IPOs relative to a matched sample of single-class IPOs, which indicates that dual-class structure confers benefits to all firm shareholders.[22] Taylor and Whittred found that Australian "second board" firms adopting dual-class structure have a higher rate of growth opportunities relative to single class firms, and so realization of these opportunities highly depends on the freedom of the founder to pursue its idiosyncratic vision.[23] Most recently, Cremers et al. and Kim and Michaely showed that dual-class structures tend to have a valuation premium over single-class firms during the first few years after the IPO.[24] Furthermore, Field and Lowry found that when insiders' power is smaller and less permanent, larger institutional investors with long-term holdings are less likely to vote against directors, suggesting the benefits outweigh the costs for these firms in the early stage of their public life cycle.[25]

Research on the factors influencing the choice of a dual-class structure suggests a correlation with several characteristics: family or founder ownership,[26] person's name in the firm's name,[27] media firms,[28] larger firms,[29] older age,[30] higher CEO salary,[31] higher leverage,[32] higher profits and more assets.[33]

Changing to Differential Voting Rights, 36 J. BANK. AND FIN. 1244 (2012) (indicating that the change from single-class to dual-class improves firm's performance, as greater corporate risk taking are being employed).

22 Ekkehart Bohmer et al., *The Effect of Consolidated Control on Firm Performance: The Case of Dual Class IPOs*, at 95 EMPIRICAL ISSUES IN RAISING CAPITAL 95 (1996).

23 *See generally*, Steven Taylor & Greg Whittred, *Security Design and the Allocation of Voting Rights: Evidence from the Australian IPO Market*, 4 J. CORP. FIN. 107 (1998).

24 Martijn Cremers et al., *The Life-Cycle of Dual Class Firms* 1, 27 (Euro. Corp. Governance Inst., Working Paper Series in Finance, Working Paper No. 550, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3062895; Hyunseob Kim & Roni Michaely, *Sticking Around Too Long? Dynamics of the Benefits of Dual Class Structures* 1, 34 (Euro. Corp. Governance Inst., Working Paper Series in Finance, Working Paper No. 590, 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3145209.

25 Laura Casares Field & Michelle B. Lowry, Bucking the Trend: Why do IPOs Choose Controversial Governance Structures and Why Do Investors Let Them (October 6, 2020) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2889333

26 DeAngelo & DeAngelo, *supra* note 20, at 50-51; Laura Casares Field, *Control Considerations of Newly Public Firms: The Implementation of Antitakeover Provisions and Dual Class Shares Before the IPO*, at 6 (Working Paper, 1999), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=150488; Ben Amoako-Adu & Brian Smith, *Dual Class Firms: Capitalization, Ownership Structure and Recapitalization Back into Single Clas*s, 25 J. BANKING & FIN. 1083, 1096 (2001); Field & Lowry, *supra* note 25, at 9–10.

27 Gompers et al., *supra* note 16, at 1104.

28 DeAngelo & DeAngelo, *supra* note 20, at 40; Gompers et al., *supra* note 16, at 1104.

29 Kim & Michaely, *supra* note 24, at 10.

30 Scott B. Smart & Chad J. Zutter, *Control as a Motivation for Underpricing: A Comparison of Dual and Single-Class IPOs*, 69 J.FIN. ECON. 85, 111 (2003); Kim & Michaely, *supra* note 24, at 10; Field & Lowry, *supra* note 25, at 9.

31 Field, *supra* note 26, at 17.

32 Kim & Michaely, *supra* note 24, at 10.

33 Smart & Zutter, *supra* note 30, at 111.

Building on and departing from this extant body of literature, this paper analyzes possible determinants on the drivers to adopt dual-class structure in the technology sector in several unique ways. First, to-date, most research has used trading dual-class firms of all ages in their samples, this analysis examines only new dual-class IPOs. The IPO-approach to studying drivers of dual-class structure has a distinct advantage over analyzing established firms, since the considerations prior to going public, and the consequences of such a move, are different in later stages. Second, where previous research has analyzed all industries, with some additional emphasis on the media sector, this paper focuses its attention on the technology sector, and in particular on technology-based founder firms. Spotlighting the technology sector makes it possible to examine whether technology warrants unique treatment distinct from other types of industries– as many technology firm executives claim should be the case. Limiting the scope to founder firms also assists in this evaluation, as the motivations underlying the adoption of dual-class structures for founder firms might be different than those of equity carve-outs in which the superior class belongs to the parent firm. Third, most previous papers compared firms from the 1980s to 2002, a timespan limited in its ability to test the effect of the rapid and incalculable nature of innovation on the choice of stock structure, particularly in the technology sector. This paper's sample includes newer data between the years 2000 and 2015, bringing the extant research up to speed. These three crucial differences in this paper' sampling choice allow its analysis to capture the increasing trend of technology-based firms to utilize dual-class structure, as well as to better understand the motives behind such trends.

II. *DATA*

The data set analyzed here includes all IPOs of both U.S. and foreign firms registered for trade on American stock exchanges and filed with the Securities and Exchange Commission (SEC) from January 2000 through December 2015. The list of IPOs, together with their offer dates, was constructed from IPOScoop.com, Thomson Financial Venture Expert, and Securities Data Company (SDC) data.[34] Each firm's stock structure and Standard Industrial Code (SIC) were collected from Jay R. Ritter IPO database.[35]

[34] The information collected from these sources was compared and amended in Martin Kenney & Donald Patton, *Kenny-Patton Firm and Management Database of Emerging Growth IPOs: 1990–2010* (2013), https://kenney.faculty.ucdavis.edu/kenney-patton-firm-and-management-database-of-emerging-growthipos-1990-2010/.

[35] Jay R. Ritter database includes a description all IPOs from 1980–2018, including- offering with multiple share class outstanding: https://site.warrington.ufl.edu/ritter/files/2017/09/IPOs-from-1980-2017with-Multiple-Share-Classes-Outstanding.pdf. This source of information is described in more detail on Appendix D of Tim Loughran & Jay R. Ritter, *Why Has IPO Underpricing Changed over Time?* 33 FIN. MGMT. 5, 35 (2004).

Ritter and Loughran's definition of "tech stock" was used to filter technology SIC codes.[36] The data set has been updated to reflect changes in "tech stock" definition, as reported in Jay R. Ritter's IPO database. Mutual funds, real estate investment trusts (REITs), asset acquisition or blank check corporation, firms that have already gone public, and all spin-offs, carve-outs or subsidiaries of other firms were all removed from the sample.

The primary explanatory variable of interest is media coverage of a firm's founder prior to the firm's IPO, which serves as a proxy for the founder's idiosyncratic vision. "Media coverage" was constructed in several steps. To find the name that media outlets use to refer to a firm, this paper's analysis followed Tetlock et al. in matching the names on the IPO list with common firm names taken from one of the three web-based data sources: Mergent Online, the SEC website, or Factiva database.[37] Mergent Online company search tool was used to identify common names for all firms founded after 1993, and the SEC company search function was used to find names that were missing from Mergent Online. Addiitonally, Factiva's search feature provided common names of firms that existed prior to 1993.[38] Numerous names had to be abbreviated in order to improve the search by focusing on only the first word or two of the name.

The primary data source to find founders' names was the SEC filings. Going public requires a registration statement (SEC Form S-1, or SEC Form F-1 if the company is foreign), which includes an investment prospectus to share with potential investors and additional information to file with the SEC. In the prospectus, the company discloses important information about the business, its financial condition, results of operations, risk factors, and management, as well as its audited financial statements. . A Company must also file, on an ongoing basis, annual reports that provide a comprehensive view of the company during the past year (SEC Form 10-k) and quarterly reports which give an update on its financial position throughout the year (SEC Form 10-Q). The reports include information regarding the company history, organization, nature of business, equity, holdings, earnings per share, subsidiaries, and other relevant financial information.. All documents were closely examined for any biographical details on the founders. For firms with no information in the SEC filings, a company profile search was run in the Business Source Complete database, followed by a general search on the firm's website and the founder's LinkedIn page. Finally, for the few firms with no information, a search was conducted though Factiva, with priority given to major financial news sources, such as the Wall Street Journal, the New York Times, and the Financial Times. In cases where no information was available on a firm's founders, firms were excluded from the sample. To obtain a precise measure of founder's media coverage, the search query included the

36 Loughran & Ritter, *supra* note 35, at 35.

37 Paul Tetlock et al., *More than Words: Quantifying Language to Measure Firms' Fundamentals*, 63 J. FIN. 1437, 1465 (2008).

38 *Id.*

name of the founder together with the common name of the firm, so as to avoid confusion if, for example, the same founder was associated with more than one firm, or multiple people shared the founder's name. The search was also adjusted to allow for variations of the founder's first and middle names. The search window was limited to the period between incorporation date and one day prior to the offering date, counting the number of articles and news stories that mention the founder and the firm during this window. All web news and blogs, wire services, and press release wires were excluded, as they were considered unreliable news sources, easy to manipulate or "spin" in a way that might benefit the firms.[39] As a measure of news tone, the number of positive and negative news articles was obtained by the Factiva news sentiment filter, which uses a fixed set of terms to identify positive and negative stories. Each founder's media coverage was measured by three values: the overall number of news articles, the number of positive news articles, and the net number of positive news articles (positive minus negative articles).

For the media coverage variable, three alternative measurements were created: "media coverage sum," which sums the media coverage for all founders of a firm; "media coverage average," which scales the sum by the number of founders; and "media coverage max," which uses the maximum number of media mentions on one of the founders. All three measurements have their advantages, but "media coverage max" was selected as the primary measurement, since it is less sensitive to possible gaps in the collected data.

Control and robustness checks variables used in the analysis below include: EBITDA-to-price ratio, debt-to-assets ratio, R&D expenses-to-assets ratio (imputing missing values of R&D to zero),[40] income-to-assets ratio, 1-year sales growth rate (the change in revenue between the year prior to the IPO to two years prior to the IPO), total assets, revenue, firm's age (based on the founding year to capture time passed to firm's IPO), offer year, and state of incorporation (a dummy variable equals to one for incorporation in Delaware, equals to two for incorporation in non-Delaware state, and zero otherwise). The financial data was collected from Standard and Poor's Compustat database for the year prior to the IPO. The public offering price, adjusted for underwriter discount, and state of incorporation were collected from the firm's prospectus on the SEC website. Founding years were taken from Jay R. Ritter IPO database.[41]

Table 1 reports the descriptive statistics for the sample of 618 single-class IPOs and 62 dual-class IPOs for years 2000–2015. The data displays some well-known characteristics of dual-class firms. Relative to single-class

---

39 *See, e.g.,* David H. Solomon, *Selective Publicity and Stock Prices*, 67 J. FIN 599, 611-615 (2012), indicating that IR firms are likely to directly influence media coverage.

40 For studies making similar adjustments, *see, e.g.*, Alon Brav et al., *How Does Hedge Fund Activism Reshape Corporate Innovation*, 130 J. FIN. ECON. 237, 240 n.7 (2018); Kim & Michaely, *supra* note 24, at 45.

41 Jay R. Ritter database, *supra* note 35.

firms, dual-class firms tend to be older, while providing a higher offering price. They also have more assets, profits, debt, and leverage, and less R&D expenses (scaled by assets) than single-class firms. Furthermore, the data shows that a larger proportion of dual-class firms, relative to single-class firms, went public during the second half of the sample period, which might be explained by the increasing trend of technology-based firms to employ a dual-class structure at the IPO stage.[42]

The most illuminating finding relates to differences in media coverage on founders of dual-class and single-class firms prior to their IPO. Founders of dual-class firms have significantly higher media coverage relative to founders of single-class firms, with a much more significant difference in the means than in the medians of the media coverage variable. These results are consistent for the different media coverage variables (media coverage max, media coverage average, media coverage min, positive coverage, net coverage). A possible explanation might be that the right-tail of relatively high media coverage is substantially larger among dual-class firms relative to single-class firms; the 75th percentile of dual-class firms, for instance, has about the same media coverage as the 95th percentile of single-class firms.

Another relevant finding concerning the use of dual-class structure involves the states in which firms choose to incorporate. Among U.S. incorporated firms, 490 firms were incorporated in Delaware (81 percent of firms), and 112 firms were incorporated outside of Delaware (19 percent of firms). Strikingly, the proportion of dual-class and single-class firms incorporated in Delaware is identical to the overall proportion of firms (81 percent of the dual-class and the single-class firms). The entire sample reveals a similar pattern. 490 firms were incorporated in Delaware (72 percent of firms) and 190 were incorporated outside of Delaware (28 percent). Slightly less than 76 percent of the dual-class firms and slightly less than 72 percent of the single-class firms were incorporated in Delaware. Apart from an earlier IPO year, no substantial differences were found between firms formed in Delaware and firms incorporated outside of Delaware. Beyond this IPO year, these findings do not indicate any strong general tendency towards selection across Delaware and non-Delaware firms.

---

42 Bernard Sharfman, Robert J. Jackson, Jr., also SEC, Investor Advisory Committee, Discussion Draft: Dual Class and Other Entrenching Governance Structures in Public Companies; Charles M. Elon & Craig K. Ferrere, *supra* note 10.

III. *DETERMINANTS OF A DUAL-CLASS STRUCTURE IN THE TECHNOLOGY INDUSTRY*

Building upon previous finance literature,[43] a proxy of founders' idiosyncratic vision was constructed to analyze determinants of dual-class structure in the technology sector. This proxy measured the number of articles and news stories on a firm's founders prior to the IPO.

The paper assumes two dimensions of quality: one is the distinction between founders of technology-based firms who intend to extract private benefits of control and those who do not. For both groups, destruction of a firm's value is possible, but for the former, it is driven by a desire to extract private benefits of control, whereas for the latter, it is driven by the desire to realize founders' idiosyncratic vision (even if they turn out to be mistaken ex post). Insofar as there is an intrinsic motivation to realize idiosyncratic value—which cannot be consciously created by founders—it provides some degree of assurance that the founders will not self-deal, even if they have the opportunity to do so in the future. Among those who do not intend to expropriate, some founders are of higher quality, or their control is at least more important to firm value. Media-based measures of reputation are a reasonable proxy for this dimension of quality, since "good" firms are likely to automatically acquire favorable media coverage in the course of their activities. As stronger founder reputation leads to lower discounts for dual-class stocks, the strength of the founders' reputation should lead to the choice of dual-class structure. If founders' reputation is weak, a dual-class choice would be inferred by investors as a way to use control for self-dealing, and thus a given firm should prefer a single-class structure. If founders' reputation is strong, investors can observe the reputation and distinguish the firm from self-dealers, which will lead to its use of dual-class structure.

Media coverage is a novel measure of founders' idiosyncratic value. While the finance literature has documented the impact of media coverage on market trading,[44] prices of large and widely followed stocks,[45] and initial returns and long-term value,[46] not much is known about the relationship

---

43 See, e.g., Gompers et al., *supra* note 16; DeAngelo & DeAngelo, *supra* note 20; Lehn et al., *supra* note 21; Taylor & Whittred, *supra* note 23; Field, *supra* note 26; Amoako-Adu & Smith, *supra* note 26; Smart & Zutter, *supra* note 30.

44 *See* Brad M. Barber & Terrance Odean, *All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors*, 21 REV. FIN. STUD. 785, 787–89 (2008); Utpal Bhattacharya et al., *The Role of the Media in the Internet IPO Bubble*, 44 J. FIN. & QUANTITATIVE ANALYSIS 657, 659 (2009); Joseph E. Engelberg & Christopher A. Parsons, *The Causal Impact of Media in Financial Markets*, 66 J. FIN. 67, 68 (2011).

45 *See* Paul C. Tetlock, *Giving Content to Investor Sentiment: The Role of Media in the Stock Market*, 62 J. FIN. 1139 (2007); Paul C. Tetlock, *All the News That's Fit to Reprint: Do Investors React to Stale Information*, 24 REV. FIN. STUD. 1481 (2011).

46 *See* Laura Xiaolei Liu et al., *The Long-Run Role of the Media: Evidence from Initial Public Offerings*, 8 MGMT. SCI. 1945, 1947 (2014).

between media coverage on founders and their firms' decisions to adopt a dual-class structure. Measuring media coverage involves the risk of capturing a high degree of "noise," as articles and news stories can be manipulated by firms to advance their own interests. Regardless of intentional manipulation, the media also can mistakenly provide favorable coverage to "bad" firms.[47] As noted above, to reduce the risk associated with manipulation, the search query was limited to exclude blogs, wire services and press releases, that were considered as less reliable news outlets. As for the risk of media mistakes or misreporting, a validation check was implemented to test the importance of media coverage to the IPO price. The baseline assumption is that while it might be that the media detects idiosyncratic value imperfectly, there is a positive correlation between the two. The empirical specification used here included regressing IPO price on media coverage, while controlling for firm characteristics, including EBITDA-to-price, debt-to-assets, R&D expenses-to-assets, income-to-assets, and 1-year sales growth rate. The results are reported in Table 2, with robust standard errors. Column 1 presents the results of the regression with no controls, while Column 2 includes the full set of controls. The findings imply that media coverage has a positive and highly significant association with IPO price.

Based on the correlation between media coverage and IPO price, an empirical specification was constructed to model media coverage as a determinant of dual-class structure, while controlling for firm characteristics. The empirical specification included regressing dual-class status on media coverage along with variables from earlier regression, firm's age, IPO year, and state of incorporation as controls. Table 3 displays the ordinary least squares (OLS) results, using robust standard errors. The estimated 0.510015 coefficient implies that the media coverage variable has a positive and highly significant association with the probability of having a dual-class status. Moreover, the results are robust to measures of firm size, such as total assets and revenue, to mitigate the concern that larger firms have a higher degree of familiarity and thus could potentially acquire more media coverage. Controlling for firm's age addresses the problem of measuring media coverage for longer windows for older firms relative to shorter windows for younger firms. In addition to the positive and highly significant association, the magnitude of the estimated relationship between media coverage and the probability of having a dual-class status is substantial: a one standard deviation increase in media coverage (corresponding to 477 articles and news stories) increases the probability of having a dual-class structure from approximately 9 percent to approximately 11.5 percent. The baseline analysis includes media coverage max, but media coverage average and media coverage sum lead

47 For active attempts by firms to influence their stock prices though the media, *see generally* Kenneth R. Ahern & Denis Sosyura, *Who Writes the News? Corporate Press Releases During Merger Negotiations*, 69 J. FIN. 241 (2014) (showing that bidders in stock mergers originate substantially more new stories after the start of merger negotiations, but before the public announcement to generate a short-lived run-up in bidders' stock prices during the period when the stock exchange ratio is determined).

to consistent results. The correlation is much stronger with positive media coverage and net media coverage.

Separately regressing dual-class status on media coverage before and after Google's IPO year displays that the overall results are mostly driven by the post-2004 period. Table 4 reports OLS results, using robust standard errors. Column 1 presents the results of regressing dual-class status on media coverage prior to the year 2004. Column 2 reports the results for the years from 2004. The results are only statistically significant for the years from 2004. The point estimate for this period is similar to the one reported for the overall regression. This is robust to the different measures of firm size to address the possibility that the results are driven by smaller firms going public in the pre-2004 period relative to post-2004 IPO firms.

It should be noted that while the probit model supports the correlation between media coverage and dual-class structure, the logit model leads to insignificant results. The different results, in this case, may be attributed to the relatively small sample size, particularly of dual-class firms. Figure 1 backs the results reported in Table 3. The bars depict means of media coverage of dual-class and single-class firms. The bar chart portrays clear and significant differences in the means of dual-class firms and single-class firms, which suggests that firms with higher media coverage are more likely to adopt a dual-class structure. Figure 2 and Figure 3 depict the means of media coverage of dual-class and single-class firms for the periods pre- and post-2004, respectively. These bar charts illustrate that significant differences in means appear to exist only for the years from 2004. The difference in means portrayed in Figure 3 is much larger than the one in Figure 1, which reinforces the finding that the correlation between media coverage and dual-class structure is significant.

The observed correlation between media coverage on founders prior to the IPO and dual-class structure suggests that a major factor in the choice of dual-class structure is founders' idiosyncratic vision. Conventional private benefits of control would presumably not be related to the media coverage variable, as "bad" firms should not be expected to automatically acquire favorable coverage in their course of activities. Moreover, even if firms can invest in generating media coverage, it would be much more expensive for "bad" firms to acquire such coverage in traditional and reliable news outlets.

The effect of media coverage on the probability of having a dual-class status was also analyzed with regard to the state of incorporation. Column 1 of Table 5 presents the results of regressing a dummy stock structure variable on the interaction term of Delaware incorporation and the media coverage variable, while controlling for firm characteristics. The sample was limited to U.S. incorporated firms. Interestingly, the estimated correlation between media coverage and dual-class structure is considerably stronger for firms incorporated outside of Delaware. The results are robust to the different measurements of firm size. Column 2 and Column 3 of Table 5 illustrate such differences: the 0.045 coefficient reported in Column 2 relative to the

0.64 coefficient reported in Column 3 suggests that for firms incorporated outside of Delaware, the effect of high media coverage on the probability of having a dual-class structure is more than 14 times higher. Table 5 reports OLS results, using robust standard errors, but probit and logit probability models lead to highly consistent results. The results were also found to be significant for the entire sample of firms, which includes also firms incorporated outside of the US.

It may well be the case that Delaware corporate law, together with its specialized corporate courts and the broad discretion in enforcement of fiduciary duties, provide a better protection for minority shareholders relative to corporate law in other states.[48] Under this view, founders' vision might be less important in the choice of dual-class and single-class structure among firms incorporated in states with stronger corporate law, as the choice of incorporation provides an alternative signal that the founders are unlikely to extract private benefits of control. Among firms incorporated in weaker corporate law states, however, media coverage may be of greater importance. A low-quality firm will tend to choose a single-class structure, or risk being perceived as a self-dealer with a high discount assigned to its value, whereas a high-quality firm will tend to choose a dual-class structure despite a lower discount for its use.

The strength of states' corporate law has been previously used in the literature to analyze determinants of firms' decisions regarding where to incorporate,[49] to evaluate incorporation's effect on stock returns,[50] and to examine determinants of the dual-class structure.[51] Gompers et al., hypothesized that for firms incorporated in states with stronger anti-takeover laws, dual-class structure might be less valuable; their study, however, was unable to find a statistically significant correlation between state law and dual-class structure.[52] Regressing a dual-class dummy on incorporation in Delaware dummy, while controlling for firm characteristics, leads to consistent results. Technology-based firms at the IPO stage were not found to have statistically significant correlation between Delaware incorporation and dual-class stock structure. If the main reason to adopt a dual-class structure is to extract private benefits of control, and Delaware is in fact better at protecting minority shareholders, then expropriators should prefer to incorporate outside of Delaware, and to choose a dual-class structure. While there is clear evidence that media coverage is of greater importance in the dual-class choice outside of Delaware, there is no evidence that non-Delaware incorporated firms are

---

48 See, e.g., Elon & Ferrere, *supra* note 10.

49 *See* Lucian A. Bebchuk & Alma Cohen, *Firms' Decision Where to Incorporate*, 46 J. L. & ECON. 383, 387 (2003); Field & Lowry, *supra* note 25, at 34.

50 Paul Gompers et al., *Corporate Governance and Equity Prices*, 118 Q. J. ECON. 107 (2003); *See* Robert Daines, *Does Delaware Law Improve Firm Value?* 62 J. FIN. ECON. 525, 534-540 (2001). .

51 Gompers et al., *supra* note 16, at 1063.

52 Gompers et al., *supra* note 16, at 1101-1105; State law was proxied by an index of state takeover laws, defined in Gompers et al., supra note 42, from the firm's state of incorporation in the previous year.

more likely to use dual-class structure. This might indicate that expropriation is not the only driver for firms to adopt a dual-class structure.

The premise that Delaware corporate law provides a better protection for minority shareholders has been controversial. Though it has gained support in the literature, it has also been contested by prominent corporate scholars. Daines, for example, reported that incorporation in Delaware between 1981 and 1996 was associated with higher Tobin's Q relative to incorporation outside of Delaware. Tobin's Q, the ratio between a firm's market value and its book value, is a widely used measure of the value of firm's assets in relation to their replacement costs.[53] Based on this finding, it was inferred that minority shareholders are willing to pay more for the share of Delaware firms in exchange for better minority protection.[54] However, as suggested elsewhere, the correlation might have been due to selection effect, in which firms with higher value tend to incorporate in Delaware.[55] The authors demonstrated that controlling for firm age and for firm-specific corporate governance arrangements, no correlation exists between Delaware incorporation and higher Tobin's Q in 1990 to 2001 data.[56]

It is also possible that the stronger effect of media coverage on the probability of having a dual-class structure for firms incorporated outside of Delaware is led by network externalities. Romano and Klausner argued that the value of a state's corporate law depends on the number of firms incorporated in that state.[57] The large number of firms incorporated in Delaware results in diverse and continuously developing legal precedents, as well as a large and comprehensive base of legal opinions. Investors tend to not only have previous familiarity with Delaware corporate law, but can also expect to benefit from its predictability going further. Under these circumstances, founders' vision might be less important in the choice of dual-class and single-class among firms incorporated in Delaware, as network externalities provide investors with reassurance as to Delaware's ability to restrain self-dealing. Among firms incorporated outside of Delaware, however, media coverage may be of greater importance, as investors are less familiar with the strength of a state's corporate law. Furthermore, even if a state's corporate law is identical to Delaware's, investors cannot adequately predict a state court's future rulings. For these firms, media coverage serves as a much more significant signal of firm's intrinsic motivation. A lower quality firm will tend to choose a single-class structure, otherwise it will be perceived as an expropriator, whereas a high-quality firm will tend to choose a dual-class structure.

---

53 *See* Daines, *supra* note 50, at 530.

54 *See* Daines, *supra* note 50, at 537-540.

55 *See* Bebchuk et al., *Does the Evidence Favor State Competition in Corporate Law?* 90 CALIF. L. REV. 1775, 1788–90 (2002); Bebchuk & Cohen, supra note 42.

56 *Id.*

57 *See* Roberta Romano, *Law as a Product: Some Pieces of the Incorporation Puzzle*, 1 J. L. ECON. & ORG. 225 (1985); Michael Klausner, *Corporations, Corporate Law, and Networks of Contracts*, 81 VA. L. REV. 757, 774-778 (1995).

CONCLUSION

This paper has aimed to contribute to the discussion on dual-class structure utilization, by focusing on technology-based firms. Its analysis has demonstrated that extraction of private benefits of control alone cannot adequately explain the tendency of technology-based firms to adopt dual-class structure. Rather, its findings complement the private benefits theory by showing that, at least in the technology sector, the desire to protect founders' idiosyncratic vision also drives the choice of dual-class structure. Founders' reputation prior to the IPO, measured by media coverage, is being used by the market to evaluate firm's intentions post-IPO. For firms with a weak founders' reputation, a dual-class choice would be inferred by investors as a way to use control to extract private benefits, thus, firms would tend to settle for a single-class structure. Firms with strong founders' reputation, however, would not be perceived as expropriators and would tend to choose a dual-class structure. Further analysis has also indicated that founders' reputation matters more for firms incorporated outside of Delaware than for firms incorporated in Delaware. This result might be driven by the strength of corporate law in and outside of Delaware, as the choice to incorporate in strong corporate law states can provide an alternative signal that the founders are unlikely to extract private benefits of control. Another possible explanation might be that Delaware corporate law is associated with stronger network externalities, which provide additional reassurance of a firm's inability to self-deal.

These findings highlight some of the costs associated with the penalties and obstructions imposed by institutional investors to discourage the use of a dual-class structure. If the incentive of technology-based firms in adopting a dual-class structure is to implement their founder's vision to benefit all shareholders, then imposing penalties or other obstructions is counterproductive. Whereas media coverage can assist large or well-established technology-based firms to obtain dual-class structure, smaller or newer technology-based firms with a less established reputation are pushed to stay private, to be acquired by larger firms, or to settle for a stock structure that does not maximize their value.

TABLE 1. DESCRIPTIVE STATISTICS

Variable	Dual-Class Firms			Single-Class Firms		
	Mean (Median)	Std. Dev.	Obs.	Mean (Median)	Std. Dev.	Obs.
Media Coverage Max.	645.73 (69.5)	3145.43	62	66.39 (23)	209.12	618
Media Coverage Ave.	560.70 (46)	3118.03	62	47.99 (16)	131.32	618
Delaware Inc.	0.7580645	0.4317514	62	0.7168285	0.450904	618
EBITDA-to-Price	0.043 (0.012)	0.217	62	0.039 (-0.006)	1.038	615
Debt-to-Assets	0.226 (0.057)	0.347	62	0.173 (0.051)	0.308	614
R&D-to-Assets	0.136 (0.088)	0.142	62	0.220 (0.171)	0.236	614
Income-to-Assets	-0.107 (-0.034)	0.307	62	-0.297 (-0.149)	0.605	614
1-Y Sales Growth Rate	1.473 (0.639)	3.393	61	677.63 (0.613)	6327.15	610
Age	12.268 (9)	10.434	62	8.012 (8)	9.324	618
Offer Year	2007.081 (2006)	5.451	62	2005.655 (2005)	5.157	618
Price	17.263 (14.88)	11.123	62	12.873 (12.09)	5.249	618
EBITDA (per share)	0.621 (0.177)	2.507	62	0.693 (-0.093)	16.887	615
Total Debt	149.948 (8.633)	312.391	62	46.272 (1.694)	295.508	614
Assets	677.753 (196.344)	1361.823	62	133.858 (44.639)	483.599	614
Revenue	447.457 (128.641)	966.649	62	110.574 (38.580)	384.688	614
R&D Expenses	69.233 (11.32)	209.643	62	11.806 (6.983)	34.541	614
Net Income	-8.768 (-1.708)	171.332	62	-7.410 (-5.750)	44.185	614

TABLE 2. THE EFFECT OF MEDIA COVERAGE ON IPO PRICE IN TECHNOLOGY SECTOR

Dependent variable:	IPO PRICE	
	(1)	(2)
Media Coverage Max	1.334136 (0.3646178)***	1.319458 (0.3681714)***
EBITDA-to-Price		0.1475134 (0.0386488)***
Debt-to-Assets		-0.8099503 (0.565874)
Income-to-Assets		0.5628848 (0.427605)
R&D-to-Assets		-0.9989577 (1.115608)
1-Y-Sales Growth Rate		0.0000717 (0.0000337)**

Note: This table reports estimated coefficients from OLS model. Robust standard errors are in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE 3. DETERMINANTS OF DUAL-CLASS STATUS IN TECHNOLOGY SECTOR

Dependent variable:

	DUAL-CLASS STATUS (=1)	
	(1)	(2)
Media Coverage Max	0.0510015 (0.0140756)***	0.0479913 (0.0127925)***
Inc. in DE		0.0474294 (0.031363)
Inc. in non-DE US state		0.0515739 (0.0401337)
EBITDA-to-Price		-0.0006963 (0.033127)
Debt-to-Assets		0.0302111 (0.0417716)
Income-to-Assets		0.0195476 (0.0188941)
R&D-to-Assets		-0.0852577 (0.045553)*
1-Y-Sales Growth Rate		-0.0000008 (0.0000004)*
Offer Year		0.0027743 (0.0024025)
Age		0.0027743 (0.0020159)

Note: This table reports estimated coefficients from OLS model, in which the dependent variable equals one for dual-class status, and zero otherwise. Robust standard errors are in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE 4. MEDIA COVERAGE AS A DETERMINANT OF DUAL-CLASS STATUS IN TECHNOLOGY SECTOR – PRE AND POST 2004

Dependent variable:

	DUAL-CLASS STATUS (=1)	
	(1) Pre-2004	(2) Post-2004
Media Coverage Max	0.1562644 (0.2996451)	0.0476426 (0.0133391)***
Inc. in DE	-0.0926101 (0.0867195)	0.1035357 (0.0294305)***
Inc. in the non-DE state	-0.0999972 (0.0888915)	0.1178411 (0.0499843)**
EBITDA-to-Price	0.4659789 (0.5378212)	-0.0010996 (0.0031456)
Debt-to-Assets	0.1635176 (0.0942753)*	-0.03192 (0.0477118)
Income-to-Assets	0.0049511 (0.0247306)	0.0128433 (0.0310022)
R&D-to-Assets	-0.0970117 (0.0542336)*	-0.1268777 (0.0817697)
1-Y-Sales Growth Rate	-0.00000354 (0.00000229)	-0.0000001 (0.0000003)
Offer Year	0.0237098 (0.0249649)	0.0049939 (0.0042472)
Age	0.0019954 (0.0038158)	0.0023926 (0.0022264)

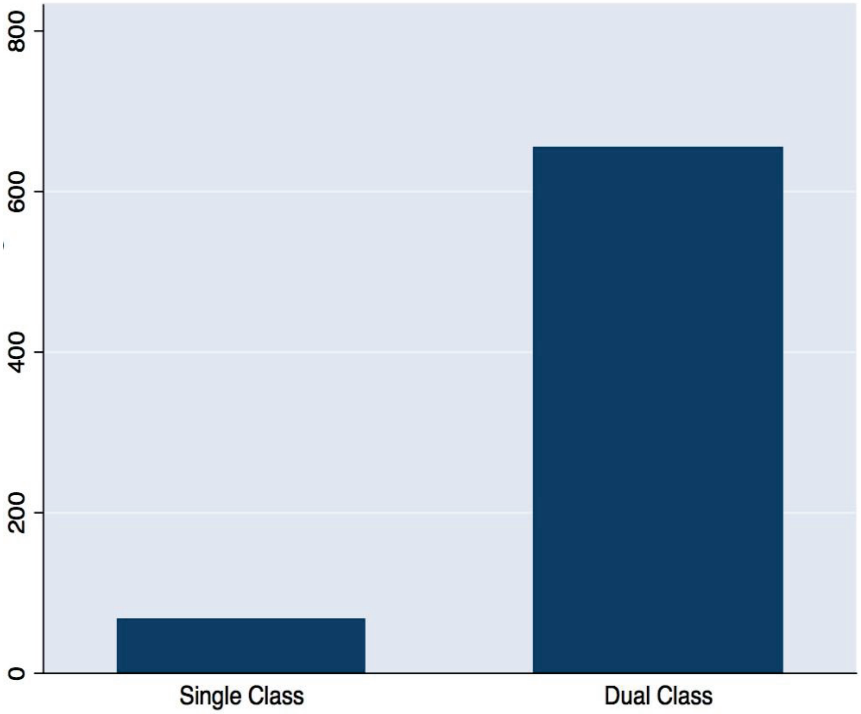
Note: This table reports estimated coefficients from OLS model, in which the dependent variable equals one for dual-class status, and zero otherwise. Robust standard errors are in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%.

TABLE 5. DUAL-CLASS STATUS AND MEDIA COVERAGE BY STATE

	(1) All States	(2) Delaware	(3) non-Delaware States
Delaware Inc. (=1)* Media Coverage Max	-0.5882058 (0.1476571)***		
Media Coverage Max	0.6337844 (0.1474421)***	0.0454417 (0.0105931)***	0.640046 (0.155364)***
Delaware Inc. (=1)	0.0431873 (0.0302308)		
EBITDA-to-Price	-0.0003869 (0.0035326)	-0.0005718 (0.0032937)	0.6927978 (0.8161789)
Debt-to-Assets	0.0175971 (0.0434942)	-0.0187695 (0.0574035)	0.0055121 (0.049944)
R&D-to-Assets	-0.067661 (0.0463143)	-0.0910958 (0.0583038)	0.002799 (0.0657533)
Income-to-Assets	0.0216326 (0.020358)	0.0174704 (0.0250993)	-0.0021435 (0.0448929)
1-Y Sales Growth	-0.0000027 (0.0000012)**	-0.00000255 (0.00000129)**	-0.0000873 (0.0000987)
Age	0.0038499 (0.0026267)	0.004297 (0.0029135)	-0.0016951 (0.005095)
Offer Year	0.002307 (0.0025494)	0.0024391 (0.0028627)	0.0033159 (0.0055307)

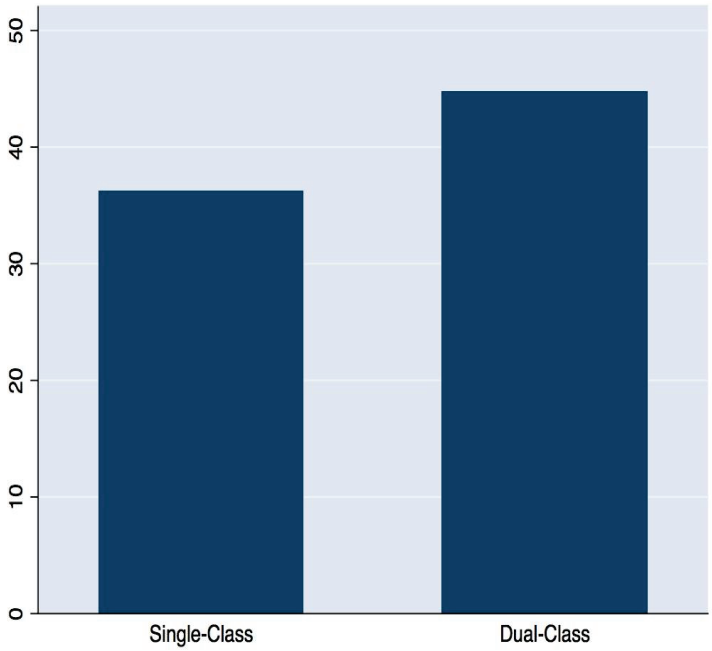
Note: This table reports estimated coefficients from OLS model, in which the dependent variable equals one for dual-class status, and zero otherwise. Robust standard errors are in parentheses; * significant at 10%; ** significant at 5%; *** significant at 1%

FIGURE 1. MEDIA COVERAGE FOR SINGLE- AND DUAL-CLASS FIRMS



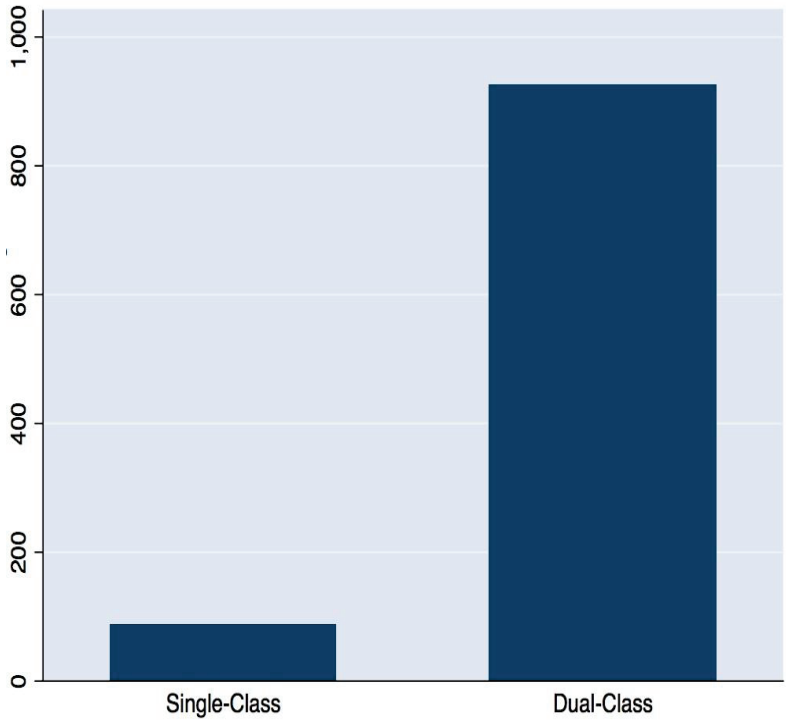
Note: The bars describe mean media coverage (max) on firms' founders prior to the IPO for dual-class and single-class structure firms

FIGURE 2. MEDIA COVERAGE FOR SINGLE- AND DUAL-CLASS FIRMS PRE-2004



Note: The bars describe mean media coverage (max) on firms' founders prior to the IPO for dual-class and single-class structure firms

FIGURE 3. MEDIA COVERAGE FOR SINGLE- AND DUAL-CLASS FIRMS POST-2004



Note: The bars describe mean media coverage (max) on firms' founders prior to the IPO for dual-class and single-class structure firms

File: Chaisse 17.1 To Publish_Edited Created on: 6/13/2022 10:40:00 PM Last Printed: 6/14/2022 9:26:00 PM

THE RISE AND FALL OF FREE TRADE AGREEMENTS: ANALYTICAL EVIDENCE FROM INDIA'S PRACTICE

Debashis Chakraborty, Julien Chaisse and Bibek Ray Chaudhuri[*]

INTRODUCTION

The inception of WTO in 1995 was expected to facilitate the process of free trade in merchandise products and movement of factors of production across borders. Over 2001-18, global exports of merchandise products has increased from USD 6.12 trillion to USD 19.28 trillion, underlining the success of the WTO-led reforms. Nevertheless, multiple barriers on merchandise[1] and services[2] trade as well as investment flows[3] remain a reality. Currently the WTO Doha Round negotiations, launched in 2001, are progressing at a slow pace, given the difference in perspectives among the Member countries.[4] The attractiveness of the multilateral body suffered a further setback in recent past, when the US, under the Trump Administration, decided to block the appointment of new Appellate Body jurists to record its discontent over the 'persistent overreaching' of the forum, by regularly 'interpreting WTO agreements in ways not envisioned by the WTO Members who entered into those agreements'[5]

Given this background, despite significant deepening of WTO reforms, the attractiveness and spread of regional trade agreements (RTAs) have intensified since launch of the Doha Round. The 'new generation' RTAs aspired to go beyond simple tariff reforms, attaching greater focus on mutual recognition of standards, Singapore Issues and several other WTO-Plus

[*] Debashis Chakraborty is Associate Professor at the Indian Institute of Foreign Trade (IIFT). The author can be reached at: debashis@iift.edu. Julien Chaisse is a Professor at the City University of Hong Kong (CityU), School of Law and President, Asia Pacific FDI Network (APFN). The author can be reached at: julien.chaisse@cityu.edu.hk. Bibek Ray Chaudhuri is Associate Professor at the Indian Institute of Foreign Trade (IIFT). The author can be reached at: brchaudhuri@iift.edu. The opinions expressed herewith are the authors' own.

[1] Josh Ederington & Michele Ruta, *Non-Tariff Measures and the World Trading* System 69, 81 (World Bank, Policy Working Paper No. 7661, 2016).

[2] Hildegun K. Nordås & Dorothée Rouzet, *The Impact of Services Trade Restrictiveness on Trade Flows*, 40 WORLD ECONOMY 1155, 1155 (2017).

[3] Stephen Thomsen & Fernando Mistura, *Is investment protectionism on the rise? Evidence from the OECD FDI Regulatory Restrictiveness Index* (OECD, Paris), 2017, at 1.

[4] Robert Wolfe, *First diagnose, then treat: what ails the Doha Round?* 10-11 (2013) (Robert Schuman Centre for Advanced Studies, EUI working paper).

[5] OFF. OF THE U.S. TRADE REPRESENTATIVE, REPORT ON THE APPELLATE BODY OF THE WORLD TRADE ORG. 1, 74 (Feb. 2020).

provisions.[6] The resulting RTAs have significantly facilitated intra-bloc trade flows and the associated benefits, e.g., deepening of the international production networks (IPNs) involving global capital inflows and rising efficiency of the regional players.[7] However empirical evidences show that the RTAs may not always lead to a welfare-enhancing outcome, particularly for the economically weaker members of the bloc.[8]

Encouraged by the prospects of the RTAs on one hand and worried by the slow progress of the WTO negotiations on the other, several countries have been drawn towards mega-regional trade agreements during the last two decades. The Trans-Pacific Partnership (TPP), involving the countries on both sides of the Pacific and Regional Comprehensive Economic Partnership (RCEP) involving ASEAN and its six FTA partners, deserve mention in this regard. Concerns have been raised whether the growth in these mega-regionals would in effect compete with the multilateral reform process and slow down the WTO negotiations further.[9] Interestingly in the recent period, we have witnessed two major pullouts from both these mega-regional negotiations, i.e., the USA from TPP in January 2017[10] and India from RCEP in November 2019[11].

It has been noted that the Indian standpoint on RCEP might have been influenced by the mismatch between expected and actual performance, i.e., less than expected export growth to partner countries, and a corresponding higher than anticipated growth of imports.[12] The current paper intends to analyze India's decision to not join RCEP in closer quarters, through the lenses of trade performance in three sectors, namely pharmaceuticals, textile and garments and automobile products. The underlying reason for focusing particularly on these three sectors are as follows. First, in compliance with the WTO provisions, India had to embrace the product patent regime from

---

6 Jo-Ann Crawford & Robert V. Fiorentino, *The Changing Landscape of Regional Trade Agreements* 5-16, (World Trade Organization, Discussion Paper No. 8, 2005).

7 Ayako Obashi and Fukunari Kimura, 'Deepening and Widening of Production Networks in ASEAN' (2016), Discussion Paper ERIA-DP-2016-09, Economic Research Institute for ASEAN and East Asia (ERIA): Jakarta.

8 Sarah Ellis Barnekow & Kishore G. Kulkarni, *Why Regionalism? A Look at the Costs and Benefits of Regional Trade Agreements in Africa*, 18 GLOBAL BUS. REV. 99, 99 (2017).

9 Shujiro Urata, *Mega-FTAs and the WTO: Competing or Complementary?*, 30 INT'L ECON. J. 231, 231-42 (2016); Chad P. Bown, *Mega-Regional Trade Agreements and the Future of the WTO* 1 (Council on Foreign Relations, discussion paper)..

10 Press Release, Office of the U.S. Trade Representative, The United States Officially Withdraws from the Trans-Pacific Partnership (2017), https://ustr.gov/about-us/policy-offices/press-office/press-releases/2017/january/US-Withdraws-From-TPP..

11 Press Release, Press Information Bureau, Ministry of Commerce & Industry, India exploring trade agreements with USA & E U; FTAs with Japan, Korea, & ASEAN being reviewed; No trade agreements in a hurry says Piyush Goyal (Nov. 5, 2019), https://pib.gov.in/newsite/PrintRelease.aspx?relid=194281.

12 Biswajit Dhar, *India's Withdrawal from the Regional Comprehensive Economic Partnership*, 54 ECON. & POL'Y WKLY. 59, 59-65 (2019).

January 1, 2005 onwards, graduating from the process patent stage. This transition was expected to facilitate innovation and competitiveness in the key Indian pharma players.[13] Second, as per the provisions of the Multi-Fibre Arrangement (MFA), from 1995 onwards the quota on textile and garment products were getting gradually phased out. [14] Since January 1, 2005, all the restrictions were removed, and the move was expected to enhance the exports from the developing countries (e.g., India) characterized by greater competitiveness.[15] Third, while the tariff rates on the automobile products declined steadily from 1995 onwards as result of both multilateral and unilateral reforms, one of the major hindrance to sectoral trade has been the existence of differing standards across countries. India joined one of the standards, the United Nations Economic Commission for Europe (UNECE) 1998 agreement in April 2006, and henceforth started aligning its domestic automobile standards with the UN global technical regulations (GTRs).[16] Given these developments, India is expected to have gained a better competitive edge in these product categories from 2005-06 onwards, which in turn, should have been reflected in the country's trade performance in general and with RCEP partners in particular. In addition, all the three sectors are part of the 'Make-in-India' initiative[17], which was launched in 2014 with the goal of enhancing domestic production, attracting foreign investment and technology transfer, as well as boosting export competitiveness.[18]

The paper is arranged along the following lines. First, India's standpoints during the RCEP negotiations are briefly mentioned. The overall trade performance of the country vis-à-vis the RCEP partners is noted next. The following section discusses the Indian policy frameworks involving the three selected sectors. India's trade performance in these three sectors are analyzed next by exploring the presence of structural breaks in both export and import figures and their possible explanations. The observed trade outcome with RCEP partners, i.e., the trade balance scenario, is scrutinized next with the help of tariff barriers and import growth rates. Finally, based on the findings, certain policy conclusions are drawn.

---

13 Aman Raj Khanna & Hemant Krishan Singh, *India's IPR Regime: Reconciling Affordable Access with Patent Protection*, STRATEGIC STUDIES PROGRAMME (Indian Council for Research on International Economic Relations, New Delhi), Aug. 2015, at 5.

14 Meenu Tewari, *Post-MFA Adjustments in India's Textile and Apparel Industry: Emerging Issues and Trends* 12-13, 32 (India Council for Research on International Economic Relations, Working Paper No. 167, 2005).

15 *Id*.

16 Marathe, S. (2008), 'WP 29 1958 agreement: a view from India', Presentation at the Special Session to commemorate the 50th Anniversary of the 1958 agreement, Geneva.

17 *Make in India: The Vision, New Processes, Sectors, Infrastructure and Mindset*, MAKE IN INDIA (2017), http://web.archive.org/web/20210112013324/http://www.makeinindia.com/article/-/v/make-in-india-reason-vision-for-the-initiative.

18 *Id*.

India's Gradual Move towards RTAs and Participation in RCEP

Southeast and East Asia witnessed emergence of several trade blocs since the nineties, namely – Association of Southeast Asian Nations (ASEAN), Free Trade Agreement (FTA) (1992), Sino-ASEAN FTA (2005), South Korea-ASEAN FTA (2007), Japan-ASEAN FTA (2008), Australia-New Zealand-ASEAN FTA (2010) and so on.[19] On the contrary, India primarily relied on the multilateral trade reforms for export growth in the initial years of WTO, though the country joined regional partnerships in South Asia, namely - SAARC Preferential Trading Arrangement (SAPTA) in 1993[20] and India-Lanka Free Trade Agreement (ILFTA) in 2001[21], guided by both trade and non-trade objectives, during this period. However, concerns over future export potential grew after the WTO Members failed to reach a common ground at the Doha Ministerial (2001) and Cancun Ministerial (2003) meetings. From 2004, India started actively pursuing the RTA route for export promotion.[22] Subsequently the country entered into the Early Harvest Programme (EHP) under the Indo-Thai FTA (2004), Indo-Singapore Comprehensive Economic Cooperation Agreement (CECA) (2005), ASEAN-India FTA (2010), India-South Korea Comprehensive Economic Partnership Agreement (CEPA) (2010), India-Japan CEPA (2011) and Indo-Malaysia CECA (2011).[23] It also joined the negotiations for India-New Zealand FTA (2010), India-Australia FTA (2011), Indo-Indonesia CECA (2011) and Bay of Bengal Initiative for Multi-Sectoral Technical and Economic Cooperation (BIMSTEC) FTA (2014), respectively. India's growing urge for integration with the East and Southeast Asian countries became apparent, when in 2014 the country replaced the existing 'Look East Policy' (launched in 1991) with the 'Act East Policy'.[24]

ASEAN and its six bilateral FTA partners launched the RCEP forum to establish a single trade bloc architecture in November 2012. RCEP aimed to minimize possible grievances among members during the trade reform by

---

19 Innwon Park, *Regional Trade Agreements in East Asia: Will They Be Sustainable?*, 23 ASIAN ECON. J. 169, 172-73. (2009); Misa Okabe, *Impact of Free Trade Agreements on Trade in East Asia* 1-2 (Economic Research Institute for ASEAN and East Asia, Discussion Paper No. 2015-01, 2015).

20 Michael Ewing-Chow & Md. Rizwanul Islam, *South Asian Free Trade Agreement and the Possibility of Regional Integration within the SAARC: A Historical, Legal and Economic Analysis*, 2 ASIAN J. COMPARATIVE L. 1, 3-4 (2007).

21 SUBRATA K. BEHERA, ET AL., DEEPENING ECONOMIC COOPERATION BETWEEN INDIA AND SRI LANKA 2 (Indra Nath Mukherji & Kavita Iyengar, eds. 1st ed., 2013).

22 Debashis Chakraborty and Dipankar Sengupta, 'Learning through Trading? India's Decade Long Experience at WTO', (2005) 12(2) South Asian Survey 223-246.

23 Julian Chaisse & Debashis Chakraborty & Biswajit Nag, *The Three-Pronged Strategy of India's Preferential Trade Policy: A Contribution to the Study of Modern Economic Treaties*, 26 CONN. J. INT'L L. 415, 420 (2011).

24 Amitendu Palit, *India's Act East Policy and Implications for Southeast Asia*, SOUTHEAST ASIAN AFFAIRS 81, 81 (2016).

taking into consideration the development divergence and, accordingly, through technical cooperation to solve the problems of the developing members.[25] Negotiations for creation of cohesiveness were concluded before finalizing the trade in goods provisions, given the dominance of SME firms and diverging internationalization level in constituent developing economies.[26] A cooperative RECP means a powerful boost to the rules-based global trading system and the new form of leadership in regional and world trade.[27] It encompasses first-ever ambitious agreements negotiated by the developing countries, covering China, India, Japan and South Korea, and can gradually deepen trade links and the associated familiarity effects through the existing bilateral agreements.[28] However, several practical concerns from Indian perspective slowed RCEP progress.[29]

First, as the negotiation aimed for high level of tariff liberalization both in agriculture and manufacturing segments, it became clear that India, whose average tariff is relatively higher, would face significant challenges from the reforms.[30] For instance, given the growing imports from China in manufacturing sectors, the domestic considerations in India were aggravated.[31] Similarly, the demand for deep tariff reforms in agricultural sectors (e.g., wheat, dairy) from Australia and New Zealand put India in a difficult position.[32] India's initial offer to have country-specific tariff reform schedules was rejected by all the RCEP partners, and the country was forced to re-submit a single tariff reform plan for all countries.[33] In addition, the missing linkages in the trade agreement architecture required time to establish; it therefore has

---

25 *See* ASEAN Secretariat, *Guiding Principles and Objectives for Negot. the Reg'l Comprehensive Econ. P'ship*, May 2012, at 1 https://asean.org/wp-content/uploads/2012/05/RCEP-Guiding-Principles-public-copy.pdf.

26 *See generally*, Junji Nakagawa, *Investment in the Trans-Pacific Partnership: Possible Impacts on ASEAN Member States (chapter 1),* in EMERGING GLOBAL TRADE GOVERNANCE 15-54 (Lurong Chen, Shujiro Urata, Junji Nakagawa and Masahito Ambashi (eds) 2019).

27 *See* Yoshifumi Fukunaga, *ASEAN's Leadership in the Reg'l Comprehensive Econ. P'ship*, ASIA & THE PACIfiC POLICY STUDIES, 2(1): 103, 104 (2015).

28 Julien Chaisse & Richard Pomfret, *The RCEP and the Changing Landscape of World Trade: Assessing Asia-Pacific Investment Regionalism Next Stage*, 12 L. & DEV. REV. 1 159, 159–90 (2019) (explaining RCEP genesis and membership as well the expected consequences on international trade and investment).

29 Sanschita Basu Das, *The Regional Comprehensive Econ P'ship: New Paradigm or Old Wine in a New Bottle?*, 29(2) ASIAN-PACIFIC ECONOMIC LITERATURE, 68, 73 (2015).

30 See Dhar (2019), *op. cit.*; Akarsh Bhutani, *India's Reluctance in Joining the RCEP — A Boon or a Bane in the Long-Run?*, OBSERVER RESEARCH FOUNDATION (Feb. 10, 2021), https://www.orfonline.org/expert-speak/india-reluctance-joining-rcep-boon-bane-long-run/.

31 Rahul Mishra, *RCEP: Challenges and Opportunities for India*, RSIS COMMENTARIES NO.140, 1, July 25, 2013.

32 Ganeshan Wignaraja, *What does RCEP mean for insiders and outsiders? The Experiences of India and Sri Lanka*, ARTNET WORKING PAPER NO. 181, July 2018 at 11-12.

33 *See* Amiti Sen, *Time for India to Exit RCEP Trade Pact*, HINDU BUS. LINE, (Mar. 9, 2018), www.thehindubusinessline.com/opinion/time-for-india-to-exit-rcep-trade-pact/article22134775.ece1.

been difficult to arrange a satisfactory tariff reform between China and India without any existing preferential agreement.[34]

Second, the recent mega-regional trade negotiations have focused on strengthening of the IPR regime. These negotiations often moved towards TRIPS-Plus provisions and mainly focused on two areas: protecting corporate patent rights and maintaining data exclusivity.[35] For instance, the data exclusivity provisions discussed in TPP may delay the time-to-market for generic drugs and increase their retail price.[36] India, therefore, was initially much more comfortable when RCEP discussed only how to reduce IPR-related obstacles to trade and investment among members.[37] However, RCEP members later evolved an adverse proposal for India after Japan and South Korea began demanding more business-friendly IPR provisions in the bloc, which may hinder the access to medicines and public health policy choices.[38] This 'business-friendly' proposal might have two underlying points, as Japan and South Korea, like Australia, New Zealand and Singapore, are Anti-Counterfeiting Trade Agreement (ACTA) and RCEP members. First, proposed RCEP text on border enforcement is TRIPS-Plus in nature, thereby targeting the goods (particularly generic medicines) in transit instead of the provisions against trademark counterfeiting and copyrights piracy (in line with WTO code).[39] Second, the text requires simply 'any information' for border enforcement, much lower requirements than TRIPS provisions, which may facilitate possible misuse of the same.[40] Such provisions might deter India's export potential of generic medicines, vis-a-vis firm located in Japan and South Korea, who are already part of ACTA.[41] This is the

---

34 *See* Rajeev Jayaswal, *At RCEP Meet, India Seeks 10% Advantage over China in Tariff Removal*, HINDUSTAN TIMES, (Oct. 19, 2019), https://www.hindustantimes.com/india-news/at-rcep-meet-india-seeks-10-advantage-over-china-in-tariff-removal/story-PhJgyiW3OcrxbmCPmMPXsJ.html

35 *See* Medecins Sans Frontier, *Regional Comprehensive Economic Partnership: Intellectual Property Chapter and the Impact on Access to Medicines*, at https://msfaccess.org/sites/default/files/IP_RCEP_%20IPChapteranalysis_MSF_Access%20to%20Medicines_2016_ENG.pdf (November 2016).

36 *See id.*

37 Young-Chan Kim, *Chinese Global Prod. Networks in ASEAN, Understanding China*, (Young-Chan Kim eds, 2016), 19, 24.

38 *See* Rachel Tansey, Sam Cossar-Gilbert, et al, *RCEP: A secret deal*, TRANSNATIONAL INSTITUTE (July 2018) at 11. https://www.tni.org/files/publication-downloads/foe-rcep-secret-deal-2-web.pdf (Japan's stance might be guided by perspectives from business groups like Keidanren, whose members include pharma majors from both Japan (e.g., Takeda and Astellas Pharma) and the global theatre (e.g., Pfizer and Sanofi)).

39 *See* Belinda Townsend et. al., *Regional Comprehensive Economic Partnership: Intellectual Property Chapter and the Impact on Access to Medicines*, 28(8) ASIA PACIFIC JOURNAL OF PUBLIC HEALTH, 682, 689 (2016).

40 Belinda Townsend, Deborah Gleeson, et al,*Regional Comprehensive Econ. P'ship: Intellectual Property Chapter and the Impact on Access to Medicines*, 28(8) ASIA PACIFIC JOURNAL OF PUBLIC HEALTH, 682, 689 (2016).

41 Op cit, Medecins Sans Frontier.

reason why India has been reluctant to accommodate strong TRIPS-plus provisions in RECP.[42]

Finally, it has been acknowledged that while India's gain from RTA with 'East' is expected to be comprehensive, opportunities under services would be substantially effected.[43] India has, accordingly, attempted to liberalize trade in services within RCEP in areas of interest, e.g., easing visa procedures on movement of skilled professionals for short-term work.[44] However enthusiasm in RCEP partners remained low, with limited number of offers in response to India's requests.[45] ASEAN members were particularly concerned over loss of local jobs from allowing temporary movement of Indian service providers.[46] This considerably lowered the incentive for India to join the bloc.

Despite these interim hiccups, the negotiations progressed forward and the 2018 Joint Leaders' Statement announced conclusion of negotiations on provisions for five key provisions, namely: Customs Procedures and Trade Facilitation (CPTF), Government Procurement, Institutional Provisions, Sanitary and Phytosanitary (SPS) Measures and Standards, Technical Regulations and Conformity Assessment Procedures (STRACAP).[47] However, in 2019 while India continued to participate at the negotiations, the rift widened with respect to the concerns over preferential tariff reforms[48], and their implications for dairy[49] and industrial sectors.[50] Finally in November 2019, on the eve of East Asia and RCEP Summit in Bangkok, India decided not to finalize the negotiations in a hurry, citing economic interests and national

---

42 Amiti Sen, *India on stronger footing to resist IPR framework at RCEP*, HINDU BUSINESS LINE, (13 January, 2018) at https://www.thehindubusinessline.com/economy/india-on-stronger-footing-to-resist-ipr-framework-at-rcep/article9557719.ece.

43 *See* Neha Raman, *Strengthening ASEAN-India P'ship: Trends and Future Prospects* EXPORT-IMPORT BANK OF India,at 59-80, (2018).

44 Arun, S, *India to Flag Worry on the Pace of Services Talks at RCEP*, THE HINDU, Sept. 6 2017, at https://www.thehindu.com/business/india-to-flag-worry-on-pace-of-services-talks-at-rcep/article19625890.ece.

45 *See* Sen (2017), *op. cit*.

46 *See* Wignaraja (2018), *op. cit*.

47 *See* Press Release, ASEAN Secretariat, Joint Leaders' Statement on the Regional Comprehensive Econ. P'ship (RCEP) Negotiations, (Nov. 14, 2018) https://www.meti.go.jp/press/2020/11/20201115001/20201115001-1.pdf. (accessed March 23, 2020).

48 *See* Amiti Sen, *India Should Act Tough and Exit RCEP*, HINDU BUSINESS LINE (Aug. 6, 2019), https://www.thehindubusinessline.com/opinion/india-should-act-tough-and-exit-rcep/article28826387.ece.

49 *See RCEP Deal Would Hurt Dairy Sector Heavily, Cautions SJM*, HINDU BUSINESS LINE (Oct. 5, 2019), https://www.thehindubusinessline.com/economy/sjm-urges-modi-government-to-oppose-dairy-sector-inclusion-in-rcep/article29575133.ece.

50 Vatasala Gaur,, *Inclusion of steel in RCEP talks worries industry captains*, Economic Times, (Sept. 19, 2019), At https://economictimes.indiatimes.com/news/economy/foreign-trade/inclusion-of-steel-in-rcep-talks-worries-industry-captains/articleshow/71193969.cms.

priorities.[51] In November 2020, the RCEP countries decided to conclude negotiations without India, but assured the country an observer membership anytime to formally begin the process of joining the bloc.[52] However the Indian External Affairs Minister rationalized the pull-out decision of the country on the grounds of unaddressed concerns.[53]

India-RCEP Trade Trends

The evolving importance of India and RCEP in each other's trade basket over 2001-18 can be observed from Tables 1 and 2. The tables have been constructed by obtaining data from Trade Map database.[54] For interpreting the transitions in trade flows in light of policy changes, the time period has been divided in four ranges, namely: 2001-04 (India's reliance on multilateral reforms), 2005-09 (inclination towards RTA participation), 2010-13 (deepened RTA participation) and 2014-18 (RCEP negotiation phase). Table 1 shows how India's export and import flows have gradually oriented towards RCEP markets (in percentage terms).[55] A few interesting observations deserve mention. First, the importance of ASEAN in India's export basket has increased in line with expectations.[56] The rise in India's export share to ASEAN can be particularly attributed to growing orientations towards Myanmar and Vietnam. Second, the export shares to Singapore and Malaysia, ASEAN partners with whom India is integrated through bilateral RTAs as well, behaved differently during 2014-18. While export share to Malaysia increased during the last reported period, the corresponding figure for Singapore declined.[57] Third, the shares of Indian exports to the other two existing RTAs, namely, Japan and South Korea, have increased only incrementally.[58] Fourth, export shares to Australia and New Zealand, with whom RTA negotiations are in progress, have also registered a modest rise.[59] Five,

51 *See*, Press Information Bureau, Government of India (2019), *op. cit*.

52 For details, see *Ministers' Declaration on India's Participation in the Regional Comprehensive Economic Partnership*, available at: https://www.meti.go.jp/press/2020/11/20201115001/20201115001-3.pdf (11 November 2020).

53 *See India Pulled out of RCEP as Concerns Not Addressed: EAM S Jaishankar*, BUSINESS STANDARD (Nov. 18, 2020, 3:07 PM), https://www.business-standard.com/article/economy-policy/india-pulled-out-of-rcep-as-concerns-not-addressed-eam-s-jaishankar-120111801459_1.html.

54 For analysis the data has been drawn from the following source: International Trade Centre, *Trade Map Database*, http://www.trademap.org/ (last visited August 12, 2021). The database contains the trade data as collected and reported by the designated authorities in each country. Based on the analysis with *Trade Map* data, the tables in the current paper has been prepared.

55 *See infra Table 1.*

56 Debashis Chakraborty, *The Upcoming Indo-ASEAN CECA: Of Great Expectations and Areas of Concern*, 50(3) CHINA REP., 259, 261–62, 273-74 (2014).

57 *See infra Table 1.*

58 *Id.*

59 *Id.*

proportional importance of China as an export market for India is on decline since 2009.[60] A major underlying reason is that India's export basket to China is dominated by primary products.[61]

A few interesting observations evolve from RCEP's presence in India's import basket as well. First, the share of ASEAN in India's import basket remained largely unchanged during the first three periods, and increased sharply after 2014, which can be part attributed to the phased tariff reforms under India-ASEAN FTA.[62] Second, rise in import from ASEAN does not originate from the bilateral RTA partners Singapore and Malaysia, but from Brunei, Indonesia, Thailand and Vietnam. While the average import shares from Brunei remained almost constant during the last two periods, the corresponding figures increased sharply for other three countries during 2014-18.[63] The export success of ASEAN countries can be attributed to their comparative advantage in terms of competitive wage.[64] Interestingly, in more recent period while in several ASEAN countries competitiveness has increased in high-skill intensive sectors, the same has declined in low-skill intensive segments, raising the concerns of getting caught in 'Middle-income trap'.[65] However, as India had been less successful in appropriately harvesting its comparative advantages in lower-skilled, labour-intensive industrial categories, the economic pressure in more on it.[66] Third, interestingly the import share from Japan declined, while the same from South Korea oscillated over the period.[67] Fourth, import shares from Australia and New Zealand are showing a decreasing and increasing trend respectively.[68] Five, proportional importance of China as an import source has increased quite sharply over the entire period, given the rising technology-intensity of the country's export

---

60 *Id.*

61 *See* S. K. Singla, *An Analysis of India's Export Performance with China*, 50(3) *FOREIGN TRADE REV.*, 219, 223 (2015) ("India's exports to China, [between 1990 and 2010], were mainly dominated by one single commodity group, namely, ores, slag and ash.").

62 *See infra* Table 1; *cf.* H.A.C. Prasad, *Reviving and Accelerating India's Exports: Policy Issues and Suggestions* viii–ix, 37, 39, (Dep't of Econ. Aff., Working Paper No. 1/2017-DEA, Jan. 2017) ([explanatory parenthetical]).

63 *See infra Table 1.*

64 For details on Average Monthly wages in ASEAN countries and India during RCEP negotiations, see: International Labour Organization, *Wages in Asia and the Pacific: Dynamic but uneven progress*, Global Wage Report 2014/15 | Asia and the Pacific Supplement (December 2014), available https://www.ilo.org/wcmsp5/groups/public/---asia/---ro-bangkok/---sro-bangkok/documents/publication/wcms_325219.pdf.

65 Tran, V.T. 2013. The Middle-Income Trap: Issues for Members of the Association of Southeast Asian Nations. ADBI Working Paper 421. Tokyo: Asian Development Bank Institute. Available: http://www.adbi.org/working-paper/2013/05/16/5667.middle.income.trap.issues.asean/.

66 John West, *India squanders its comparative advantage*, 10 July 2020, East Asia Forum, available at: https://www.eastasiaforum.org/2020/07/10/india-squanders-its-comparative-advantage/.

67 *See infra Table 1.*

68 *Id.*

basket.[69] On the whole, it is evident from Table 1 that RCEP partner's collective presence in India's import basket has increased sharply in comparison with the corresponding export figures.

Table 2 depicts how India's presence has evolved in the export and import baskets of RCEP partners (in percentage terms). The emerging conclusions from the export flows are noted sequentially in the following sentences. First, India's presence has increased in the export basket of all ASEAN countries, irrespective of their development levels.[70] The rise has been particularly noticeable for Brunei, Indonesia, Malaysia, Thailand and Vietnam.[71] The deeper orientation towards India in Vietnam's export basket is understandable, given the country's growing manufacturing competitiveness.[72] Second, India's presence grew in the export baskets of all five FTA partners of ASEAN, with the rise being sharpest for China.[73] This depicts growing importance of India in the export decision of partners, given its market size and GDP growth rate.[74] On the other hand, India's presence in RCEP partner's imports reveal interesting results. First, India's presence has increased in the import basket of all ASEAN countries, barring the exception of Cambodia and Laos, both of whom being LDCs, opened markets to India gradually.[75] Second, India's presence has deepened in the import basket of all FTA partners of ASEAN, barring the exception of China.[76] The result can be part attributed to the fact that unlike other FTA partners of ASEAN, India is not integrated with China through a vibrant FTA relationship; although both China and India are part of Asia-Pacific Trade Agreement (APTA), preferences granted through this arrangement are not widespread.[77]

The growing trade association of India with RCEP partners is evident from Tables 1 and 2, but a few areas of concern simultaneously emerge. First, it is apparent that while China's export orientation from India is on the rise, a corresponding import orientation has been missing.[78] Second, while

---

69 *See generally* Pravakar Sahoo and Abhirup Bhunia, *China's Manufacturing Success: Lessons for India* 134–140 (Inst. Of Econ. Growth Working Paper No. 344, 2014) (explaining China's federal direct investments and regional initiatives in technological growth and their influence on India).

70 *See infra* Table 2.

71 *Id*.

72 Nguyen Thi Tue Anh, Luu Minh Duc, and Trinh Duc Chieu, *The Evolution of Vietnamese industry* 24 (U.N. Univ. - World Inst. for Dev. Econs. Rsch. Working Paper No. 2014/076, 2014).

73 *See infra* Table 2.

74 *C.f.* POONAM GUPTA, WORLD BANK, INDIA DEVELOPMENT UPDATE: INDIA'S GROWTH STORY 88(2018) (discussing India's GDP and export growth, and the expected continued acceleration of growth).

75 Prasad, *supra* note 55 at .; *Cf.* DAVID SINATE ET AL. EXP.-IMP. BANK OF INDIA STRENGTHENING ASEAN-INDIA PARTNERSHIP: TRENDS AND FUTURE PROSPECTS 23–32 (2018).

76 *See infra* Table 2.

77 *See generally*, Zhang Yunling, *People's Republic of China* in ASIA'S FREE TRADE AGREEMENTS: HOW IS BUSINESS RESPONDING? 106-129 (Masahiro Kawai & Ganeshan Wignaraja eds., 2011) (discussed proposed FTAs between China and India that would include thirty-three firms, but the support for the agreement has not yet been sufficient).

78 *See infra* Tables 1 and 2.

Singapore, Japan, Malaysia and South Korea have become India's bilateral trade partners, their trade orientation with India is also not necessarily deepening.[79] India's modest trade orientation with ASEAN underlines the country's limited presence in Asian Integrated Production Networks (IPNs).[80] Third, import from RCEP partners has risen sharply after 2014-18, reflected in diverging trade balance,[81] which, in turn, might have acted as a possible stress during negotiations. This imbalance in regional orientation pattern is explained by lower technology-intensity[82] and modest competitiveness pattern of Indian exports, vis-à-vis RCEP partners on one hand[83] and possible trade barriers in ASEAN markets on the other[84].

79 *See* Chaisse, *supra* note 23.

80 *See* Biswajit Nag, *Emerging Production Network between India and ASEAN: An Analysis of Value-Added Trade in Select Industries* in TRADE, INV. AND ECON. DEV. IN ASIA: EMPIRICAL AND POL'Y ISSUES 41(Debashis Chakraborty & Jaydeep Mukherjee, eds., 2016) (describing India's engagement in Asian production networks as slow, steady).

81 The worsening of trade balance for India vis-à-vis RCEP is particularly noticeable in case of manufacturing sectors. *See*, Debashis Chakraborty and Julien Chaisse, *Tightrope Walk between Faith and Skepticism: India's 'Contingency Plan' for Free Trade*, 15 ASIAN J. OF WTO & INT'L HEALTH L. AND POL'Y, 91, 150-53 (2020).

82 *See* Rahul Anand, Kalpana Kochhar, and Saurabh Mishra, *Make in India: Which Exports Can Drive the Next Wave of Growth?*, Figure 5 (Int'l Monetary Fund Working Paper No. 15/119, 2015) (showing low-tech manufacturing has been at or near fifty percent of total manufacturing exports in India, the highest percentage of countries surveyed).

83 *See* Namita Kaur and Vishal Sarin, *Comparative Advantages and Competitiveness of Indian Agricultural Products Exports to ASEAN in Context of India's Look East Policy*, 13 *INT'L J. OF AGRIC. AND STAT. SCI.*,, 159, Figure 1 (2017) (showing that India's revealed comparative advantages in 2015 are largely similar across sampled industries to where they stood in 2001).

84 *See* LILI YAN ING, SANTIAGO FERNANDEZ DE CORDOBA, AND OLIVIER CADOT, NON-TARIFF MEASURES IN ASEAN, (Econ. Rsch. Inst. for ASEAN and E. Asia et al. eds., 2016) (observed technical barriers to trade as the most prevalent alternative to tariffs in ASEAN countries as of 2015).

Table 1: Importance of RCEP Partners in India's Total Trade Flows (Percentage Share)

Partner Countries	Export Share (%)				Import Share (%)			
	2001-04	2005-09	2010-13	2014-18	2001-04	2005-09	2010-13	2014-18
Brunei	0.007	0.014	0.080	0.015	0.001	0.104	0.142	0.142
Cambodia	0.029	0.029	0.034	0.048	0.001	0.001	0.002	0.009
Indonesia	1.491	1.475	1.980	1.288	2.289	2.255	2.971	3.419
Laos PDR	0.005	0.006	0.009	0.014	0.000	0.000	0.019	0.031
Malaysia	1.496	1.434	1.454	1.762	2.361	2.239	1.958	2.261
Myanmar	0.139	0.116	0.169	0.355	0.554	0.368	0.290	0.224
Philippines	0.621	0.435	0.380	0.518	0.180	0.108	0.100	0.127
Singapore	3.058	4.715	4.548	3.193	2.523	2.649	1.735	1.953
Thailand	1.301	1.078	1.083	1.193	0.744	0.922	1.130	1.428
Vietnam	0.612	0.847	1.329	2.238	0.051	0.107	0.406	0.903
Australia	0.923	0.767	0.766	1.130	2.799	3.636	2.831	2.606
China	3.721	6.308	5.855	4.116	4.818	9.746	11.484	15.227
Japan	3.104	2.162	2.107	1.597	3.367	2.583	2.392	2.442
New Zealand	1.170	1.865	1.475	1.434	2.789	2.820	2.743	3.310
South Korea	0.132	0.179	0.090	0.113	0.124	0.148	0.158	0.135
ASEAN	8.759	10.148	11.067	10.624	8.703	8.753	8.752	10.498
ASEAN + Japan + South Korea	13.034	14.176	14.649	13.656	14.859	14.157	13.887	16.251
Australia + New Zealand (ANZ)	1.054	0.946	0.856	1.243	2.923	3.784	2.989	2.741
RCEP	17.809	21.430	21.360	19.015	22.600	27.687	28.360	34.218

Source: Constructed by the authors from Trade Map data

Table 2: Importance of India in RCEP Partners' Total Trade Flows (Percentage Share)

Partner Countries	Export Share (%)				Import Share (%)			
	2001-04	2005-09	2010-13	2014-18	2001-04	2005-09	2010-13	2014-18
Brunei	2.292	4.422	7.528	8.902	0.497	0.848	0.882	1.448
Cambodia	0.005	0.049	0.146	0.091	0.797	1.072	1.222	0.666
Indonesia	2.509	4.533	6.638	7.541	2.006	2.164	2.309	2.299
Laos PDR	-	-	0.009	0.804	-	-	0.321	0.229
Malaysia	2.119	3.231	3.771	3.926	0.940	1.438	1.938	2.554
Myanmar	-	-	14.963	6.624	-	-	3.402	4.805
Philippines	0.245	0.372	0.690	0.678	0.836	0.933	1.080	1.672
Singapore	2.094	3.262	3.202	2.951	1.155	2.241	3.114	2.052
Thailand	0.773	1.729	2.292	2.618	1.163	1.308	1.311	1.527
Vietnam	0.268	0.475	1.579	1.833	1.679	2.135	2.089	1.774
Australia	2.942	5.882	5.264	3.894	0.750	0.828	0.992	1.645
China	0.824	1.865	2.444	2.750	0.967	1.492	1.185	0.828
Japan	0.493	0.849	1.262	1.301	0.599	0.671	0.814	0.788
New Zealand	1.173	1.885	2.235	2.412	0.780	1.215	1.343	1.028
South Korea	0.579	1.119	1.783	1.271	0.584	0.684	0.860	1.108

Source: Constructed by the authors from Trade Map data

THE THREE SECTORS: THE EVOLVING POLICY FRAMEWORK IN INDIA

For the purpose of the current sectoral analysis, three product groups have been selected, namely – pharmaceuticals (HS 30), textile and clothing (HS 52-63), and automobiles (HS 87). The relative transition of these product groups in India's trade basket has been summarized in Table 3, which shows an interesting dynamic. First, it is observed that the importance of pharma products have steadily increased in India's export basket, while a marginal increase has been noted for the corresponding imports. Second, in cases of textile and garment products, the contributions of these product groups are gradually declining in India's export basket, while the import shares have changed only marginally. Finally, the shares of the automobile sector in India's trade basket have shown a rising trend. A brief discussion on trade-related policy formulations in each of these sectors in the WTO era are provided in the following sections.

Pharmaceuticals

Protection of intellectual property was ensured in India during the colonial period by the Patents and Designs Act (1911), which created a strong product patent regime. However, concerns over availability of essential drugs at an affordable price, given the monopoly right enjoyed by foreign pharmaceuticals on patented medicines, surfaced during the sixties.[85] The policy deliberations resulted in the Indian Patent Act (1970), where, through introduction of a process patent regime, Indian pharma companies were allowed to reverse engineer and produce the drugs invented elsewhere through a different process. While the policy enhanced domestic pharmaceutical production,[86] several foreign firms avoided Indian markets to protect their invention.[87] In the aftermath of WTO inception, USA[88] and EC[89] have challenged

85 For a perspective, see EPW Editorial, *No Room for Product Patents*, 50 ECON. & POL'Y WKLY. 108, 108 (1965).

86 Nilesh Zacharias & Sandeep Farias, *Patents and the Indian Pharmaceutical Industry*, BUSINESS BRIEFING: PHARMATECH 2002 (Nishith Desai Associates, Mumbai), Apr. 2002, at 44.

87 Aman Raj Khanna & Hemant Krishan Singh, *India's IPR Regime: Reconciling Affordable Access with Patent Protection*, STRATEGIC STUDIES PROGRAMME (Indian Council for Research on International Economic Relations, New Delhi), Aug. 2015, at 5.

88 United States v. India, WTO Dispute Settlement 50 IP/D/5, Patent Protection for Pharmaceutical and Agricultural Chemical Products, ¶ 4 (July 9, 1996); for details, *see* Julian Chaisse & Debashis Chakraborty, *Dispute Resolution in the WTO: The Experience of India, in* BEYOND THE TRANSITION PHASE OF WTO: AN INDIAN PERSPECTIVE ON EMERGING ISSUES 507, 521-22 (Dipankar Sengupta & Debashis Chakraborty & Pritam Banerjee eds., 2006).

89 European Communities v. India, WTO Dispute Settlement 79 IP/D/7, Patent Protection for Pharmaceutical and Agricultural Products, ¶ 4 (May 6, 1997); for details, *see* Julian Chaisse & Debashis

the level of market access in India on this front. As decided under the Uruguay Round negotiations leading to inception of the WTO,[90] India moved to a product patent regime from January 1, 2005 onwards, thereby extending protection to a firm's invented molecule, test data, and so on.[91] The introduction of Science and Technology Policy (STP) (2003) can be seen as a conscious step to encourage domestic firms to intensify research and development (R&D) efforts and convert them into innovations to prepare for the product patent regime.[92]

With the introduction of the product patent regime and re-entry of foreign pharma companies, R&D spending increased in Indian pharma majors, which led to development of improved generic drugs and expansion of retail formulations market.[93] India gradually emerged as a hub for R&D services for global corporates, either through contracts or collaborative programs.[94] The growing comparative advantage of the country can be noted from increased exports of both bulk drugs and formulations on the one hand,[95] and a rise in approvals for Indian medicines in foreign markets[96] on the other. To promote the culture of innovations further, the Science, Technology and Innovation Policy (STIP) was introduced in 2013, which aspired to facilitate greater private sector participation in R&D activities.[97] R&D expenses are showing an increasing trend in recent times, due to the growth of co-development arrangements between Indian and global pharma corporations.[98]

Chakraborty, *Dispute Resolution in the WTO: The Experience of India, in* BEYOND THE TRANSITION PHASE OF WTO: AN INDIAN PERSPECTIVE ON EMERGING ISSUES, *supra* note 79, at 521-22.

90 Muhammad Ijaz Latif, *Uruguay Round of GATT and Establishment of the WTO*, 65 PAK. HORIZON 53, 53 (2012).

91 Biswajit Dhar and K.M. Gopakumar, *Effect of Product Patents on the Indian Pharmaceutical Industry* 1-10 (Centre for WTO Studies, 2015), https://wtocentre.iift.ac.in/Papers/3.pdf. For additional information, *see* Senpathy Krisi Gopalakrishnan & Jibak Dasgupta, *Policies to Drive Innovation in India*, *in* THE GLOBAL INNOVATION INDEX 2015: EFFECTIVE INNOVATION POLICIES FOR DEVELOPMENT 121, 121-30 (Soumitra Dutta & Bruno Lanvin & Sacha Wunsch-Vincent eds., 2015).

92 Nirupa Sen, *Science and Technology Policy—2003*, 84 CURRENT SCI. 13, 13 (2003).

93 Sudip Chaudhuri & Chan Park & K.M. Gopakumar, FIVE YEARS INTO THE PRODUCT PATENT REGIME: INDIA'S RESPONSE 10 (United Nations Development Programme ed., 2010).

94 *Global Pharma Looks to India: Prospects for Ggrowth*, PHARMACEUTICALS AND LIFE SCIENCES (PricewaterhouseCoopers, London), 2010, at 26.

95 Shibanjan Dutta and Dhirendra Gajbhiye, *Drivers of Indian Pharmaceutical Exports*, available https://rbidocs.rbi.org.in/rdocs/Bulletin/PDFs/03AR_15072021B75D322EF39B4B3A83C9B918B459A759.PDF (RBI Bulletin July 2021).

96 Samrat Sharma, *Indian Pharma Firms Set to Earn More Revenue from US; Get Series of Approvals in Recent Months*, https://www.financialexpress.com/industry/indian-pharma-firms-set-to-earn-more-revenue-from-us-get-series-of-approvals-in-recent-months/2151939/ (December 17, 2020).

97 Government of India, Ministry of Science and Technology, New Delhi, SCIENCE, TECHNOLOGY AND INNOVATION POLICY 2013 11, 13 (2013).

98 *The Indian Pharmaceutical Industry: Business, Legal & Tax Perspective* (Nishith Desai Associates, Mumbai), 2019, at 45.

India has now emerged as "one of the major vaccine producers, and is a global leader in end-to-end drug manufacturing."[99]

To facilitate the ongoing process of innovation and trade, under the 'Make-in-India' initiative several policies have been introduced since 2014. First, for promoting R&D activities, 11 National Institutes of Pharmaceutical Education & Research (NIPERs) have been set up.[100] Second, 100 percent FDI has been allowed for manufacturing of medical devices (under automatic route), encouraging global majors to relocate production units in India.[101] Third, to enhance domestic value content in generic drug exports, indigenization of Active Pharmaceutical Ingredients (APIs) is being encouraged to lower imports of raw materials from China.[102] Fourth, to ensure qualitative improvement, the government has recently increased the validity of the World Health Organisation (WHO) Good Manufacturing Practices (GMP) certificate from two to three years[103] and formulated the Medical Devices Rules (2017) in line with Global Harmonization Task Force framework.[104] Nevertheless, major challenges for Indian pharma exports include presence of an inefficient logistic supply chain and associated higher logistic costs.[105]

Cotton, Textile and Garments

The textile and garments sector in India was protected by tariff during pre-1991 period, given its importance on domestic employment generation. Even after WTO inception, India retained certain quantitative restrictions on

---

99 *Indian Pharmaceutical Industry: Challenges and Prospects*, 176 OCCASIONAL PAPER (Export-Import Bank of India, Mumbai), Aug. 2016, at 11.

100 Rajya Sabha *Unstarred Question No. 1631*, GOVERNMENT OF INDIA DEPARTMENT OF INDUSTRIAL POLICY AND PROMOTION 1, 6 (answered by Shri C.R. Chaudhary, Minister of State in the Ministry of Commerce & Industry).

101 Shri Sadananda Gowda, MINISTRY OF CHEMICALS AND FERTILIZERS PRESS INFORMATION BUREAU *Endeavor of Government is to provide Affordable Quality Healthcare for All and boost Indigenous Pharma Sector through 'Make in India': Shri Sadananda Gowda Shri Sadananda Gowda to hold Roundtable of Pharma and Medical Devices CEOs to discuss Government policy and Challenges facing the Industry,* MINISTRY OF CHEMICALS AND FERTILIZERS PRESS INFORMATION BUREAU 1, 1(2019).

102 *See* Ani, *India will be one of top 5 global pharma innovation hubs by 2020 through its PPP model: study*, BUISNESS STANDARD (Aug. 29, 2016), https://www.assocham.org/newsdetail-print.php?id=5886.

103 Prabha Raghavan, *CDSCO increases WHO GMP certificate to three years for ease of doing business*, THE ECONOMIC TIMES (May 08, 2018), https://economictimes.indiatimes.com/industry/healthcare/biotech/healthcare/cdsco-increases-who-gmp-certificate-to-three-years-for-ease-of-doing-business/articleshow/64084982.cms?from=mdr.

104 Contentmii, *Sector Survey: Pharmaceuticals*, MAKE IN INDIA (Feb. 8, 2021), https://www.makeinindia.com/sector-survey-pharmaceuticals.

105 TNN, *Pharma sector must cut logistics costs to stay in race: Pharmexcil Chief*, THE TIMES OF INDIA (July 14, 2017), https://timesofindia.indiatimes.com/business/india-business/pharma-sector-must-cut-logistics-costs-to-stay-in-race-pharmexcil-chief/articleshow/59601274.cms.

this sector on Balance of Payments (BOP) ground, which were terminated only after losing the dispute to the US. While several other countries also challenged India on the ground of quantitative restrictions, they were settled prior to the panel decision.[106] With decline in import barriers, from late nineties, India started preparing for the MFA phase-out. Several ex-ante simulations analyses noted that China and India should be the prime beneficiaries of the MFA phase-out, given raw material and labour cost advantages.[107] However, the sector suffered from several bottlenecks as well, e.g., lack of capital-intensity at various stages of production, low capacity utilization, poor supply-chain etc.[108] The problems largely resulted from the government decision to reserve the sector as a small-scale industry for a long time,[109] which prevented the firms from reaching efficient economies of scale and consequently, higher level of competitiveness.[110] To complicate matters, several developed countries back-loaded their MFA phase-out schedule by not reforming most of high value items (i.e., garments and made-ups) till December 2004 and instead liberalizing the relatively low value items (e.g., yarn and fabrics) since 1995.[111] On the other hand, imposition of anti-dumping duties and other barriers on Indian exports continued to curb the effective market access.

In order to improve the competitiveness of the sector, a number of initiatives were introduced before MFA phase-out. First, the government considered the N. K. Singh Committee report, which recommended partial restructuring of loans in this sector. Second, a support of 5 percent on the loans taken from various bodies like Industrial Development Bank of India (IDBI) was provided under 'Technological Upgradation Fund', used only for new

106 DS96: India — Quantitative Restrictions on Imports of Agricultural, Textile and Industrial Products (Complainant: EU); DS94: India — Quantitative Restrictions on Imports of Agricultural, Textile and Industrial Products (Complainant: Switzerland); DS93: India — Quantitative Restrictions on Imports of Agricultural, Textile and Industrial Products (Complainant: New Zealand); DS92: India — Quantitative Restrictions on Imports of Agricultural, Textile and Industrial Products (Complainant: Canada); DS91: India — Quantitative Restrictions on Imports of Agricultural, Textile and Industrial Products (Complainant: Australia) and DS90: India — Quantitative Restrictions on Imports of Agricultural, Textile and Industrial Products (Complainant: US).

107 Hildegunn Kyvik Nordås, *The Global Textile and Clothing Industry Post the Agreement on Textiles and Clothing*, WTO SECRETARIAT (2004).

108 Nilanjan Banik and Saurabh Bandopadhyay, *Cotton Textile Industry in India, In the Wake of MFA Phase-Out* RAJIV GANDHI INSTITUTE OF CONTEMPORARY STUDIES, WORKING PAPER NO. 9, 1, 16. (2000).

109 Anindya Sen and Partha Ray, *The Ascent and Decline of Reservation in Indian Small Scale Industries: Evolution of the Policy Environment*, INDIAN INSTITUTE OF MANAGEMENT CALCUTTA, WORKING PAPER SERIES WPS NO. 759, 1, 16-17 (2015).

110 Samar Verma, *Export Competitiveness of Indian Textile and Garment Industry* INDIAN COUNCIL FOR RESEARCH IN INTERNATIONAL ECONOMIC RELATIONS, WORKING PAPER NO. 94, 1, 18-23 (2002).

111 *Id*., at 6, 15.

capacity creation.[112] Third, the decision to withdraw restriction on import of second hand machinery helped technological upgradation of mid and low-segment firms.[113] Fourth, the National Textile Policy (2000) and National Jute Policy (2005) were introduced to strengthen the sector. As a result of these initiatives, India witnessed a rise in exports in the post MFA phase-out period,[114] though the same has come down in recent period.

The major challenges for Indian players is the emergence of low-cost competitors like Bangladesh and Vietnam, long response time to global fashion trends,[115] poor skill-set of workers, inefficiencies across the value chain,[116] limited diversification of export basket[117] and so on. Taking note of the concerns, supports provided to this sector under the 'Make-in-India' initiative (2014) and other policies include: investment facilitation (100 per cent FDI under automatic route), technology upgradation support, integrated skill development scheme,[118] fiscal supports,[119] establishment of textile parks and steps to improve backward and forward integrations etc.

Automobiles

The automobile sector in India before 1991 was also protected through high tariff barriers.[120] While tariff reforms followed since launch of the economic reforms in 1991, policy interventions guided by the objective of trade balancing requirements were terminated only after losing disputes at the WTO. Since then India aspired to become a major player in global auto market through a set of coordinated policies, starting with National Auto Policy

112 IDBI, *Fund scheme for textile and Jute industries-TUF Scheme*, IDBI BANK, https://www.idbibank.in/tufs-textile-jute-industries.asp (accessed Mar. 22, 2020).

113 Anuj Bhagwati, *Indian Textile Machinery Industry: Current Scenario and Future Outlook*, THE TEXTILE MAGAZINE (October 2, 2011), https://www.indiantextilemagazine.in/indian-textile-machinery-industry-current-scenario-and-future-outlook/.

114 Meenu Tewari, *Post-MFA Adjustments in India's Textile and Apparel Industry: Emerging Issues and Trends* INDIAN COUNCIL FOR RESEARCH IN INTERNATIONAL ECONOMIC RELATIONS, WORKING PAPER NO. 167, 1, 16-17 (2005).

115 India, *supra* note 13.

116 *See* generally, Saon Ray & Smita Miglani, *Global Value Chains and the Missing Links: Cases from Indian Industry*. 1st Ed., Routledge (2018).

117 T. E. Narasimhan, *India's Garment Exports Stagnant on High Costs, Compliance Burden*, BUSINESS STANDARD, (Apr. 19, 2019) https://www.business-standard.com/article/economy-policy/india-s-garment-exports-stagnant-on-high-costs-compliance-burden-119041800484_1.html (Apr. 19, 2019).

118 *Textile and Garments*, MAKE IN INDIA, http://www.makeinindia.com/sector/textiles-and-garments (accessed Mar. 20, 2020).

119 PTI, *Government's Rs 6,000 cr Package to Boost Apparel Sector: Smriti Irani*, ECONOMIC TIMES, https://economictimes.indiatimes.com/industry/cons-products/garments-/-textiles/governments-rs-6000-cr-package-to-boost-apparel-sector-smriti-irani/articleshow/62676513.cms (Jan. 27, 2018).

120 Harsha Vardhana Singh, *Trade Policy Reform in India since 1991*, 1, 10-11 BROOKINGS INDIA, WORKING PAPER 2 (2017).

(2002). However, it was soon realized that participation in automobile integrated production networks (IPN) would deepen only when the domestic production can conform to global standards relating to vehicle and passenger safety. The harmonization of global regulations on this front are managed by the WP.29 Forum through various agreements under United Nations Economic Commission for Europe (UNECE), which has enormous implications on trade.[121] Currently there are three major agreements under UNECE, namely - Vehicle Regulations (1958)[122], Periodical Technical Inspections of Vehicles in use (1997)[123] and Global Technical Regulations (GTR) on Vehicles (1998).[124] India entered WP 29 as observer in 2003 and joined UNECE (1998) in April 2006[125], whereupon it accordingly modified Central Motor Vehicle Rules (CMVR), i.e., domestic standards, in line with UN GTRs[126] through a series of reforms, '... India has more than 70% safety regulations which are either partially or fully aligned with GTRs and UN Regulations while keeping in view the Indian specific driving and environmental conditions

It is expected that joining of a UNECE provision would result in freer trade in auto-components and final vehicles among member countries, given the reciprocity enshrined in the provisions.[127] India's decision to join UNECE

---

121 *WP.29 - Introduction*, UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, https://www.unece.org/trans/main/wp29/introduction.html (accessed May 14, 2019).

122 Agreement Concerning the Adoption of Harmonized Technical United Nations Regulations for Wheeled Vehicles, Equipment and Parts which can be Fitted and/or be Used on Wheeled Vehicles and the Conditions for Reciprocal Recognition of Approvals Granted on the Basis of these United Nations Regulations, ECONOMIC COMMISSION FOR EUROPE, https://www.unece.org/fileadmin/DAM/trans/main/wp29/wp29regs/2017/E-ECE-TRANS-505-Rev.3e.pdf (Oct. 20, 2017) [hereinafter Vehicle Regulations].

123 Economic Commission for Europe, AGREEMENT CONCERNING THE ADOPTION OF UNIFORM CONDITIONS FOR PERIODICAL TECHNICAL INSPECTIONS OF WHEELED VEHICLES AND THE RECIPROCAL RECOGNITION OF SUCH INSPECTIONS, https://www.unece.org/fileadmin/DAM/trans/conventn/conf4e.pdf (Nov. 13, 1997) [hereinafter Periodical Technical Inspections].

124 Economic Commission for Europe, AGREEMENT CONCERNING THE ESTABLISHING OF GLOBAL TECHNICAL REGULATIONS FOR WHEELED VEHICLES, EQUIPMENT AND PARTS WHICH CAN BE FITTED AND/OR BE USED ON WHEELED VEHICLES, https://www.unece.org/fileadmin/DAM/trans/conventn/globaut.pdf (accessed 6 August 2019) [hereinafter Global Technical Regulations].

125 Marahe, *supra* note 16.

126 Shri Babul Supriyo, MINISTER OF STATE IN THE MINISTRY OF HEAVY INDUS. AND PUB. ENTER., LOK SABHA UNSTARRED QUESTION NO.4313 (to be answered on 28.03.2017) (Mar. 28, 2017) ("India is a signatory of UNECE, WP-29, 1998 agreement and takes active part in the formulation of Global Technical Regulations. As such all safety norms prescribed under CMVR 1989 are based on UN regulated international standards.").

127 The official description of the UNECE 1958 agreement is, "Agreement concerning the adoption of uniform technical prescriptions for wheeled vehicles, equipment and parts which can be fitted and/or used on wheeled vehicles and the conditions for reciprocal recognition of approvals granted on the basis of these prescriptions." *See Brief Description of "the UN/ECE 1958 Agreement" on Reciprocal*

1998 (US-dominated) rather than UNECE 1958 (EU-dominated) agreement needs to be viewed through this strategic prism. The preparedness of Indian firms for an early compliance to modified global standards has been a major driver. First, UNECE 1958 member countries may adopt a new regulation or modify an existing regulation by a two-thirds majority, while UNECE 1998 does so only by consensus. Moreover, the 1998 agreement allows developing countries to partially modify domestic regulations with GTRs in line with national interests.[128] Second, a major proportion of Indian auto exports goes to Latin America and Africa, who are not part of either UNECE provisions. Therefore, going for a more stringent provision, i.e., UNECE (1958), does not make sense for India. It deserves mention that since 2006, India's sectoral trade with the US (common membership in UNECE 1998), Japan (India-Japan CEPA, 2011) and Thailand (ASEAN-India FTA, 2010) has increased considerably.[129]

Over the last decade, India has established itself as the "largest manufacturer of tractors, second largest manufacturer of two wheelers and buses, fifth largest heavy truck manufacturer, sixth largest car manufacturer and eighth largest commercial vehicle manufacturer in the world."[130] The growth of the sector is being further consolidated through the Automotive Mission Plan (AMP) 2006-26 and National Auto Policy (2018).[131] Some of the supports extended to this sector as part of the 'Make-in-India' initiative include: investment facilitation (100% FDI under automatic route), duty-free import of auto-parts,[132] removal of minimum investment criteria,[133] introduction of a series of fiscal supports,[134] increase in customs duty on commercial vehicles,[135] encouragement for R&D to enhance competitiveness through innovation,[136] etc. A major challenge for Indian players, however, would be to

---

Recognition of Type Approval of Motor Vehicles, etc., Ministry of Land, Infrastructure, Transport and Tourism (Japan), http://www.mlit.go.jp/english/mot_news/mot_news_000627/material1.pdf (accessed Aug. 11, 2019).

128 Ramos, *Supra* note 30, at 13-14.

129 D. Chakraborty et. al., *Global Auto Industry and Product Standards: A Critical Review of India's Economic and Regulatory Experience*, 19(1) J. of Int'l Trade L. and Pol'y, 8 (2019). Pincite is to the first page, should be pg. 23

130 Export Import Bank of India, *The Indian Automotive Industry: An International Trade Perspective* 11 (Mumbai: EXIM Bank, Working Paper No. 59 2017).

131 *See* SOC'Y FOR INDIAN AUTO. MANUFACTURES, *AUTOMOTIVE MISSION PLAN: 2016-26 – A CURTAIN RAISER* 3 (New Delhi: Soc'y for Indian Auto. Manufactures 2016).

132 GOV'T OF INDIA, FOREIGN TRADE POLICY 2015-20, (New Delhi: Dep't of Com., Ministry of Com. and Indus. 2015).

133 INDIA BRAND EQUITY FOUND., AUTOMOBILES SECTOR 24, 34 (New Delhi: India Brand Equity Found. 2017).

134 *Automobile: Financial Support*, MAKE IN INDIA http://www.makeinindia.com/sector/automobiles.

135 GOV'T OF INDIA, MAKE IN INDIA: AUTOMOTIVE SECTOR - ACHIEVEMENTS REPORT, 8-9 (New Delhi: Dep't of Indus. Pol'y and Promotion and Dep't of Heavy Indus. 2016).

136 GOV'T OF INDIA, *supra* note 134, at 10.

adjust to the global emission norms and evolving demand for hybrid and electric vehicles.[137]

Table 3: Importance of Select Sectors in India's Trade Basket (%)

HS Code	Product Groups	Export Share (%)				Import Share (%)			
		2001-04	2005-09	2010-13	2014-18	2001-04	2005-09	2010-13	2014-18
30	Pharmaceuticals	2.52	2.61	3.08	4.44	0.33	0.32	0.35	0.41
52	Cotton	4.02	2.57	3.01	2.58	0.64	0.24	0.14	0.21
53	Other vegetable textile fibres	0.25	0.13	0.13	0.14	0.09	0.05	0.06	0.08
54	Man-made filaments	1.29	0.90	0.86	0.76	0.46	0.23	0.17	0.19
55	Man-made staple fibres	1.07	0.79	0.71	0.72	0.19	0.11	0.12	0.17
56	Wadding, felt and nonwovens	0.09	0.08	0.11	0.14	0.06	0.04	0.04	0.07
57	Carpets etc.	1.18	0.82	0.50	0.60	0.02	0.02	0.02	0.02
58	Special woven fabrics	0.28	0.13	0.10	0.13	0.06	0.04	0.03	0.04
59	Impregnated, coated, covered or laminated textile fabrics	0.10	0.06	0.06	0.09	0.25	0.20	0.16	0.18
60	Knitted or crocheted fabrics	0.07	0.05	0.08	0.10	0.07	0.06	0.07	0.12
61	Articles of apparel and clothing accessories, knitted or crocheted	4.09	2.85	1.99	2.70	0.02	0.02	0.03	0.07
62	Articles of apparel and clothing accessories, not knitted or crocheted	6.14	3.97	2.64	3.09	0.04	0.03	0.04	0.09
63	Other made-up textile articles	2.46	1.71	1.34	1.65	0.11	0.07	0.08	0.11
87	Automobile and Transport Equipment	2.39	3.12	3.98	5.36	0.65	0.87	1.06	1.23

137 GOV'T OF INDIA, NAT'L ELEC. MOBILITY MISSION PLAN 2020, 8-10, 28 (New Delhi: Dep't of Heavy Indus., Ministry of Heavy Indus. and Pub. Enters. 2012).

	Total (Per-cent)	25.92	19.78	18.58	22.48	2.96	2.30	2.34	2.97

	Export Value (USD Billion)				**Import Value (USD Billion)**			
	2001-04	**2005-09**	**2010-13**	**2014-18**	**2001-04**	**2005-09**	**2010-13**	**2014-18**
India's Average Aggregate Trade	57.31	145.22	287.02	292.23	69.88	223.97	441.86	431.69

Source: Constructed by the authors from Trade Map data

Structural Break Analysis

The brief discussion on sectoral dynamics in the last section indicates that in all the three sectors, trade series might have displayed a disjoint, i.e., structural break, at several junctures since 1995. The multilateral factors, likely to influence India's aggregate trade flows, may be considered first. First, for the pharmaceutical sector, the structural break might have occurred after 2005, i.e., after introduction of the product patent regime, owing to the associated competitiveness-related implications. Second, for the cotton, fibre and fabric segments, the structural break is expected to occur after 2001, when the MFA phase-out reforms for the relatively low value-added products deepened. In addition, structural break may also occur after 2005, when final producers of garment and clothing intensify import of these intermediate products for enhancing export of final products. Third, for the garment and clothing products, the structural break can be expected only after the complete MFA phase out, i.e., 2005. Fourth, for the automobile segment, the structural break is expected after 2006, when the domestic players started conforming to the UNECE 1998 standards and subsequently receiving the mutual recognition in the other contracting parties to this convention. On the other hand, the trade preference patterns under RTAs are likely to get reflected in data after formation of the respective trade blocs. By this logic, the structural breaks in trade with ASEAN and South Korea may occur after 2010 and, for Japan, after 2011. As the India-Singapore and India-Malaysia CECA entered into force in 2005 and 2011 respectively, their effects may be contained within the India-ASEAN trade flows. Finally, structural breaks observed in an odd year can be explained by trade reforms/growing protectionism or the global/regional demand dynamics.

Among the three sectors considered in this analysis, two of them (pharmaceutical and automobile) show increasing trends in terms of export shares in India's trade basket and the other (textiles and garments) shows a decreasing trend. Looking at the changes in these sectors and identifying time points where these have been significant can reveal the underlying triggers, which can be economic or other shocks and policy changes. A series of policy

changes, both domestic and multilateral, as already mentioned in the earlier section, can be the candidate triggers. Identifying the points where the change has been significant for the time period considered can take two paths. First, focusing on a known structural break date, e.g., phasing away of the MFA regime from the year 2005, which is expected to have a significant impact on the sector and definitely on exports. The most common method for empirically estimating this break is Chow's test.[138] In this method, one tests whether a single regression equation is more efficient than representing the data through two separate ones, one till the identified break date and the other from the next period to the end date. Here the data needs to be split, and two separate regressions need to be reported. The F-test is performed to find out whether the break is indeed statistically significant. This approach has a few problems. One, since the data is split into two periods, it creates problems in degrees of freedom. Data requirement increases significantly which is a problem when one works with yearly time series. Second, the theoretical basis for the known break should be strong. For example, in the auto sector India adopted UNECE (1998) in the year 2006.[139] This significantly changed the auto sector policy through change in the CMVR. Does this change lead to significant change in production and hence exports? What is the transmission mechanism through which this policy can lead to significant change in export? These are not easy questions to answer. Finally, there can be multiple candidates (i.e., policy change moments) for identifying breaks throughout the selected period. Deciding which one among them should be considered for a single structural break analysis is again a difficult choice.

The splitting of data series problem can be taken care of by the dummy variable approach. Even if we are fairly confident of our choice of break date and it is chosen by observing the data series, there can be a problem. It in effect means we are estimating the break date itself and it nullifies the Chow Test critical values. An alternate approach, which we have adopted in this paper, is to find out the break date endogenously. The approach was pioneered by Quandt (1958, 1960).[140] As has been observed in case of the three sectors selected for analysis, a series of policy changes happened both through domestic (unilateral and from adopting of internationals standards) and multilateral policy changes. Hence, Chow's test for a single structural break does not seem to be the right approach. Instead, the unknown structural break test constructs a test statistic without specifying a known date. Here,

138 *See* G. C. Chow, *Tests of Equality between Sets of Coefficients in Two Linear Regressions*, 28 ECONOMETRICA 591, 591-605 (1960).

139 Marathe, supra note 16.

140 *See* R. E. Quandt, *The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes*, 53 J. OF AM. STAT. ASS'N 873, 873-80 (1958); R. E. Quandt, *Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes,* 55 J. OF AM. STAT. ASS'N 324, 324-30 (1960).

the test statistic is estimated for each possible break date in the sample and is combined. A Wald test based on the maximum of the sample Wald tests is used to determine the most significant break point. The maximum sample test is compared with what could have been expected under the null of no structural break. The advantage of this approach is it identifies the structural break at the point where it is most significant and hence allows us to link it to the policy or shock corresponding to that time period. A Quandt Likelihood Ratio (QLR) statistic[141] is calculated based on the maximum of all the Chow F-statistic in a truncated time period. To identify the break through method, usually 70% of the years considered are retained. A 15% cut is applied to the initial and final set of years. The results in Table 4 should be read and understood from this perspective. The identified break dates for each of the sectors and subsectors are the periods where the break is most significant in Chow's sense. The policy change corresponding to the dates mentioned in the table are thus responsible for most significant change in slope of the corresponding series. It may be noted that the test procedure does not rule out the presence of other breaks but only finds out the most significant one.

Table 4 summarizes the endogenous structural break analysis for India's aggregate sectoral trade with rest of the world. It is observed that for pharma (2005),[142] fabric (2002)[143] and garments (2005)[144] exports, the breaks are observed around the policy change years. In other words, the multilateral policy changes have influenced the deviation in the series (export rise in this case) from the old trend. In case of cotton (2010) and fibres (2010), the break year can be explained by the growing demand of low-value added products after revival of world growth scenario in the aftermath of the sub-prime crisis.[145] The auto sector (2004) however interestingly shows no impact of the UNECE 1998 membership, perhaps owing to the continuing dominance of several non-member countries (e.g., Africa, Latin America) in India's export basket.[146] But the general rise in auto export, after joining WP 29 as an observer in 2003, might have motivated India to go ahead for the UNECE 1998 forum membership in 2006. On the import front, the structural breaks for four segments, namely, pharma (2006), fibre (2003), fabric (2004) and garments (2007), occur after the policy changes and are along the expected lines. The break year for cotton (2009) can be explained by a sharp decline in imports,

141 Let $F(\tau)$ be the Chow's test statistic for estimating no structural break at time τ. Then the QLR Test statistic is as follows: $QLR = Max\{F(\tau_0), F(\tau_0 + 1), \ldots F(\tau_1 - 1), F(\tau_1)\}$ for $\tau_0 \leq \tau \leq \tau_1$. The critical values for QLR Statistic for the trimming used (15% as mentioned in the text) is available in all standard econometric softwares. The break is identified by comparing the calculated QLR with the corresponding critical value.

142 *See* Dhar and Gopakumar, *supra* note 86, at 1-10.

143 Verma, *supra* note 85, at 18-23.

144 *See* Tewari, *supra* note 14.

145 *See infra* Table 4.

146 *See* Chakraborty et al., *supra* note 112, at [pincite] 17. *See also infra* Table 4.

as Indian global garment exports declined due to the recession.[147] The break year for automobile (2007) can be explained by a sharp rise in imports from China and South Korea, both of whom being members of UNECE 1998 arrangement received eased import policy framework after Indian entry in the forum.[148]

Table 5 presents the results for India-ASEAN trade. In the export series, the break years for cotton (2003) and garments (2004) are in line with the MFA phase-out.[149] The scenario for the pharma sector (2004) can be explained by the diversification of the sectoral export basket to the world in general and to ASEAN in particular.[150] The break year for fibre (2016) [151] can be explained by the decline in demand due to recession and shrinking import in several ASEAN countries (e.g., Cambodia, Laos, Myanmar, and Vietnam) respectively. From Trade Map data, a reduction in demand from ASEAN is observed for the auto sector (2016) as well. The fabric (2014) export growth is explained by the rising exports to Indonesia, Myanmar, Malaysia, and Philippines.[152] In case of imports, while the break years for cotton (2005) and fibre (2003) are in line with MFA phase-out, for the other four sectors, the global, as well as regional, recession during 2016-17 is the underlying factor (2016).[153]

The analysis on trade with China is summarized in Table 6. In exports, while fabric (2005) and garments (2007) have largely displayed the MFA phase-out effect, the break for fibre (2008) can be explained by the rise in exports of HS 53 (other vegetable textile fibres), as Chinese garment exports enjoyed a growth spree till the global financial crisis.[154] In case of pharma (2015) and automobiles (2016), Indian exports to China witnessed a sharp

147 *See* D. Gopalakrishnan, S. Anandhakumar, K. Santhoshkr & U. Divya, *Global Economic Challenge & its Impact on Indian Textiles*, (Feb. 2010), https://indiantextilejournal.com/articles/FAdetails.asp?id=2680.

148 *See* Lijee Philip & Chanchal Pal Chauhan, *Cheap Chinese Imports Threaten to Derail India's Auto Parts Industry*, ECONOMIC TIMES (Dec.16, 2008), https://economictimes.indiatimes.com/cheap-chinese-imports-threaten-to-derail-indias-auto-parts-industry/articleshow/3843732.cms. *See also infra* Table 4.

149 *See infra* Table 5.

150 may be noted that to facilitate pharma sector exports, Pharmaceutical Export Promotion Council (PHARMEXCIL) was established in 2004 by the Ministry of Commerce and Industry, Government of India. *See also infra* Table 5.

151 *See* Henry Dsouza, *India's T&C Exports Have Dipped Again in FY 16-17*, TEXTILE EXCELLENCE, (Jan. 8, 2017), https://www.textileexcellence.com/news/trade-policy/indias-tc-exports-have-dipped-again-in-fy-16-17/.

152 *See infra* Table 4.

153 *See World Economic Outlook Update: A Shifting Global Economic Landscape,* INT'L MONETARY FUND 1, 4, 7 (2017), https://www.imf.org/external/pubs/ft/weo/2017/update/01/pdf/0117.pdf. *See also infra* Table 5.

154 *See* Amitendu Palit & Shounkie Nawani, *India-China Trade: Explaining the Imbalance* 1–16, 5 (Inst. of S. Asian Stud., Working Paper No. 95, 2009). *See also infra,* Table 6.

fall.[155] While the former can be explained by increased competition in Chinese market,[156] the latter can be explained by fall in exports of auto-components to China, namely parts and accessories for tractors, motor vehicles, etc. (8708), underlining the competitiveness-related challenges for India in joining Asian IPNs.[157] Cotton import from India showed a spike (2013), given the export-related requirement in China.[158] As China moved away from the cotton stockpiling programme from 2014 onward,[159] the import from India declined subsequently.[160] On the import front, while the break for pharma (2004) can be explained by rise in inflow of certain product groups, namely bandages, etc. (HS 3005) in line with expectations, the global recession effect dominates the fibre, fabric, auto (2016) and garments (2009) sectors. The break in growing demand for cotton from China (2013) can be explained by the rise in exports of garments during 2013-14.[161]

Table 7 presents the results for India-Japan trade. Garment (2004) and auto (2009) exports received a boost as a result of the MFA phase-out and UNECE 1998 agreement accession (where Japan is a member country) respectively. In the case of pharma, the sharp fall (2014) is caused by India's loss in the Japanese import market of medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses (HS 3004) to USA, Belgium and China, owing to competitiveness concerns.[162] The recession effect dominates decline in exports of cotton (2015), fibre (2009)[163], and fabric (2016)[164]. On the import front, the MFA phase-out dynamics are noted for cotton (2005) and fibre (2004), while the decline in garment imports from Japan (2015) can be explained by the rise in Indian import from China and South Korea for other made-up textile articles (HS 63). The rising import

155 *See infra,* Table 6.

156 *See* Swati Rana, *Pharma Export Records 1.96% Negative Growth in Bulk Drug and Formulations*, *Ayush Export Rises 16%*, PHARMABIZ (Dec. 2, 2016) http://www.pharmabiz.com/NewsDetails.aspx?aid=99012&sid=1.

157 *See* Biswajit Nag, Biswa Nath Bhattacharyay & Debdeep De, *Integrating India with Asian Production Networks: Prospects and Challenges* 1–31, 16–20 (Ctr. for Econ. Stud. and IFO Inst.,Working Paper No. 5616, 2015).

158 See *infra* Table 6.

159 *See* STEPHEN MACDONALD, FRED GALE & JAMES HANSEN, COTTON POLICY IN CHINA 15 (2015).

160 *See* Meenakshi Sharma, *India's Cotton Exports Hit as China Shifts Policy*, REUTERS (Apr. 17, 2014), https://www.reuters.com/article/india-cotton-exports/indias-cotton-exports-hit-as-china-shifts-policy-idUSL4N0MP2MK20140417.

161 *See* Press Trust of India, *Textile Industry on Growth Path, 2013-14 Encouraging: SIMA*, ECONOMIC TIMES (July 2, 2014), https://economictimes.indiatimes.com/industry/cons-products/garments-/-textiles/textile-industry-on-growth-path-2013-14-encouraging-sima/articleshow/37635468.cms.

162 Rana, *supra* note 155.

163 Press Trust of India, *High interest, raw material cost hit textile industry*, THE ECONOMIC TIMES (July 2, 2009, 2:23PM IST), https://economictimes.indiatimes.com/news/economy/policy/high-interest-raw-material-cost-hit-textile-industry/articleshow/4728497.cms.

164 Dsouza, *supra* note 150.

from China continued in the subsequent period as well.[165] The recession effect explains the scenario in the pharma (2013) and fabric (2009) sectors. The break for auto products (2016) can be explained by the rise in imports of motor cars and other motor vehicles (8703) and parts and accessories for tractors, motor vehicles, etc. (8708) from Japan, signifying both growing intra-industry trade (IIT) and deepening of IPNs.[166]

The analysis with South Korea is summarized in Table 8. For exports, the break years for the fibre (2004), fabric, (2004) and garments (2005) sectors can be explained by the MFA phase-out, while the same for auto (2009) depicts the UNECE 1998 (where South Korea is a member country) participation effect. The underlying reason for the break in pharma (2012) is the rise in the export of several products, namely medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses (HS 3004) and medicaments consisting of two or more constituents (HS 3003). On the other hand, the break for cotton (2010) has been primarily caused by the rise in the export of cotton yarn other than sewing thread, containing >= 85% cotton by weight (HS 5205). Apart from a surge in demand in South Korea and rise in Indian export competitiveness[167], the CEPA with South Korea (2010) might have helped India to gain market access in pharma and cotton segments. On the import front, the sectoral breaks can largely be explained by multilateral policy changes, namely for: pharma (2006), cotton (2004), fabric (2005) and garments (2005). The breaks for fibre (2008) can be explained by a rise in domestic demand in India. The auto sector (2016) presents interesting dynamics. While India's overall import of this category (HS 87) declined due to the recession effect, the same from South Korea increased by displacing Germany, USA (non-RTA partner countries) and Thailand (part of Indo-ASEAN FTA) in India's import basket. The observation underlines a growing participation of South Korea in Indian auto IPNs.[168]

A few interesting observations emerging from the structural break analysis involving India's selected trade partners deserve mention here. First, India's aggregate as well as bilateral trade flows have often been influenced by multilateral reforms (i.e., MFA phase-out, product patent and UNECE 1998). Second, the general global trade dynamics, i.e., both boom (2009-10) and recession (2009, 2016) periods, have been reflected in the identified break years. Third, the RTA effect has not been strongly revealed in trade with trade partners for the selected sectors. Only in the subsequent period,

165 Rajendra Jadhav and Sudarshan Varadhan, *India doubles import tax on over 300 textile products to 20%, may hit China*, Live Mint (Aug. 07, 2018), https://www.livemint.com/Politics/pL5q4ZP7LwrOdYYfZEgzKL/India-doubles-import-tax-on-some-textile-products-to-20.html.

166 Nag, *supra* note 71.

167 V. S. SESHADRI, INDIA-KOREA CEPA: AN APPRAISAL OF PROGRESS, 10, 39 (ASEAN-India Ctr. & Rsch. and Info. Sys. for Dev. Countries eds., 2015).

168 D. Chakraborty, *Picking the Right Alternative: Should India participate in TPP instead of RCEP?*, *in* THE TRANSPACIFIC PARTNERSHIP: A PARADIGM SHIFT IN INTERNATIONAL TRADE REGULATION, 501-519, 514-15 (Chaisse et al. eds., 2018).

can export growth in certain sectors be linked with the preference obtained in partner markets, e.g., fabric (ASEAN), auto (Japan and South Korea). The absence of break points after entry into force of the RTAs indicates continuation of bilateral trade along the historical trend
line.

The result underlines the limited gains enjoyed by India through the engagement with RTA partners in Southeast and East Asia.[169] On the other hand, in China, the structural breaks have often been noticed as a sharp reversal of trend in data series, e.g., involving pharma, auto, cotton. The observations are in line with the reported competitive edge of China vis-à-vis India.[170]

Table 4: Structural Breaks in India's Trade with Rest of World

Product Groups (HS Code)	Export break Year	Underlying Driver	Import break Year	Underlying Driver
Pharma (30)	2005	Product patent effect	2006	Product patent effect, with subsequent tariff reforms and rise in component imports
Cotton (52)	2010	Raw material specialization, revival of world growth after Sub-prime crisis	2009	Global recession effect
53-55 (Fibre)	2010	Raw material specialization, revival of world growth after Sub-prime crisis	2003	MFA phase-out effect, growth in Clothing export, need for raw material
56-60 (Fabric, Intermediate and Low Value Added)	2002	MFA phase-out effect, specialization	2004	MFA phase-out effect, growth in Clothing export, need for raw material
61-63 (Garments and Other Made up Textile)	2005	MFA phase-out effect, specialization	2007	MFA phase-out effect, tariff reforms
Auto (87)	2004	Rise in export, given participation in UNECE 1998	2007	Post UNECE participation, foreign automakers and OEMs start importing components, particularly from UNECE 1998 contracting parties

Source: Author's Estimation and Policy Analysis

169 Manoj Pant & Anusree Paul, *The Role of Regional Trade Agreements: In the Case of India*, 33(3) J. ECON. INTEGRATION 538-571, 561 (2018).

170 KLAUS SCHWAB, THE GLOBAL COMPETITIVENESS REPORT 2019 15-16 (World Econ. F. ed., 2019).

Table 5: Structural Breaks in India's Trade with ASEAN

Product Groups (HS Code)	Export break Year	Underlying Driver	Import break Year	Underlying Driver
Pharma (30)	2004	Product patent effect, export basket diversification	2016	Global recession effect
Cotton (52)	2003	MFA phase-out effect, specialization	2005	MFA phase-out effect
53-55 (Fibre)	2016	Global recession effect	2003	MFA phase-out effect
56-60 (Fabric, Intermediate and Low Value Added)	2014	RTA effect, export growth to Indonesia, Myanmar, Malaysia and Philippines	2016	Global recession effect
61-63 (Garments and Other Made up Textile)	2004	MFA phase-out effect, specialization	2016	Global recession effect
Auto (87)	2016	Global recession effect	2016	Global recession effect

Source: Author's Estimation and Policy Analysis

Table 6: Structural Breaks in India's Trade with China

Product Groups (HS Code)	Export break Year	Underlying Driver	Import break Year	Underlying Driver
Pharma (30)	2015	Competitiveness-related challenges in Chinese markets	2004	Demand growth for medical supplies
Cotton (52)	2013	Demand growth, given stage of cotton stockpiling programme in China	2013	Rise in exports of garments, demand for raw materials
53-55 (Fibre)	2008	Growth in China and rising demand for intermediate products	2016	Global recession effect
56-60 (Fabric, Intermediate and Low Value Added)	2005	MFA phase-out effect, specialization	2016	Global recession effect
61-63 (Garments and Other Made up Textile)	2007	MFA phase-out effect, specialization	2009	Global recession effect
Auto (87)	2016	Competitiveness-related challenges in Chinese markets	2016	Global recession effect

Source: Author's Estimation and Policy Analysis

Table 7: Structural Breaks in India's Trade with Japan

Product Groups (HS Code)	Export break Year	Underlying Driver	Import break Year	Underlying Driver
Pharma (30)	2014	Loss of competitive edge to USA, Belgium and China	2013	Global recession effect
Cotton (52)	2015	Global recession effect	2005	MFA phase-out effect, growth in Clothing export, need for raw material
53-55 (Fibre)	2009	Global recession effect	2004	MFA phase-out effect, growth in Clothing export, need for raw material
56-60 (Fabric, Intermediate and Low Value Added)	2016	Global recession effect	2009	Global recession effect
61-63 (Garments and Other Made up Textile)	2004	MFA phase-out effect, specialization	2015	Rise in import from China and South Korea
Auto (87)	2009	Rise in export, given participation in UNECE 1998	2016	Growing intra-industry trade (IIT) and deepening of IPNs

Source: Author's Estimation and Policy Analysis

Table 8: Structural Breaks in India's Trade with South Korea

Product Groups (HS Code)	Export break Year	Underlying Driver	Import break Year	Underlying Driver
Pharma (30)	2012	Specialization, CEPA formation	2006	Product patent effect
Cotton (52)	2010	Specialization, CEPA formation	2004	MFA phase-out effect, growth in Clothing export, need for raw material
53-55 (Fibre)	2004	MFA phase-out effect, specialization	2008	Growth in Clothing export, need for raw material
56-60 (Fabric, Intermediate and Low Value Added)	2004	MFA phase-out effect, specialization	2005	MFA phase-out effect, growth in Clothing export, need for raw material
61-63 (Garments and Other Made up Textile)	2005	MFA phase-out effect, specialization	2005	MFA phase-out effect, demand effect, tariff reforms
Auto (87)	2009	Rise in export, given participation in UNECE 1998	2016	Growing intra-industry trade (IIT) and deepening of IPNs

Source: Author's Estimation and Policy Analysis

SECTORAL TARIFF REFORMS AND TRADE BALANCE SCENARIO

To gauge India's liberalization and competitiveness in the selected sectors an analysis on the country's tariff and trade scenario is provided next. Table 9 shows the sectoral (at HS 2-digit level) tariff scenario in India during the four periods, as obtained from the WITS data.[171] The table first reports the simple and weighted average effective applied tariffs, which enables one to understand the actual reform. If the weighted tariff is found to be lower than the simple tariff, it implies higher import happening under lower tariff lines, i.e., signifying effective trade reforms.[172] The occurrence of domestic and international peak tariffs,[173] which indicate sectoral protection inclination, is reported next. Imposition of tariff rates above 15 percent on HS 6-digit tariff lines within a HS 2-digit product group are termed as international tariff peaks, depicting possible protectionist intents. On the other hand, the domestic tariff peaks exist if a tariff line at HS 6-digit level faces a tariff greater than three times the aggregate average tariff. In the current context, the mere presence of domestic and international peak tariffs in a sector for any year has been considered as a protectionist inclination. The domestic and international peak tariffs are summarized in last two columns of Table 9. If the entry in a row reads 'No', then the corresponding sector has not witnessed high tariff spikes, either by international or domestic benchmarks, over the 2001-18 period. Conversely, presence of peak tariff on either count, if any, has been clearly specified by indicating the relevant years under the appropriate columns.

A few observations emerge from the analysis. First, a declining linear trend is noted both in simple and weighted average effective applied tariffs across sectors, signifying gradual reforms. Second, however, a relatively higher weighted average tariff indicates possible hindrances in certain sectors, namely pharma (entire period), special woven fabric (2010-13 and 2014-18), textile fabrics (2005-09, 2010-13 and 2014-18), knitted or crocheted fabrics (2005-09, 2010-13 and 2014-18) and apparels (2010-13). Third, domestic peak tariff for a prolonged period is observed only in the case of man-made staple fibres (2005-16) and auto-products (2001-18). On the other hand, international peak tariff for a long duration has been observed in case of cotton, other vegetable textile fibres and auto-products for the entire period (2001-18). In a wide range of sectors, the country has not imposed an international peak tariff after 2006. In other words, the observed tariff protection effect across sectors has generally come down.

---

171 *See generally* World Integrated Trade Solution, *Trade Statistics by Product (HS 6-digit)*, https://wits.worldbank.org/ (WITS data compiled by authors to create Table 9).

172 Douglas A. Irwin, *Trade Restrictiveness and Deadweight Losses from US Tariffs,* 2(3) Am. Econ. J.: Econ. Pol'y 111, 120, 129 (2010).

173 André M. Nassar, Zuleika Arashiro & Marcos S. Jank, *Tariff Spikes and Tariff Escalation*, *in* HANDBOOK ON INTERNATIONAL TRADE POLICY 222-36 (William A. Kerr & James D. Gaisford eds., 2007).

Table 10 reports sectoral import growth and trade balance scenarios for the four periods using Trade Map data. It is observed from the table that India's import growth has been significantly higher during 2001-02 to 2004-05 and 2009-10 to 2012-13 periods, underlining the policy reforms during the period. Over 2001-02 to 2004-05, India relied more on multilateral trade reforms and gradually reduced tariff rates. In addition, since 2008 India allowed Duty Free Quota Free (DFQF) tariff preference for LDCs, paving the way for imports from low-cost partners.[174] The relatively lower import growth rate over 2013-14 to 2017-18 results from the higher import base as well as the global recession effect.[175] Interestingly, only impregnated, coated, covered or laminated textile fabrics (HS 59) and knitted or crocheted fabrics (HS 60), have shown a declining trade balance over the period. On the other hand, several sectors, e.g., pharma (HS 30), carpets (HS 57), special woven fabric (HS 59), garments, knitted or crocheted (HS 61), garments, not knitted or crocheted (HS 62), other made-up textile articles (HS 63) and auto-products (HS 87), have shown positive and improving trade balance. The results indicate a rise in competitiveness in these categories.[176]

Tables 11 and 12 show the trade balance scenario for India in the selected sectors vis-à-vis the RTA partners, namely: Singapore, Malaysia, ASEAN, Japan and South Korea over the four periods. For comparability, the trade balance scenario for China, the RCEP negotiation partner, is also reported. The emerging observations are noted in the following. First, in the pharma sector, India has witnessed a negative trade balance with only China and South Korea, which worsened from 2014 onwards. While the growing trade balance in ASAN has been an indication of the rising competitiveness on this front, the threat of Chinese competition is apparent. Second, in case of cotton, India is enjoying a trade surplus with the partners, which underlines specialization in a low value-added category. Third, the scenario for fibres and fabrics shows a dismal outcome; for each partner, at some point India has had a negative trade balance. In the ASEAN market, over 2010-13 in the fabric segment, India improved the trade balance which can be attributed to the preferential access following Indo-ASEAN FTA reforms. However, over 2014-18, trade balance has been negative and worsening for Malaysia,

174 Murali Kallummal, et al., *Utilising India's Duty-Free Preference Scheme for LDCs: Analysis of the Trade Trends*, 8, 11 (Centre for WTO Studies, Indian Institute of Foreign Trade, Working Paper No. CWS/WP/200/10, 2013).

175 *See generally*, *Exports decline 6 per cent to $26.13 billion; trade deficit narrows*, Times of India 1,2 (Sept 14, 2019 8:12 IST), https://timesofindia.indiatimes.com/business/india-business/exports-decline-6-to-26-13-billion-trade-deficit-narrows/articleshow/71115960; Jonathan Eaton, J. et al., *Trade and the Global Recession*, 106(11) Am. Econ. R. 3428-29 (2016) (showing the export and import growth decline for India continued in 2019 as well).

176 Misu Kim, *Export Competitiveness of India's Textiles and Clothing Sector in the United States*, Economies, 3,6,11-12 (2019); Manisha Dhiman and Sandeep Kaur Bhatia, *Changing Pattern of India's High Technology Exports: A Study of Competitiveness of Pharmaceutical Products*, Business Analyst, 37(1): 118, 126-37 (2018); Export-Import Bank of the United States, 2017 Annual Report: Exports Create U.S. Jobs, 23 (2017).

ASEAN, China and South Korea. Though the trade balance for fabrics remained positive, it declined against Singapore and Japan as well. In all, the result indicates sliding competitiveness for India in the mid-segment of the garment value-chain. Fourth, in case of garment exports India enjoys trade surplus with all the partners, barring China, which receive no trade preference in Indian market. The result indicates that while India remains competitive in the garment segment against the current RTA partner countries, a tussle with China for market access within RCEP could have posed a major challenge. Finally, in case of auto products India has witnessed an interesting transition. While India's average trade balance against Singapore, Malaysia and ASAN have remained positive, the margins have declined in the post-2014 period. On the other hand, the trade balance has been negative and declining against China and Japan over the

entire period. In the case of South Korea, though the trade balance remains negative, it somewhat improved after 2014. The results indicate that after accession to UNECE 1998, the imports from China, Japan and South Korea (all being contacting parties to the agreement) have increased faster compared to the corresponding export figures. The mixed trade balance outcome against the RCEP negotiating partners (particularly, China) fuels

the perceived need for protection and justifies India's decision to tread cautiously at the negotiations. [177]

[177] Debashis Chakraborty et al., *Is It Finally Time for India's Free Trade Agreements? The ASEAN "Present" and the RCEP "Future"*, 9(2) Asian J. of Int'l L. 359, 377, 381, 391 (2019).

Table 9: Tariff Scenario on Indian Imports for Select Sectors

HS Code	Product Groups	Effective Applied Tariff (Simple Average)				Effective Applied Tariff (Weighted Average)		Domestic Peaks			International Peaks
		2001-04	2005-09	2010-13	2014-18	2001-04	2005-09	2010-13	2014-18		
30	Pharmaceuticals	28.84	11.67	9.48	8.83	29.55	11.89	9.85	9.34	No	2001-04
52	Cotton	25.86	12.75	8.82	7.66	12.90	11.20	3.02	1.50	2006	2001-18
53	Other vegetable textile fibres	29.60	12.46	9.01	7.41	27.30	11.79	8.69	4.08	No	2001-18
54	Man-made filaments	22.17	14.99	8.60	8.08	20.72	14.10	8.60	7.87	2006	2001-04, 2006
55	Man-made staple fibres	24.66	13.79	8.49	8.16	22.03	12.03	8.44	8.05	2005-16	2001-06
56	Wadding, felt and nonwovens	24.03	11.87	9.58	8.18	23.30	11.42	9.43	7.45	No	2001-04
57	Carpets etc.	30.00	13.05	9.52	8.31	30.00	14.26	9.62	6.27	No	2001-04, 2006
58	Special woven fabrics	28.26	15.43	9.53	7.93	26.76	15.16	9.62	8.59	2006	2001-04, 2006
59	Impregnated, coated, covered or laminated textile fabrics	27.56	11.95	9.57	8.69	26.18	11.99	9.72	9.15	No	2001-04
60	Knitted or crocheted fabrics	29.46	11.48	9.50	8.06	28.45	11.57	9.54	8.71	2006	2001-04, 2006
61	Articles of apparel and clothing accessories, knitted or crocheted	29.69	13.53	9.69	7.09	29.41	13.34	9.81	6.89	2006	2001-04, 2006
62	Articles of apparel and clothing accessories, not knitted or crocheted	29.69	16.10	9.63	6.58	29.08	15.61	9.66	3.40	2006	2001-04, 2006
63	Other made-up textile articles	29.19	11.72	9.23	8.20	26.75	11.02	8.91	7.57	No	2001-04, 2006
87	Automobile and Transport Equipment	43.10	24.60	21.32	21.73	38.80	20.86	15.65	13.86	2001-18	2001-18

Source: Constructed by the authors from WITS data

Table 10: Import Growth and Trade Balance Scenario in India for Select Sectors

HS Code	Product Groups	Import Growth Rate (%)				Trade Balance (USD Million)			
		2001-02 to 2004-05	2005-06 to 2008-09	2009-10 to 2012-13	2013-14 to 2017-18	2001-04	2005-09	2010-13	2014-18
30	Pharmaceuticals	24.00	30.50	12.25	4.60	1219.48	3115.11	7391.15	11120.96
52	Cotton	3.50	3.00	19.00	7.60	1813.94	3148.19	8023.34	6643.78
53	Other vegetable textile fibres	30.00	16.75	22.50	4.40	77.60	65.54	102.55	68.14
54	Man-made filaments	19.00	3.25	11.00	4.00	428.32	823.90	1711.90	1394.03
55	Man-made staple fibres	19.50	12.50	23.50	9.60	482.84	902.11	1509.40	1369.62
56	Wadding, felt and nonwovens	24.50	4.50	20.00	14.20	7.76	29.62	138.75	110.58
57	Carpets etc.	35.25	11.00	14.25	11.00	648.87	1085.46	1344.97	1653.62
58	Special woven fabrics	37.25	6.00	11.00	5.20	100.36	93.14	150.85	183.72
59	Impregnated, coated, covered or laminated textile fabrics	30.25	7.50	12.50	3.60	-114.51	-352.56	-530.09	-515.36
60	Knitted or crocheted fabrics	26.25	15.00	25.50	8.80	-5.86	-42.64	-112.56	-208.94
61	Articles of apparel and clothing accessories, knitted or crocheted	11.75	33.75	39.50	23.60	2301.85	4046.59	5582.69	7521.19
62	Articles of apparel and clothing accessories, not knitted or crocheted	18.25	22.75	40.75	20.40	3384.00	5492.60	7356.50	8527.35
63	Other made-up textile articles	15.75	27.75	13.00	6.00	1330.75	2211.10	3513.92	4318.98
87	Automobile and Transport Equipment	41.50	33.00	15.75	6.40	952.90	2511.32	6750.87	10355.52

Source: Constructed by the authors from Trade Map data

Table 11: Trade Balance Scenario for India against Select Partners (USD Million)

Product Groups (HS Code)	Trade Balance against Singapore				Trade Balance against Malaysia				Trade Balance against ASEAN			
	A	B	C	D	A	B	C	D	A	B	C	D
Pharma (30)	13.64	13.35	41.21	37.90	7.56	18.13	32.70	45.93	98.24	214.05	413.29	648.09
Cotton (52)	15.39	11.79	14.71	5.29	31.23	27.81	82.28	59.27	98.46	227.43	550.54	743.28
53-55 (Fibre)	42.67	16.31	24.42	-8.97	-1.07	2.75	41.70	15.41	11.81	13.12	105.74	-41.94
56-60 (Fabric, Intermediate and Low Value Added)	4.18	12.31	22.83	11.49	3.74	6.25	17.63	-4.23	-16.82	-38.88	6.76	-81.92
61-63 (Garments and Other Made up Textile)	42.87	55.71	77.10	52.60	40.00	50.70	84.32	117.44	105.40	118.60	204.18	200.98
Auto (87)	0.36	43.22	198.16	8.58	13.77	32.80	51.12	24.37	27.45	140.32	643.10	491.53

Source: Constructed by the authors from WITS data
Period A, B, C and D represent 2001-04, 2005-09, 2010-13 and 2014-18 respectively

Table 12: Trade Balance Scenario for India against Select Partners (USD Million)

Product Groups	Trade Balance against China				Trade Balance against Japan				Trade Balance against South Korea			
	A	B	C	D	A	B	C	D	A	B	C	D
Pharma (30)	18.55	-4.63	-61.64	-106.92	4.17	-1.43	23.23	21.31	3.16	0.51	-19.87	-27.65
Cotton (52)	46.50	520.37	3146.81	1596.79	72.59	60.37	67.84	59.29	163.71	184.18	247.08	209.61
53-55 (Fibre)	-60.23	-248.26	-490.53	-638.69	-14.19	-32.43	-64.91	-73.02	-56.51	-41.72	-39.50	2.97
56-60 (Fabric, Intermediate and Low Value Added)	-71.05	-368.31	-769.00	-1077.48	22.54	21.25	21.34	3.98	-25.89	-23.56	-45.62	-38.59
61-63 (Garments and Other Made up Textile)	-10.37	-51.58	-134.63	-339.22	110.00	149.19	237.78	240.55	-2.49	2.48	19.72	18.05
Auto (87)	3.35	-258.52	-830.09	-1238.69	-80.55	-348.81	-468.27	-401.69	-106.59	-436.12	-743.04	-665.93

Source: Constructed by the authors from WITS data

Period A, B, C and D represent 2001-04, 2005-09, 2010-13 and 2014-18 respectively

CONCLUSION

India made a conscious move to promote RTA-led exports in the post-2005 period, given the slow progress of the multilateral negotiations.[178] Several *ex ante* simulation analyses indicated significant export growth prospects for India resulting from these arrangements, namely: India-ASEAN FTA[179], India-Japan RTA[180], India-South Korea RTA[181] and India-China FTA[182]. The benefits for India was expected not only through growing merchandise and service exports leading to integration with Asian IPNs, but also through increasing FDI inflows, associated with much-aspired technology transfer.[183] The benefits were also expected through growing regional trade integration and deeper participation in Asian GVCs, which indeed has shown an upward trend.[184] In several sectors, despite receiving a tariff preference, India has however witnessed a worsening trade balance scenario against current FTA partners (e.g., Japan, South Korea) as well as potential RCEP partners (i.e., China). As the current paper ascertains, the expectations on RTA-led trade expansion remain under-fulfilled in three key sectors in India's trade basket in general and for pharmaceuticals, textile and garments and automotive sectors, in particular. India's November 2019 decision therefore needs to be viewed in this wider context.

The import threat drivers played a crucial role in India's RCEP pull-out. First, as India has already been integrated with all the RCEP countries barring the exception of China, entry into the trade bloc would have resulted in preferential trade only in that front. The worsening trade balance trend with China indicates that India could have faced a further deterioration by joining RCEP, after granting tariff preference to imports from the dragon. In addition, in the past wherever China failed to export a product directly to India due to anti-dumping and other policies, they invested in Vietnam and other low-cost ASEAN economies to route the same indirectly, taking advantage

178 Chaisse et al., *supra* note 23, at 416.

179 Chandrima Sikdar & Biswajit Nag, *Impact of India-ASEAN Free Trade Agreement: A cross-country analysis using applied general equilibrium modelling 31-32*, (U.N. Econ. and Soc. Comm'n for Asia and the Pac., Working Paper No. 107, 2011).

180 Geethanjali Nataraj, *India-Japan Investment Relations: Trends & Prospects,* 2-3 (Indian Council for Rsch. in Int'l Econ. Rel.: New Delhi, Working Paper No. 245, 2010).

181 Shahid Ahmed, *India-Korea CEPA: An Assessment*, 12 Kor. and the World Econ. 45, 52 (2011).

182 Swapan K. Bhattacharya & Biswa N. Bhattacharyay, *Free Trade Agreement between People's Republic of China and India: Likely Impact and Its Implications to Asian Economic Community*, 2006 Asian Dev. Bank Inst., 4.

183 Mridula Manjari Moitra Roy & Rupa Chanda, *The trends in FDI inflows from Japan to India*, (India Japan Study Centre, Indian Inst. of Mgmt., Bengaluru, Working Paper No. 1, 2019).

184 U.N. Econ. and Soc. Comm'n for Asia and the Pac., ESCAP Digital and Sustainable Regional Integration Index at 25, 76, 91, https://www.unescap.org/resources/DigiSRII (last visited Apr. 21, 2020).

of the preferential access.[185] The lobbying by key domestic industries, threatened by growing Chinese imports, played a major role in arriving at the final decision.[186] Second, though the negotiations with Australia and New Zealand for trilateral FTA are ongoing, the RCEP would have granted their exports, particularly in the dairy sector, tariff-free access for the first time. The dairy industry noted that the price of skimmed milk powder (SMP), cheese, butter and other derivatives in Australia and New Zealand would be significantly lower than the corresponding Indian varieties.[187] In 2019, sensing the pace of RCEP negotiations, opposition grew in India over duty-free import of these products. New Zealand recently reiterated its interest to conclude the bilateral FTA negotiations with India, in case the country decides not to move forward with RCEP[188], but India is not likely to react positively too soon. Finally, it is observed that imports have increased from the RTA partners like Japan and South Korea in the automobile sector, as well as non-RTA partner, China, given the eased entry owing to common UNECE 1998 membership. It can be argued that India judged the cost of not joining RCEP against the possible respite for the industrial sector.

The threat perceptions on the export front needs to be considered next. First, for example, even after joining UNECE 1998, India witnessed a decline in sectoral trade balance for auto products against China, Japan and South Korea (all contracting parties). In the ASEAN market, India's trade balance has also declined since 2014. Interestingly, in 2015, ASEAN members decided that their sectoral mutual recognition agreement (MRA) would be based on UNECE 1958 standards[189] and several countries have already started implementing these provisions[190]. This development would throw a major challenge for auto-components and vehicles exports from India, as Australia, Japan, New Zealand and South Korea, being members of both UNECE 1998 and 1958 provisions, would be indifferent to this policy

185 Sudip Chaudhuri, *Import Liberalisation and Premature Deindustrialisation in India*, Economic and Political Weekly, 64, Oct. 29, 2015.

186 Tanya Thomas, *Steel industry, hurt by Chinese imports, relieved on India not joining RCEP*, Live Mint (Nov. 5, 2019, 1:13 PM), https://www.livemint.com/politics/policy/steel-industry-hurt-by-chinese-imports-relieved-on-india-not-joining-rcep-11572939588486.html.

187 Dilip Kumar Jha, *Industry miffed as Govt mulls dairy import from New Zealand, Australia*, Business Standard (July 23, 2019 01:12 IST), https://www.business-standard.com/article/economy-policy/industry-miffed-as-govt-mulls-dairy-import-from-new-zealand-australia-119072000739_1.html.

188 ET Bureau, *New Zealand trade ministry allays India's dairy import fears*, Economic Times (Feb. 28, 2020 09:11 IST), https://economictimes.indiatimes.com/news/economy/foreign-trade/new-zealand-trade-ministry-allays-indias-dairy-import-fears/articleshow/74368893.cms.

189 *See* Manfred Lottig, *ASEAN MRA and Implementation Status* at 13, June 24, 2015, http://www.thaiauto.or.th/2012/Automotive-Summit/doc/ppt/2015/25-6-15/203_PM/10.ASEAN%20MRA%20and%20status%20of%20Implementation%20Rev%201.pdf.

190 ECON. RSCH. INST. FOR ASEAN AND EAST ASIA, Harmonization of Standards and Mutual Agreements on Conformity Assessment in Indonesia, Malaysia, Thailand and Vietnam, 7-8 (2016).

decision of ASEAN. China, despite being a member of UNECE 1998, is also on a superior technology plane, and therefore unlikely to face a major difficulty in complying with this transition. However, for India, this move may lower the expected exports to ASEAN further. Second, the proposal by the five ACTA member countries within RCEP (i.e., Australia, Japan, New Zealand, Singapore and South Korea) to make IPR provisions within the bloc 'business-friendly' makes India uncomfortable, given their earlier track record of targeting goods (particularly generic medicines) in transit. Inclusion of similar provisions in RCEP poses a potential threat for Indian exports in the future.[191] Finally, to compensate the modest performance of the merchandise exports, India needed a deeper commitment on service front in RCEP partners, which is not easily obtained.[192] The Indian decision therefore can also be interpreted as a function of unfulfilled export ambitions from the regional integration.

It is apparent that relatively modest competitiveness pattern has limited India from accessing the full benefits of preferential entry opportunities in RTA partner countries, which in turn caused the RCEP pull-out consideration to emerge. Even though India left RCEP for the time being, it still continues to be part of a complex RTA universe involving ASEAN (and separately with Singapore and Malaysia), Japan and South Korea. It has also recently explored the idea to have bilateral RTAs with Australia[193], EU[194], and the US[195]. Enhancing competitiveness therefore should be the priority for the country through a series of pending reforms, namely: strengthening the IPR regime[196], removing regulatory and administrative hassles and hidden costs, improving trade infrastructure[197] and the similar steps. These initiatives, once implemented, would improve the supply chains of Indian firms and the associated

191 MEDECINS SANS FRONTIERS, Regional Comprehensive Economic Partnership: Intellectual Property Chapter and the Impact on Access to Medicines, 41-42 (2016).

192 R. V. Anuradha, *RCEP: Redefining India's trade in services agenda*, Financial Express (Nov. 19, 2019 03:00), https://www.financialexpress.com/opinion/rcep-redefining-indias-trade-in-services-agenda/1768380/

193 Business Standard, *India, Australia agree to conclude free trade agreement by 2022-end*, 2 (Sept. 30, 2021 22:02 IST) https://www.business-standard.com/article/economy-policy/india-australia-agree-to-conclude-free-trade-agreement-by-2022-end.

194 Times of India, *Progress made on resumption of India-EU FTA negotiations, formal talks to start this month: EAM*, 2 (Sept. 8, 2021 23:58) https://timesofindia.indiatimes.com/india/progress-made-on-resumption-of-india-eu-fta-negotiations-formal-talks-to-start-this-month-eam/articleshow/86048487.cms.

195 Vikas Dhoot, *Piyush Goyal urges U.S. firms to push for trade agreement*, The Hindu, 1 (Sept. 29, 2021 23:53) https://www.thehindu.com/business/Industry/piyush-goyal-urges-us-firms-to-push-for-trade-agreement/article36743266.ece.

196 Khanna & Singh, *supra* note 13 at 52-54.

197 Pritam Banerjee, *Development of East Coast Economic Corridor and Vizag–Chennai Industrial Corridor: Critical Issues of Connectivity and logistics*, 6, 8-12, 14, 16-17, 22-23, (Asian Dev. Bank, South Asia Working Paper No. 50, 2017).

competitiveness gains would deepen their participation in the global IPNs.[198] Once India scales up the competitiveness ladder and starts gaining prominence in GVCs, the demand for re-joining Asian RTAs would emerge from within, which will politically be much easier for the policymakers to align with.

198 SAYON RAY & SMITA MIGLANI, GLOBAL VALUE CHAINS AND THE MISSING LINKS: CASES FROM INDIAN INDUSTRY, 256-59, LONDON ROUTLEDGE, 2018).

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

REALISTIC ASSUMPTIONS, ECONOMIC MODELS, AND THE ADMISSIBILITY OF EXPERT TESTIMONY IN THE CLASS ACTION LAWSUIT *DOVER V. BRITISH AIRWAYS*

Hannah Faulkner

INTRODUCTION

Dover v. British Airways is a breach of contract suit brought by lead plaintiff Russell Dover on behalf of an estimated 168,259 class members[1] who were part of the airline's frequent flyer program, the Executive Club.[2] Members of the club alleged that the airline imposed illegitimate fuel surcharges on rewards-redeemed flights, and thus violated the Executive Club contract.[3] While the airline's right to levy fuel surcharges ("YQ charges") was explicitly stated in the contract, plaintiffs asserted that the level of charges was not reasonably related to the actual cost of jet fuel.[4] Each party enlisted the testimony of economics experts to answer the central question of whether there existed a strong correlation between these YQ charges and the market price of jet fuel.[5] Econometricians Jonathan Arnold (representing the plaintiffs) and Andrew Hildreth (representing British Airways) reached opposite conclusions regarding whether this relationship existed.[6] In addition, plaintiff's expert Arnold proffered multiple damages models that estimated the class members' entitlement if British Airways was found to have breached the contract.[7] Cross-motions were filed to dismiss the expert testimony, involving challenges to the reliability of the expert's method and the realisticness of his model's assumptions.[8]

The use of economics experts—along with these challenges to their testimony and underlying models— is increasingly common in suits involving employment discrimination, business torts, fraud, antitrust violations, and property damage.[9] Experts use statistical techniques such as regression analysis to establish correlation between the defendant's harmful act and the

---

1 Dover v. British Airways, PLC, 321 F.R.D. 49, 53 (E.D.N.Y. 2017).

2 *Id.* at 52.

3 *Id.* at 52-53.

4 *Id.*

5 Dover v British Airways, PLC, 254 F. Supp. 3d 455, 465 (E.D.N.Y. 2017).

6 *Id.* at 460, 461-63.

7 *Id.* at 461-62.

8 *Id.* at 457-65.

9 Rebecca Haw Allensworth, *Law and the Art of Modeling: Are Models Facts?*, 103 GEO. L.J. 825, 835 (2015).

plaintiff's injury.[10] Econometric models also help experts perform complex valuation of assets and corporations as well as estimate the amount of damages.[11]

The intended audience for this expert testimony is the jury.[12] As fact-finders, jurors determine how much weight to give to evidence, which involves "questions of credibility and choice among competing inferences."[13] When confronted with these models showing causation, discrimination, or damages, the jury evaluates them in the context of other evidence and lay witness testimony.[14] The economic models and accompanying narrative therefore represent a key thread in the jury's understanding of the dispute.[15]

In addition to the increasing necessity of economic models,[16] motions to exclude expert testimony are becoming routine due to the high returns that exclusion can bring for the opposing party.[17] Experts and their models often comprise the only proof of causation and damages, so plaintiffs are far less likely to prevail without them.[18] Partial exclusion can "devastate" a plaintiff's case by increasing the relative strength of the defendant's evidence,[19] while full exclusion often results in summary judgement since the plaintiff has no proof of damages.[20]

Following motions to exclude expert testimony, judges engage in admissibility determinations that aim to establish the relevance and reliability of the testimony.[21] However, judges struggle to evaluate economic models for objective scientific validity because models are both science and art.[22]

---

10 John E. Lopatka & William H. Page, *Economic Authority and the Limits of Expertise in Antitrust Cases*, 90 CORNELL L. REV. 617, 687-88 (2005).

11 Anthony J. Casey & Julia Simon-Kerr, *A Simple Theory of Complex Valuation*, 113 MICH. L. REV. 1175, 1178 (2015).

12 *See* Jeff Todd, *An Interdisciplinary Perspective on Economic Models in Complex Litigation*, 46 HOFSTRA L. REV. 971, 1019-23 (2018).

13 Todd, *supra* note 12, at 988 (quoting Geoffrey C. Hazard, Jr. et al., CIVIL PROCEDURE 478, 485 (6th ed. 2011)); David L. Faigman, Christopher Slobogin, & John Monohan, *Gatekeeping Science: Using the Structure of Scientific Research to Distinguish Between Admissibility and Weight in Expert Testimony*, 110 NW. U. L. REV. 859, 861, n. 1, 884 (2016).

14 Jennifer L. Mnookin, *Atomism, Holism, and the Judicial Assessment of Evidence*, 60 UCLA L. REV. 1524, 1577 (2013).

15 John W. Hill et al., *Increasing Complexity and Partisanship in Business Damages Expert Testimony: The Need for a Modified Trial Regime in Quantification of Damages*, 11 U. PA. J. BUS. L. 297, 334 (2009).

16 *Id.* at 317-18 (noting that "expert testimony is highly desirable in cases involving business damages" and that some courts will not accept damages estimates from non-experts).

17 Roger D. Blair & Jill Boylston Herndon, *The Implications of Daubert for Economic Evidence in Antitrust Cases*, 57 WASH. & LEE L. REV. 801, 802 (2000).

18 *Id.*

19 Todd, *supra* note 12, at 996-97.

20 Mnookin, *supra* note 14, at 1569.

21 FED. R. EVID. 401-403; FED. R. EVID. 702.

22 *See infra* Part III.C.; *see also* Allensworth, *supra* note 9, at 830-834 (stating that this "straddle between art and science has made for the awkward and at times inconsistent treatment of modeling as

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

Economists wield substantial discretion in the construction of their models by choosing which variables to include and exclude, as well as what idealizations and omissions to make.[23] The realism of a model's simplifying assumptions are frequent points of attack, placing the judge in a position to weigh competing notions of realisticness.[24] This evaluation becomes problematic when the judge encroaches upon the jury's role by deciding the credibility of the testimony.[25]

Existing legal precedent provides little guidance for admissibility decisions on model-based testimony.[26] The vague language of Federal Rule of Evidence (hereinafter FRE) 702[27] and the misapplication of scientific standards to the non-scientific aspects of a model[28] leads to unpredictable and often unfounded decisions.[29] These admissibility determinations are high-stakes, often defining the outcome of multimillion-dollar cases such as *Dover v. British Airways*.[30]

Legal scholars have attempted to fill in the gaps by contextualizing models in the current legal framework, classifying them as issues of law to be decided by the judge or issues of fact to be weighed by the jury, or a combination of the two.[31] However, comments to FRE 702 advise courts to consider the relevant standards of an expert's field, so these paradigms are misguided because they do not consider economic models in the context of their own discipline.[32]

Heeding these instructions, Todd surveyed the methodological literature to arrive at an understanding of how economic models are built and used within a rhetorical context, and how this knowledge can inform admissibility

---

factual in the eyes of the law"); *and see* Mark Klock, *Contrasting the Art of Economic Science with Pseudo-Economic Nonsense: The Distinction Between Reasonable Assumptions and Ridiculous Assumptions*, 37 PEPP. L. REV. 153, 196-98 (2010).

23 Allensworth, *supra* note 9, at 829; *id* at 832 (noting that "like a mapmaker, a modeler makes choices about what are the essential elements (and what are inessential, such as mailboxes and trees) with reference to the task the model is to perform.").

24 Hill et al., *supra* note 15, at 330-32.

25 *See* Faigman et al., *supra* note 13, at 862 (noting how the complexity of scientific evidence creates confusion among courts regarding the boundary between admissibility and weight).

26 Joni Hersch & Blair Druhan Bullock, *The Use and Misuse of Econometric Evidence in Employment Discrimination Cases*, 71 WASH. & LEE L. REV. 2365, 2377 (2014).

27 *Id.*

28 Jeff Todd, *Realistic Assumptions in Economic Models*, 47 HOFSTRA L. REV. 231, 252 (2018).

29 Allensworth, *supra* note 9, at 863-64; *see also* Hill et al., *supra* note 15, at 311.

30 Todd, *supra* note 12, at 992-95 (citing, *inter alia,* Polymer Dynamics, Inc. v. Bayer Corp., 67 Fed. R. Serv. 201 (E.D. Pa. 2005); *see also* Nebraska Plastics, Inc. v. Holland Colors Americas, Inc., 408 F.3d 410, 422-23 (8th Cir. 2005).

31 Ronald J. Allen & Michael S. Pardo, *The Myth of the Law-Fact Distinction*, 97 NW. U. L. REV. 1769, 1769-70 (2003); Allensworth, *supra* note 9, at 864; Casey & Simon-Kerr, *supra* note 11, at 1187; *and see* D. H. Kaye, *The Dynamics of* Daubert*: Methodology, Conclusions, and Fit in Statistical and Econometric Studies*, 87 VA. L. REV. 1933, 1958-62 (2001); *and see* Hill et al., *supra* note 15, at 311.

32 *See* FED. R. EVID. 702 advisory committee's note to 2000 amendment.

decisions on model-based testimony in complex litigation.[33] In a subsequent article, he surveyed the literature on how the realisticness of simplifying assumptions affects a model's validity, a frequent point of contention among both economics methodologists and litigants.[34] This survey revealed that all models necessarily involve false or unrealistic assumptions, such as assuming away a factor that exists but is negligible in the real world.[35] However, targeting these assumptions as unrealistic disregards the reasons for which they are imposed, which may be justified for the purposes of isolation and abstraction.[36] Todd proposes a theoretical framework that synthesizes typologies from economics methodologists that classify different types of assumptions and realisticness.[37] The framework prescribes that judges evaluate the realism of assumptions relative to their context and purpose, rather than in isolation.[38]

In principle, the framework guides judges' decision-making regarding model-based testimony, which may lead to greater consistency and coherence of admissibility rulings.[39] Application to the judge's rulings in the British Airways class action lawsuit can show whether the framework is effective in practice. Further, *Dover* presents an opportunity to explore the unique dynamic of competing experts, a need that has been expressed in legal scholarship.[40]

Part II of this article summarizes the multi-decade debate among economics methodologists about what models are and how economists use them.[41] This Part also charts the parallel debate within the discipline regarding whether the realism of assumptions is relevant to a model's validity, as well as recent contributions to this "assumptions controversy" that classify types of assumptions and realisticness.[42] Part III describes the current legal framework for admissibility decisions, how these standards are lacking, and proposed solutions in legal scholarship.[43] Part IV describes the framework proposed by Todd and applies it to the competing expert testimony in *Dover v. British Airways*.[44] Part V evaluates the framework's applicability and

---

33 Todd, *supra* note 12, at 973-76.

34 *See infra* Part II.B., Part III.C.

35 Todd, *supra* note 28, at 272-75.

36 Uskali Mäki, *On the Method of Isolation in Economics*, 26 POZNAN STUD. PHIL. SCI. & HUMAN. 316, 329-30 (1992).

37 Todd, *supra* note 28, at 239-40.

38 *Id.* at 273-75.

39 *See id*.

40 Jeff Todd & R. Todd Jewell, Dubious Assumptions, Economic Models, and Expert Testimony, 42 DEL. J. CORP. L. 279, 320 (2018) (concluding that "Additional articles could explore assumptions in such multi-expert situations and what role argumentation and evidence play in resolving that battle.").

41 *See infra* Part II.

42 *See infra* Part II.

43 *See infra* Part III.

44 *See infra* Part IV.

usefulness for each expert's testimony.[45] Application to explicit assumptions that are challenged as unrealistic demonstrates a straightforward application of the framework, providing greater depth and precision to the judge's analysis.[46] Application to tacit assumptions such as choice of methodology requires more inference, but reveals that the framework is also useful for evaluating choices that are external to a model.[47] However, the latter application requires an understanding of the often ill-defined boundary between the scientific and artistic aspects of a model, a distinction that would need to supplement the framework in order to be applicable to a wide range of challenges to expert testimony.[48] This article concludes in Part VI.[49]

II. ECONOMIC MODELS IN THE METHODOLOGICAL LITERATURE

A. Models are Analogical Devices and Economists are Storytellers

Economic models are simplified representations of a more complex system,[50] but there is substantial methodological debate regarding the metaphysical relationship between models and the real-world systems they represent. A survey of the economics literature will illustrate how there is no single conception of what models are or how they should be built. This lack of consensus is not fatal, however, because one may distill general points of agreement which will help inform an analysis of economic models in the context of litigation.

Some methodologists posit a high degree of similarity between economics and other sciences. Founder of neoclassical economics Alfred Marshall noted that economic inquiries, like those in other sciences, aim to study an essential cause-and-effect relationship, under the condition that all other things are equal.[51] Mäki extends this reasoning to something more concrete, likening theoretical modelling to the material experiments that are characteristic of natural sciences.[52] The essential similarity between the two lies in the notion of manipulation.[53] Reality is so complex that the ceteris paribus

---

45 *See infra* Part V.

46 *See infra* Part IV.B.

47 *See infra* Part IV.C.

48 *See infra* Part V.

49 *See infra* Part VI.

50 Robert M. Solow, *How Did Economics Get that Way and What Way Did It Get?*, 126 DAEDALUS 39, 43 (1997).

51 *See* Mäki, *supra* note 36, at 317.

52 *See* Uskali Mäki, *Models are experiments, experiments are models*, 12:2 J. ECON. METHODOLOGY 303, 306-9 (2005).

53 *See id.* at 306.

condition, i.e. other things being equal, is not naturally occurring.[54] Thus, in order to isolate a certain causal relationship and examine its properties, investigators must impose a series of controls to craft an "artificial world" that is "free from [the] complications... of the rest of the world."[55] Scientists in laboratories impose these controls through material manipulation, whereas economists use idealizing assumptions to achieve the same effect, that is, of neutralizing or standardizing the myriad elements in a complex system.[56] As a result, Mäki sees models as "surrogate systems," the properties of which are examined directly in order to indirectly gain knowledge about the systems they represent.[57] This relationship is akin to animal subjects functioning as surrogates for human beings.[58]

Sugden disagrees with the isolationist approach, and instead aims to construct an autonomous model-world that parallels rather than simplifies the real world.[59] For this reason, Sugden's method is sometimes described as the constructionist approach, since it does not aim to mirror reality but instead crafts from scratch an explicitly counterfactual yet credible world.[60] More succinctly, these model-worlds are "imaginary but imaginable."[61] A model's credibility derives from the coherence of its assumptions and whether they construct a world that could be true, given our understanding of how the real world works.[62] Sugden also compares models to novels.[63] Realistic novels do not claim to be anything other than fiction, but they contain certain characters and situations that could conceivably be true.[64] Perhaps most importantly, the story and its characters' behaviors (like a good model's assumptions) cohere to logic and eal world causal processes.[65]

Cartwright writes that "models are like fables, and the lesson derived from the model is its moral."[66] The key is to translate the concrete, specific results of the model into abstract results that can be more generally applied to other cases.[67] In this way, the model is both true to reality and useful for

---

54 *See* Mäki, *supra* note 36, at 317.

55 Mäki, *supra* note 52, at 308.

56 *See id.*

57 *Id.* at 304.

58 *See id.*

59 *See* Robert Sugden, *Credible Worlds: The Status of Theoretical Models in Economics*, 7 J. ECON. METHODOLOGY 1, 25 (2000) (arguing that the "model world is not constructed by starting with the real world and stripping out complicating factors: although the model world is simpler than the real world, the one is not a simplification of the other.").

60 *See id.* at 25, 28.

61 Robert Sugden, *Credible Worlds, Capacities and Mechanisms*, 70 ERKENNTNIS 3, 5 (2009).

62 *See* Sugden, *supra* note 59, at 25.

63 *See id.*

64 *See id.*

65 *See id.* at 26.

66 Nancy Cartwright, *Models: Parables v Fables, in* BEYOND MIMESIS AND CONVENTION: REPRESENTATION IN ART AND SCIENCE 19, 26 (Roman Frigg & Matthew C. Hunter eds., 2010).

67 *Id.* at 28.

inductive inference.[68] Gibbard and Varian use similar literary language, describing some models as "caricatures" that do not purport to approximate the real world.[69] Instead, the model is a "deliberate distortion" of reality that aims to exaggerate or illuminate a particular feature of that reality.[70] This magnification helps the economist to "tell a simple story" about select aspects of the real world, rather than attempt to recount reality in all its complexity.[71]

Finally, Morrison and Morgan see models as autonomous instruments that nevertheless help mediate between theory and the real world.[72] Like any other tool, models must be "put to work, used, or manipulated" by an external entity in order to be useful.[73]

Clearly, economists lack a single conception of what model-building entails. Nevertheless, a few points of similarity may be gleaned from the divergent literature. Perhaps most importantly, models are tools that explain real-world phenomena through the mechanism of analogy. Whether they are called surrogate systems, parallel worlds, fables, or caricatures, economists clearly see models as inherently metaphorical devices that help shed light on real-world processes.[74] Though there may be gaps between the model-world and the real world it represents (due to isolation, distortion, and abstraction), the bridge between the two lies in inductive inference.[75] This process of induction requires taking the specific propositions of a model - say, a certain factor R causes a change in factor F - and generalizing those to more general situations, inferring that the same causal relationship exists in the real world.[76]

Crucially, models cannot make these "inductive leaps" themselves and require human interpretation to be useful.[77] In addition to interpretation, economists must communicate the quantitative results of a model as a qualitative, coherent narrative. The explanations that accompany models are essentially stories, with the modeler as a storyteller.[78] The modeler must

---

68 *See id.* at 29-30 (concluding that increasing the level of abstraction permits "generalizable conclusions" that are "true of new target situations").

69 Allan Gibbard & Hal R. Varian, *Economic Models*, 75 J. PHIL. 664, 673 (1978).

70 *Id.* at 673–676.

71 *Id.* at 674.

72 Margaret Morrison & Mary S. Morgan, *Models as Mediating Instruments*, in MODELS AS MEDIATORS: PERSPECTIVES ON NATURAL AND SOCIAL SCIENCE 10, 10 (Mary S. Morgan & Margaret Morrison eds., 1999).

73 *Id*. at 32.

74 *See* Cartwright, *supra* note 66 (fables); Gibbard & Varian, *supra* note 69 (caricatures); Mäki, *supra* note 52, at 304 (surrogate systems); Sugden, *supra* note 59 (parallel worlds).

75 *See* Sugden, *supra* note 61, at 4.

76 Sugden, *supra* note 59, at 20.

77 *Id.*

78 *See* Mäki, *supra* note 36, at 330-31; *see also* Mary S. Morgan, *Models, Stories and the Economic World*, 8 J. ECON. METHODOLOGY 361, 361 (2001) (describing how economists use models "to explain or to understand the facts of the world by telling stories about how those facts might have arisen."); *see id.* at 366 (writing that models and stories "go hand in hand").

communicate the model and its results to a particular audience, whether they be other economists, academics from other disciplines, or lay audiences.[79] There is also a rhetorical aspect to these stories, wherein economists aim to convince the audience of the model's credibility and its similarity with the real world.[80] As part of this persuasion, the modeler engages in "storied idealizations" to describe the reasons behind his simplifications and omissions.[81]

B. Simplifying Assumptions and Realisticness

While the metaphysical relationship between the real world and the model-world can be described in a variety of ways, it is undisputed that unrealistic assumptions are a ubiquitous and inescapable component of these model-worlds.[82] Simplification and idealization transform a complicated reality into a tractable, useful model that ignores irrelevant elements to shed light on a particular relationship or phenomenon.[83] Since the model-world is supposed to inform us about the real-world in some way, the issue of realisticness is perhaps the "most chronic ongoing methodological controversy in economics."[84] Historical approaches to this dispute have generally tended towards two methodological camps: instrumentalism and realism.[85] Instrumentalists are concerned with the realism of outputs, and posit that the accuracy of predictions is the only criterion by which a model should be judged.[86] Conversely, realists are focused on inputs and see verisimilitude or "truthlikeness" of assumptions as intrinsically valuable and desirable.[87]

---

79 *See* Mäki, *supra* note 36, at 330-31 (holding that the story attached to a model "may vary somewhat from audience to audience.").

80 *See* Itzhak Gilboa et al., *Economic Models as Analogies*, 124 ECON. J. F513, F518 (2014) (stating that "the similarity judgement is often hinted at by the economist," but the audience or "readers" of a model may not necessarily agree with these judgements).

81 Mäki, *supra* note 36, at 330-31 (writing that "storied idealizations" are particularly important for audiences of non-economists).

82 *See* David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, in FED. JUDICIAL CTR., REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 211, 272 (3d ed. 2011) (noting that from a certain perspective, models are simply a "set of assumptions"); Uskali Mäki, *Reorienting the Assumptions Issue*, in NEW DEVELOPMENTS IN ECONOMIC METHODOLOGY 236, 241 (Roger Backhouse ed., 1994).

83 *See* Uskali Mäki, *Aspects of Realism about Economics*, 13 THEORIA 301, 308 (1998).

84 Mäki, *supra* note 36, at 319.

85 *See* Bruce J. Caldwell, *A Critique of Friedman's Methodological Instrumentalism,* 47 S. ECON. J. 366, 367 (1980).

86 *See id.*

87 *See* Mäki, *supra* note 82, at 240.

1. The Origins of the Assumptions Controversy

The most notable contribution to this debate is staunch instrumentalist and Nobel Prize laureate Milton Friedman, and his essay, *The Methodology of Positive Economics*.[88] In the essay Friedman proclaims that unrealistic assumptions are not an unfortunate byproduct of simplification, but rather a necessary and welcome element of important theories.[89] Important theories should "explain much by little," he says, and in general, "the more significant the theory, the more unrealistic the assumptions."[90] Accordingly, whichever assumptions yield those predictions - no matter their degree of "conformity [] to 'reality'"[91] - have proven themselves to be "sufficiently good approximations for the purpose at hand."[92]

This bold thesis provoked the response of Nobel Prize winner and realist Paul Samuelson, who dubbed Friedman's "principle of unreality" the "F-Twist," and used deductive logic to show that false assumptions necessarily imply false conclusions.[93] While Samuelson concedes that models cannot perfectly mirror reality, he sees lack of realism as a defect rather than a virtue and denounces Friedman's flagrant disregard for realism in the name of parsimony.[94] Samuelson feared that the tolerance (or celebration) of unrealistic assumptions could be a slippery slope into complete neglect of empirical validity.[95] He admits that some abstract models may have a certain "psychological usefulness" for understanding some latent patterns of reality, however, this usefulness is entirely different from the empirical accuracy that Friedman claimed to achieve with his instrumentalist models.[96]

2. A "Reorientation" of the Assumptions Controversy

Following the polarized debate between Friedman and Samuelson, the assumptions controversy has evolved beyond the binary issue of whether or

---

88 *See* MILTON FRIEDMAN, *The Methodology of Positive Economics, in* ESSAYS IN POSITIVE ECONOMICS 3 (1953).

89 *See id.* at 14.

90 *Id.*

91 *Id.*

92 *Id.* at 15 (asserting that the appropriate test for an assumption's realism is the accuracy of predictions it yields).

93 Paul Samuelson, *Problems of Methodology—Discussion*, 53 AM. ECON. REV. 227, 232-36 (1963).

94 *See* Paul Samuelson, *Theory and Realism: A Reply*, 54 AM. ECON. REV. 736, 736 (1964) (arguing that "the doughnut of empirical correctness in a theory constitutes its worth, while its hole of untruth constitutes its weakness."); *see id.* at 736 (calling it a "monstrous perversion of science to claim that a theory is *all the better for its shortcomings*") (emphasis in original).

95 *See* Samuelson, *supra* note 93, at 236.

96 *Id.*

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

not assumptions need to be realistic. More recent developments recognize that assumptions should be evaluated relative to their context and purpose, rather than in isolation.[97] As a result, several methodologists have attempted to categorize assumptions based on this more nuanced understanding.[98]

Alan Musgrave aimed to "un-twist" Friedman's F-Twist by specifying three main types of assumptions: negligibility, domain, and heuristic.[99] Negligibility assumptions are statements that a certain factor X has no effect – or at least no detectable effect – on Y, the phenomenon under study.[100] As a result, the omission of factor X from the model will not substantially change its results.[101] An example would be Galileo's assumption of zero air resistance when investigating the motion of free-falling objects.[102] It would be "plain silly," Musgrave says, to discount Galileo's theory simply because air resistance does exist and the objects were not, in fact, falling through a vacuum.[103] Instead, the proper focus should be on Galileo's statement about the negligibility of air resistance on the object of study, which is a potentially true statement.[104]

Domain assumptions specify where a theory may be applied.[105] In contrast to the previous type, factor X admittedly has non-negligible effects on Y, so the theory only applies when the factor is absent.[106] Musgrave states that domain assumptions should be true of as many actual situations as possible, because if domain assumptions are never true, they can never be tested, and the theory or model loses its utility.[107]

Finally, heuristic assumptions are early simplifications that ease the logical development of a theory.[108] Musgrave provides an example from physics: Newton's early approximations assumed that only one planet orbited the sun

---

97 *See, e.g.,* Yew-Kwang Ng, *Are Unrealistic Assumptions/Simplifications Acceptable? Some Methodological Issues in Economics*, 21 PAC. ECON. REV. 180, 181-182 (2016).

98 *See, e.g.,* Frank A. Hindriks, *Tractability Assumptions and the Musgrave-Mäki Typology*, 13 J. ECON METHODOLOGY 401 (2006); Uskali Mäki, *Kinds of Assumptions and Their Truth: Shaking an Untwisted F-Twist*, 53 KYKLOS 317 (2000); Alan Musgrave, *"Unreal Assumptions" in Economic Theory: The F-Twist Untwisted*, 34 KYKLOS 377 (1981).

99 Musgrave, *supra* note 98, at 378-82.

100 *See id.* at 378.

101 *See id.* at 380 (describing how the truth of negligibility assumptions becomes apparent by "examining the consequences of the theory in which they are embedded.").

102 *See id.* at 378.

103 *Id.* at 379.

104 *See id.* at 380 (writing that negligibility assumptions are true descriptions of reality because they "do not assert that present factors are absent but rather that they are 'irrelevant for the phenomena to be explained'").

105 *Id.* at 381.

106 *See id.*

107 *See id.* at 382 (negating Friedman's claim that "the more significant the theory, the more unrealistic the assumptions" because the significance of a theory is dependent on how widely it can be applied, i.e. how often its domain assumptions are true).

108 *See id.* at 383.

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

and did not take into account the effects of inter-planetary gravitational forces.[109] Implicit in these assumptions is a promise to relax them later on, as they are only intermediate steps towards more precise predictions.[110] Since they are only temporary, some descriptively false assumptions are permissible.[111]

To understand how each of these assumptions functions, consider how the same idealization of factor X can take on various interpretations:

> An economist who says 'assume the government has a balanced budget' may mean that any actual budget imbalance can be ignored because its effects on the phenomena he is investigating are negligible. But he may also mean precisely the opposite: that budget imbalance would have significant effects, so that his theory will only apply where such an imbalance does not exist.[112]

A heuristic assumption would be to assume the government has a balanced budget at first, with a subsequent theory that takes the possibility of budget imbalance into account.[113] In all three cases, the status of the government's budget is "assumed away," (factor X is absent) but the implications of each type of assumption represent critical differences that are often imperceptible.[114] If a successor theory were to build upon this economist's model, it is crucial that he know whether continuing to assume a balanced budget is appropriate.[115] An applicability assumption that is misconstrued as a negligibility assumption may cause a violation of the proper domain, resulting in an invalid theory.[116] Therefore, Musgrave urges economists to be explicit regarding which type of assumptions they are imposing.[117]

Uskali Mäki supplemented this typology, with improvements that were both semantic and material.[118] Musgrave's heuristic assumption was replaced by an essentially identical "early-step" assumption.[119] Mäki also clarified that Musgrave's domain assumption is simply one component of what is in fact an applicability assumption.[120] He notes that the domain assumption merely identifies the relevant domain, while the higher-level applicability

---

109 *See id.*

110 *See id.* (naming this process "a *method of successive approximation*") (emphasis in original).

111 *See id.*

112 *Id.* at 381 (emphasis in original).

113 *Id.* at 386.

114 *Id.* at 381 (noting that identical phrasing for the different types of assumptions can mask their divergent implications).

115 *See id.*

116 *See id.* at 385 (speculating on whether changes in the status of assumptions (*e.g.*, from negligibility to domain) have gone unnoticed in the development of economic theory at large).

117 *Id.*

118 Mäki, *supra* note 98, at 317-18.

119 *Id.* at 325.

120 Id.

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

assumption does the work in restricting the theory to only that domain.[121] As a result, applicability assumptions typically involve an if-then formulation: If the domain assumption is true, then the theory applies.[122] As for the third class of assumptions, Mäki cautioned against the conflation of negligibility with undetectability, the latter of which he feared Musgrave was actually referencing.[123]

On the topic of realism, Mäki feared that Musgrave's artful paraphrasing is too flexible, to the point where nearly any statement can become "a potentially true assertion if it is suitably 'meta-paraphrased'."[124] Mäki aims to limit the powers of meta-statements by requiring that they transform a statement into a "factual claim" about economic reality.[125]

In the latest contribution to the assumptions typology, Frank Hindriks introduces the notion of a tractability assumption, the successor of "heuristic" or "early-step" assumptions.[126] However, this change was not mere re-packaging of the previous methodologists' terminology. Instead, Hindriks' addition better encompasses a third primary reason why an economist might impose a given assumption. Tractability assumptions arise when a problem is unmanageable but-for a certain simplification.[127] Hindriks distinguishes between two types of tractability: theoretical and empirical.[128] The former is reminiscent of Musgrave's heuristic assumption and refers to the logical development of a theory.[129] Certain "exogenous constraints," such as the level of sophistication of mathematics, may necessitate a theoretical tractability assumption.[130] Such an assumption may also be imposed due to limitations on cognitive capacity, both of the theorists and their audiences alike.[131]

Empirical tractability, on the other hand, concerns more pragmatic constraints.[132] Frequently, data are not available for a given variable or factor X, either due to non-collection or confidentiality reasons.[133] Even when data are available, their method of collection may be inconsistent across time, making

---

121 *Id.*

122 *See id.* at 323.

123 *Id.* at 320.

124 *Id.* at 331.

125 *Id.* at 331-32.

126 Hindriks, *supra* note 98, 411-14.

127 *Id.* at 412 (stating that because tractability assumptions involve non-negligible factors, "one would prefer to avoid relying on it, if it were not for the fact that it makes the problem under investigation (more) tractable."); Frank A. Hindriks, *Unobservability, Tractability and the Battle of Assumptions*, 12 J. ECON. METHODOLOGY 383, 399 (2005) (asserting that despite their potentially "distorting effects," the imposition of tractability assumptions is "usually unavoidable").

128 Hindriks, *supra* note 98, at 413.

129 *Id.*

130 *Id.* at 414.

131 *Id.*

132 *Id.*

133 *Id.*

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

their use in an empirical model problematic.[134] Another empirical tractability concern is unobservability, which is often the case when theoretical constructs in economics have no corresponding real-world data.[135] As such, both theoretical and empirical models frequently incorporate these assumptions in order to reduce the number of unknown variables and make models more tractable.[136]

Both Musgrave and Mäki recognize that the content of an assumption and its purpose constitute two separate concepts whose "truth values...behave differently."[137] While Musgrave and Mäki hint at the dual identity of assumptions, Frank Hindriks articulates this distinction by introducing first-order and second-order assumptions.[138] The first-order assumption is an idealization statement concerning factor X – whether it is absent, constant, or infinite in the model.[139] The second-order assumption is a meta-statement that identifies the purpose of imposing the first-order assumption.[140] While the first-order assumption concerns only the model-world, the second-order assumption is a statement about the real-world.[141] For example, a negligibility assumption might take the following form:

> [A] Factor F is absent or has no effect on the phenomenon under investigation.
>
> [N] The factor F mentioned in first-order assumption A has a negligible effect on the phenomenon under investigation relative to the purpose for which the theory is used.[142]

This formulation demarcates the instances in which the truth-value of an assumption is an important concern.[143] First-order assumptions are often false or unrealistic, but this is unproblematic so long as the second-order assumption is approximately true relative to its purpose.[144] For example, assuming that air resistance is absent is a descriptively false first-order

---

134 *Id.*

135 Hindriks, *supra* note 127, at 399 (2005) (providing an example: "The value of marginal cost cannot easily be computed from cost data that are reported. Economists often assume that the relation between marginal cost and the number of products produced is nonlinear. This means that marginal cost is not the same as average cost").

136 *Id.* at 392 (emphasizing that "tractability is a matter of solubility or of the efficiency of a solution.").

137 Mäki, *supra* note 98, at 325; Musgrave, *supra* note 98, at 380 (differentiating between the absence of a factor as a "descriptively false" statement versus the negligibility of that factor as a "descriptively realistic" statement).

138 Hindriks, *supra* note 98, at 406.

139 *Id.* at 407.

140 *Id.*

141 *Id.*

142 *Id.*

143 *See id.* at 421.

144 *See id.* at 410 (rephrasing Musgrave's argument).

assumption.[145] The second-order assumption stating that air resistance has a negligible effect may be approximately true, as in the case of a falling baseball.[146] However, this negligibility assumption may be a false in other contexts, such as a falling feather.[147]

The assumptions typology demonstrates how a lack of overt realism in face-value assumptions need not be in conflict with a realist approach to economics.[148] The framework thus offers a sort of reconciliation between the realist and instrumentalist camps; on one hand, economists may use simplifying first-order assumptions as the tool that they are, and models will not be held to the impossible standard of being "photographic reproduction[s]" of reality.[149] On the other hand, realistic second-order assumptions ensure that a model is still meaningfully connected to reality. The requirement that the second-order assumption be at least approximately true may allay Samuelson's fears concerning the descent of economic science into empirical invalidity.[150] It also prevents indiscriminate acceptance of unrealistic assumptions, a concern had by many following Friedman's seminal essay.[151]

Most importantly, these methodologists approach the realisticness of assumptions as a dynamic phenomenon rather than a "dichotomous notion."[152] This analysis can be further developed by an awareness of the various interpretations that the terms "realistic" and "unrealistic" may take.

Mäki sees the debate about realism of assumptions as "plagued by multiple ambiguity."[153] The culprits of this ambiguity, he says, are binary thinking and lack of specificity in language.[154] As a remedy, he catalogs different kinds of realisticness which include truth, confirmability, plausibility and partiality, among others.[155]

Perhaps the most obvious conception of realisticness is truth; truth is self-evident and its antithesis is falsehood.[156] In this sense, a statement is

---

145 *See* Musgrave, *supra* note 98, at 378 n.2.

146 *See id.*

147 *See* Gibbard & Varian, *supra* note 69, at 671.

148 *See* Tarja Knuuttila, *Isolating Representations versus Credible Constructions? Economic Modelling in Theory and Practice*, 70 ERKENNTNIS 59, 61; *see also* Todd, *supra* note 28, at 269 (summarizing that "Although idealization involves some false assumptions, that does not undermine the truth of the model").

149 Friedman, *supra* note 88, at 35 (contrasting the instrumentalist and realist approaches); Mäki, *supra* note 52, at 308 (emphasizing that "Unrealistic assumptions are the indispensable tools of the experimental theorist.").

150 *See, e.g.,* Mäki, *supra* note 98, at 332 (clarifying that paraphrased assumptions must involve factual claims about economic reality).

151 *See* Hindriks, *supra* note 98, at 410-411 (noting how the truth of second-order assumptions is important for a theory's validity).

152 *E.g.,* Mäki, *supra* note 36, at 324.

153 *Id.* at 320.

154 Mäki, *supra* note 82, at 239.

155 *Id.* at 241-43.

156 *Id.* at 241-42.

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

realistic if true and unrealistic if false.[157] However, the truth is not always observable and may be difficult to ascertain after-the-fact.[158] Empirical evidence can provide hints about this truth, but belongs to a different type of realisticness, which is confirmation and disconfirmation.[159] This type concerns matters that are not only observable, but testable and confirmed by empirical evidence.[160] Confirmability is often conflated with truth but the two are not synonymous; evidence can speak for a false statement (confirmable but not true) and not all truths are supported by observable evidence (true but not confirmable).[161] A third type of realisticness is plausibility, which is a matter of being believed by people.[162] The criterion for realisticness here is not evidence but instead human logic and reasoning, so a representation may be realistic if plausible and unrealistic if implausible.[163]

The final class of realisticness concerns partiality, of which there are a few sub-types. Partiality may refer to isolation, which "focuses on the influence of only one factor...to the exclusion of others."[164] Another subset of partiality is abstraction, where a universal or quasi-universal concept is stripped from its particularities.[165]

Partiality can relate to realisticness in two opposite ways. In one sense, a partial representation may be deemed unrealistic because it is not comprehensive (in the case of isolation)[166] or concrete (in the case of abstraction).[167] Alternatively, partiality may help a modeler "attain the truth about the essential features" of economic phenomena, and thus be more "realistic" than a model weighed down by innumerable and irrelevant details.[168] "An isolating theory or statement is true if it correctly represents the isolated essence of the object; otherwise it is false."[169] Similarly, abstraction can help "facilitate the attainment of truth" by permitting greater scope.[170]

Mäki's enumeration of various types of realisticness provides yet another dimension to the previously superficial controversy surrounding the realism of assumptions. For one, an assumption may be simultaneously

157 *Id.*

158 *Id.* at 242.

159 *Id.*

160 Mäki, *supra* note 82, at 242.

161 *Id.*

162 *Id.*

163 *Id.*

164 *Id.* at 243.

165 Mäki, *supra* note 36, at 322 (referencing the example of a production function, where an 'L' represents the aggregate labor input without any reference to "spatio-temporally specified instances of labor.").

166 *Id.* at 321.

167 *Id.* at 323.

168 Mäki, *supra* note 83, at 311.

169 Mäki, *supra* note 36, at 344.

170 Jack Melitz, *Friedman and Machlup on the Significance of Testing Economic Assumptions*, 73 J. POL. ECON. 37, 41 (1965).

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

realistic in one sense and unrealistic in another, such as a false assumption that may still be plausible.[171] Furthermore, realisticness does not always have to be binary as in the case of truth and falsehood or confirmation and disconfirmation, as plausibility and partiality can exhibit differences of degree.[172]

These various conceptions of realism are also present in the context of litigation. In the "quest for truth," juries make plausibility judgments on competing testimony and parse evidence that may confirm or disconfirm certain statements.[173]

III. ECONOMIC MODELS IN LITIGATION

A. *The Critical Role of Economics Expert Testimony in Complex Litigation*

Economics experts are frequently called upon to certify a class, prove causation, and estimate damages.[174] Their model-based testimony is prevalent in various types of cases involving employment discrimination, business torts, fraud, antitrust violations, and property damage.[175] Proving causation is no straightforward task given the myriad of confounding variables that can affect an employment decision or a firm's share price and profits.[176] Because economists do not have the benefit of controlled experiments,[177] they must use econometrics - the application of statistics to analyze economic data - to study phenomena like correlation and causation.[178] These econometric models therefore help the jury make sense of vast amounts of disorderly data.[179]

Statistical techniques such as regression analysis can establish and isolate the causal link between an independent variable - the defendant's alleged wrongful conduct - and a dependent variable, the plaintiff's injury.[180] Experts

---

171 *See* Mäki, *supra* note 82, at 241-43; *see also* Mäki, *supra* note 36, at 346 (describing how a theory can also be simultaneously realistic (it shows nothing-but-the-truth) and unrealistic (it does not show the whole truth)); Todd, *supra* note 28, at 263, 275, 278-79.

172 Mäki, *supra* note 36, at 321.

173 Casey & Simon-Kerr, *supra* note 11, at 1184.

174 *See* Allensworth, *supra* note 9, at 835; Hersch & Bullock, *supra* note 26, at 2373-76.

175 *See* Allensworth, *supra* note 9, at 835.

176 *See* Todd & Jewell, *supra* note 40, at 286-87.

177 *See* Malcolm B. Coate & Jeffrey H. Fischer, *Daubert, Science, and Modern Game Theory: Implications for Merger Analysis*, 20 SUP. CT. ECON. REV. 125, 151 (2012).

178 *See generally* Alan O. Sykes, *An Introduction to Regression Analysis* (Coase-Sandor Inst. for Law & Econ., Working Paper No. 20, 1993).

179 *See* Todd, *supra* note 28, at 243; Todd & Jewell, *supra* note 40, at 286-87 (describing how experts must sort through "voluminous data, such as market information showing prices and sales, or financial records showing revenues and expenses" as well as account for the variable of time).

180 *See* Lopatka & Page, *supra* note 10, at 687 (stating that economic models help determine whether "the alleged harm bears the necessary causal link" in antitrust inquiries).

also perform complex valuation of assets and corporations for cases involving business torts.[181] Additionally, they use empirical models to construct a counterfactual past that estimates what a plaintiff's position would have been but-for the defendant's wrongful conduct.[182] These estimations provide a basis for calculating the amount of damages, which is the difference between the plaintiff's current position and their but-for position.[183]

In sum, economics experts use their knowledge and technical expertise to "fill[] that evidentiary void" between complex, raw data and actionable information that the jury can understand and evaluate.[184] Econometric models frequently comprise the only proof of causation and damages, so many cases' outcomes are contingent on the jury seeing this testimony.[185]

B. *The Jury's Role in Assessing the Credibility of an Expert's Story*

As factfinders, juries are tasked with determining the facts in issue of a case given the available evidence.[186] By assessing witness credibility and weighing competing evidence, juries decide whether the defendant is liable and if so, the amount of damages.[187] However, jurors do not approach evidence with a blank state.[188] Instead, they filter the evidence through the lens of their past experiences.[189] These experiences engender a common sense and intuition about how the world works.[190] With the advantage of numerosity, juries are the entity best equipped to make common-sense judgments regarding "credibility and choice among competing inferences."[191]

Furthermore, juries evaluate evidence holistically rather than in isolation, searching for continuity among the various elements presented to them.[192] With these evidentiary fragments, the jury constructs "alternative

---

181 *See* Casey & Simon-Kerr, *supra* note 11, at 1178.

182 *See* Allensworth, *supra* note 9, at 837.

183 *See* Roger D. Blair & William H. Page, *"Speculative" Antitrust Damages*, 70 WASH. L. REV. 423, 435-36 (1995); *see also* Lopatka & Page, *supra* note 10, at 687 (writing that the "causal link [between injury and harm] forms the basis for any damage model.").

184 Casey & Simon Kerr, *supra* note 11, at 1179; *see* FED R. EVID. 702(a) ("the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue").

185 *See* Blair & Herndon, *supra* note 17, at 802.

186 *See* Casey & Simon-Kerr, *supra* note 11, at 1185.

187 *See* Faigman et al., *supra* note 13, at 884.

188 *See* Neil Vidmar & Shari Seidman Diamond, *Juries and Expert Evidence*, 66 BROOK. L. REV. 1121, 1137 (2001).

189 *See id.* at 1138; Mnookin, *supra* note 14, at 1540-41.

190 *See* Lisa Kern Griffin, *Narrative, Truth, and Trial*, 101 GEO. L.J. 281, 294 (2013).

191 Allensworth, *supra* note 9, at 848-49; *see also* Faigman et al., *supra* note 13, at 884 (specifying the jury's role as the "assessors of witness credibility").

192 Mnookin, *supra* note 14, at 1540; *see also* Griffin, *supra* note 190, at 286 (discussing how different pieces of evidence "interact in ways that alter their individual significance").

interpretations, or 'stories,' about the events that led to the dispute now on trial."[193] Each party also has a story about these events supported by their respective evidence.[194] The parties present these opposing narratives to the jury, and use rhetoric and persuasion to convince this audience to share their respective conceptualizations of reality, essentially, whether or not the defendant caused harm to the plaintiff. [195]

This model of jury factfinding is no different for expert testimony. Economics experts present their models to the jury and aim to persuade this audience of their model's credibility.[196] Juries use their past experiences as "buyers and sellers, parties to contracts, and business owners" to gauge the plausibility of the expert's testimony.[197] Because the jury is a lay audience, experts rely on qualitative explanations and natural language to communicate the model's results.[198] The role of economists as storytellers becomes especially prominent in the context of litigation.[199] In isolation, models are but "skeletal representations" whose substantive contributions are revealed only when they are accompanied by the modeler's story.[200] The expert must explain his idealizations and omissions,[201] and aims to convince the audience that the model bears sufficient resemblance to relevant aspects of the real world.[202] If the expert is successful in his storytelling role the jury will accept the conclusions of his model, such as the causal link between the defendant's conduct and the plaintiff's injury, or a given damages estimate.[203]

Rhetoric and persuasion are particularly important when juries are confronted with dueling experts, as parties aim to dissuade the jury from accepting the opposing expert's model and conclusions.[204] The jury may or may not agree with these attacks, so each expert must defend his modelling choices in order to maintain his story's credibility in the eyes of the factfinder.[205]

While expert testimony is a critical piece of evidence in many lawsuits, it is but one element embedded in a greater narrative that each party

193 Vidmar & Diamond, *supra* note 188, at 1138.

194 *See* Lopatka & Page, *supra* note 10, at 622-24.

195 *Id.* at 622.

196 *See* Todd, *supra* note 12, at 1018-22.

197 *Id.* at 1034.

198 *See* Mäki, *supra* note 36, at 330-31; Morgan, *supra* note 78, at 361, 366.

199 *See* Todd, *supra* note 12, at 1018-20.

200 Mäki, *supra* note 36, at 330-31.

201 *See id.*

202 *See* Gilboa et al., *supra* note 80, at F518.

203 *See* Lopatka & Page, *supra* note 10, at 687; Casey & Simon-Kerr, *supra* note 11, at 1178; Allensworth, *supra* note 9, at 837; Blair & Page, *supra* note 183, at 435-36.

204 *See* Hill et al., *supra* note 15, at 342 (referencing the "battle of experts"); *see id.* at 358-64 (describing courts' struggles with expert partisanship).

205 *See* Todd & Jewell, *supra* note 40, at 313 (writing that the plausibility of assumptions partly rests on the "persuasiveness of the [expert's] reasoning").

constructs.[206] Econometric models showing causation and damages are thus evaluated in relation to other evidence and lay witness testimony.[207]

C. *The Ill-Defined Gatekeeper Role for Judges Ruling on Admissibility*

Before a jury can assess the weight of expert testimony, the judge must first act as a "gatekeeper" and evaluate the evidence for admissibility.[208] This division of responsibility aims to prevent scientifically invalid expert testimony from reaching the jury and skewing verdicts.[209] Admissibility determinations thus require weighing the relative competency of the jury in evaluating complex scientific evidence.[210]

In the case of model-based testimony, delineating the separate domains of judge and jury is both important and troublesome. If the admissibility threshold is too low, judges may admit "junk science" testimony that misleads or "bamboozles" the jury.[211] Conversely, a high admissibility threshold may cause a court to "substitute its judgment for the jury" in deciding issues of credibility.[212]

Furthermore, admissibility rulings are increasingly frequent and high-stakes decisions.[213] Given the vital role of economic models in showing causation, establishing injury, and estimating damages, the absence of such evidence can be outcome-determinative.[214] One possible outcome is the case never makes it to a jury. An excluded model may mean the plaintiff is unable to establish causation between alleged misconduct and personal or financial harm.[215] Without sufficient proof of causation, the judge may grant summary judgment to the defendant.[216] Even when a suit does make it to the jury, excluded expert testimony constitutes a significant missing "thread" in the party's case.[217] Admissibility rulings allow a judge to control the "evidentiary landscape" that a jury may use in constructing a coherent narrative of what

206 *See* Hill et al., *supra* note 15, at 334.
207 Mnookin, *supra* note 14, at 1577; Vidmar & Diamond, *supra* note 188, at 1138.
208 Faigman et al, *supra* note 13, at 861.
209 *Id.* at 862.
210 *Id.* at 884.
211 Lloyd, *supra* note 31, at 423.
212 Mnookin, *supra* note 14, at 1570.
213 *Id.* at 1569.
214 *See* Blair & Herndon, *supra* note 17, at 802 (describing how a plaintiff's case may "evaporate" without expert testimony that shows injury from anticompetitive behavior or provides damages estimates); Mnookin, *supra* note 14, at 1569 (noting that admissibility determinations are often case dispositive).
215 *See* Mnookin, *supra* note 14, at 1569.
216 *Id.*
217 Hill et al., *supra* note 15, at 334.

happened.[218] Even partial exclusion can mean that the relative weight between opposing expert testimony becomes lopsided, which can result in a party losing its case.[219]

Admissibility versus weight is clearly a critical distinction, but the boundary between the two remains fuzzy for model-based testimony.[220] An overview of the existing rules and precedents will illuminate how current standards are lacking.

1. Shortcomings of Current Evidence Law and Case Law on Admissibility

The first element a judge considers when evaluating evidence is relevance, that is, whether it "has any tendency to make a fact more or less probable than it would be without the evidence."[221] If relevant, FRE 402 follows a principle of "general admissibility"[222] where evidence should only be excluded if it is "unfairly prejudicial, misleading, a waste of time, deceptive, redundant, or unreliable."[223] Expert testimony in particular has its own set of parameters in evidence law.[224] FRE 702 states that expert testimony may be admitted if:

> (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
>
> (b) the testimony is based on sufficient facts or data;
>
> (c) the testimony is the product of reliable principles and methods; and
>
> (d) the expert has reliably applied the principles and methods to the facts of the case.[225]

This rule was amended in 2000 to incorporate a trilogy of Supreme Court cases that has culminated in the *Daubert* standard.[226] *Daubert v. Merrell Dow Pharmaceuticals, Inc*. was the first of these cases, and involved a

---

218 Lopatka & Page, *supra* note 10, at 619 (explaining that "a jury's evaluation of conflicting economic opinions rarely decides cases because federal judges' choices limit the scope and force of expert testimony."); *see* Mnookin, *supra* note 14, at 1542.

219 *See* Todd, *supra* note 12, at 997.

220 *See* Faigman et al., *supra* note 13, at 884.

221 FED. R. EVID. 401.

222 FED. R. EVID. 402.

223 Faigman et al., *supra* note 13, at 875.

224 *See* FED. R. EVID. 702–705.

225 FED. R. EVID. 702.

226 FED. R. EVID. 702 Committee Notes on Rules—2000 Amendment.

mother's allegations that a prescription drug had caused her child to have birth defects.[227] The plaintiff's expert offered testimony that suggested causation, but the evidence was declined because it was based on methods that were not accepted by the general scientific community.[228] The *Daubert* ruling thus established that expert testimony must adhere to the principles, methods, and procedures of the scientific method.[229] The Court also issued a series of flexible and non-exhaustive "Daubert factors" that a judge may consider in evaluating expert testimony.[230] Importantly, this ruling restricted the judge's gatekeeping function to an expert's principles and methodology, reserving the conclusions for the jury to evaluate.[231]

The second case in the *Daubert* trilogy, *General Electric Co. v. Joiner*, involved allegations that chemical exposure at the workplace had accelerated the onset of the plaintiff's lung cancer.[232] While the reliance on animal studies was an accepted methodology, the Court still excluded the testimony because the studies were insufficient grounds for the expert's conclusion.[233] The *Joiner* ruling consequently expanded the judge's range of discretion by allowing the perceived gap between methodology and conclusions to be a basis for exclusion.[234]

The final *Daubert* case reinforced and supplemented this finding. A judge declined to admit expert testimony in *Kumho Tire Co. v. Carmichael*, where an engineer inspected an allegedly defective tire and concluded that the defect had caused the tire to explode.[235] While visual and tactile examination was an acceptable method, there was too great of an analytical gap between the method and the conclusion that the defect caused the blow out.[236] *Kumho* thus reinforced that an expert's methods should not only be reliable in the abstract, but must also be reliable in context and application.[237]

While the purpose of the *Daubert* standard was to protect juries from scientifically invalid evidence, some legal commentators fear that it set the admissibility threshold too high.[238] In evaluating the "reliability" of an

227 *Daubert v. Merrel Dow Pharmaceuticals, Inc.,* 509 U.S. 579 (1993).

228 *Id*.

229 *Id.* at 589-90.

230 *Id.* at 593-94 (listing these factors: 1. Whether the theory or technique in question can be and has been tested; 2. Whether it has been subjected to peer review and publication; 3. Its known or potential error rate; 4.The existence and maintenance of standards controlling its operation; 5. Whether it has attracted widespread acceptance within a relevant scientific community).

231 *Id.* at 595.

232 *General Electric Co. v. Joiner*, 522 U.S. 136 (1997).

233 *Id.*

234 *Id.* at 146 (holding that "conclusions and methodology are not entirely distinct from one another" and that a "court may conclude that there is simply too great an analytical gap between the data and the opinion proffered.").

235 *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999).

236 *Id.* at 139.

237 Faigman et al., *supra* note 13, at 872-73 (citing FED. R. EVID. 702(d)).

238 *See* Mnookin, *supra* note 14, at 1570; Todd, *supra* note 12, at 976, 1024-25.

expert's conclusions, courts may inadvertently encroach upon the jury's domain, which is to make credibility judgments amid competing testimony.[239] The *Daubert* standard has garnered several other critiques, most notable of which are vagueness[240] and lack of consistency in application.[241]

Furthermore, all three cases in the *Daubert* trilogy concerned expert testimony from the "hard" sciences. Consequently, the guidelines derived from these cases are couched in terminology regarding the scientific method and objective scientific validity.[242] But economic models are not a hard science,[243] making these guidelines a poor fit for model-based testimony.[244]

The next section illuminates why model-based testimony should be treated differently from the "hard" sciences in litigation. This discussion will also inform why courts struggle to distinguish between issues of admissibility versus weight with regards to economics experts and their models.

2. Common Challenges to Economic Models in Litigation

The conversation surrounding econometric models suggests a highly rigorous and straightforward scientific inquiry.[245] However, though the methods themselves (regression analysis, for example) are mathematically sophisticated, their construction is not an exact science.[246] Models exhibit an "illusion of objectivity" that masks the integral role of the modeler's subjective judgments,[247] as pure theory and data alone cannot create a model.[248] The

---

239 Mnookin, *supra* note 14, at 1570-71.

240 Hersch & Bullock, *supra* note 26, at 2377.

241 Andrew I. Gavil, *Defining Reliable Forensic Economics in the Post-Daubert/Kumho Tire Era: Case Studies from Antitrust*, 57 WASH. & LEE L. REV. 831, 874 (2000); Todd & Jewell, *supra* note 40, at 311 (observing that "some courts pay only lip service to Rule 702 and the *Daubert* trilogy and focus their analysis on pre-Daubert authority rejecting expert testimony that is too speculative").

242 *See Daubert*, 509 U.S. at 590 (1993) ("The adjective 'scientific' implies a grounding in the methods and procedures of science. Similarly, the word 'knowledge' connotes more than subjective belief or unsupported speculation. . . in order to qualify as 'scientific knowledge,' an inference or assertion must be derived by the scientific method.").

243 Allensworth, *supra* note 9, at 829 (holding that "models are not scientific in the Popperian sense of being falsifiable; as inventions designed to perform a task, they are purposive rather than positive.").

244 Todd, *supra* note 28, at 252 (explaining that "the relative clarity of the *Daubert* factors for the hard sciences become muddied when applied to the non-scientific choices of the modeler").

245 Kaye & Freedman, *supra* note 82, at 272 (noting the general perception of statistical models as "marvel[s] of mathematical rigor").

246 Allensworth, *supra* note 9, at 841 (describing how "scientifically acceptable choices are neither unique. . . nor objective").

247 Hill et al., *supra* note 15, at 334.

248 Morrison & Morgan, *supra* note 72, at 15.

combination of statistical techniques with human discretion thus makes modelling both a science and an art.[249]

Perhaps the first choice a modeler makes is the methodology he wishes to use to achieve a given purpose, such as asset valuation.[250] Multiple regression analysis is particularly popular for showing causation due to the method's power in isolating the effects of independent variables on a target variable.[251] However, established methodologies such as these do not guarantee admissibility[252] and there are alternative methodologies that can achieve the same purpose.[253]

Though they are not explicitly phrased as assumptions, model construction decisions regarding type of methodology, data, and variables are tacit assumptions.[254] Choice of methodology implicitly assumes the appropriateness of that method for the task at hand,[255] whereas choice of data assumes the similarity between a firm or market in a past study and those involved in the present case.[256] These types of assumptions often only become apparent in their negation, that is, whenever opponents attack the modeler's choices in methodology, data, or variables. Judges' treatment of these challenges has historically been inconsistent: some courts have required that experts use a certain methodology (such as regression analysis)[257] while other courts have held that choice of methodology is an issue of weight.[258]

Another implicit assumption is a modeler's choice in variables, which presumes the relevance of those included and the irrelevance of those

249 Allensworth, *supra* note 9, at 829 (holding that the art of modeling exists where experts do not agree); Klock, *supra* note 22, at 198 (stating that "The art of good model-building lies in the ability to assume well"); Morrison & Morgan, *supra* note 72, at 31 (noting that good model-building, like an art or craft, involves "acquired skills in choosing parts and fitting them together").

250 *See* Allensworth, *supra* note 9, at 841; Hill et al., *supra* note 15, at 338-39 (enumerating the various and "frequently subjective" choices involved in valuation methodologies).

251 Hill et al., *supra* note 15, at 352.

252 Lopatka & Page, *supra* note 10, at 690.

253 Allensworth, *supra* note 9, at 841 (noting how modeling goals can be achieved through more than one method).

254 *See* Todd & Jewell, *supra* note 40, at 302, 318.

255 Gavil, *supra* note 241, at 876 (emphasizing the need for "fit" between a methodology and the facts of a case); Todd, *supra* note 12, at 990 (stating that "the type of math to perform and the applicable valuation method are themselves artistic choices").

256 Todd & Jewell, *supra* note 40, at 298 (describing the underlying assumption in yardstick approaches to damages that "but for the anticompetitive behavior, plaintiff's business would have performed like the comparator"); *id. at* 318 (noting a court's rejection of growth projections based on tacit comparisons between TV markets).

257 Lopatka & Page, *supra* note 10, at 689.

258 Hill et al., *supra* note 15, at 313 (citing *Popham v. Popham*, 607 S.E.2d. 575, 576 (Ga. 2005)).

excluded.[259] Again, courts do not approach these challenges with consistency.[260] In some cases, courts determined that omitted variable were issues of weight; in others, the omission of relevant variables caused the testimony to be excluded.[261]

In addition to choosing appropriate methodology, data, and variables, modelers make a series of explicit simplifying assumptions.[262] In litigation, these assumptions often involve similarity judgments between products and markets, as well as constructing alternate pasts where the defendant's misconduct did not occur.[263] Other types of simplifying assumptions are needed to distill patterns or abstract salient features from a "chaotic reality."[264] These assumptions typically involve tradeoffs between accuracy, simplicity, and usefulness.[265] These simplifying assumptions are those described by the assumptions typology, where certain factors are idealized or omitted for a second-order purpose that may be negligibility, applicability, or tractability.[266]

Due to their subjective nature, simplifying assumptions are the elements most likely to be scrutinized in *Daubert* motions.[267] Litigants often attack these simplifying assumptions as speculative or unrealistic, which prompts a highly ambiguous determination that appears to be guided primarily by "the predilections of the individual judge."[268]

Many simplifying assumptions cannot be falsified,[269] such as whether a product or market is similar enough to the ones in the underlying studies and data.[270] Speculating on what a plaintiff's position would have been but-for the defendant's wrongful conduct is also an exercise in reasoning rather than

---

259 Mäki, *supra* note 82, at 248; Todd & Jewell, *supra* note 40, at 302 (describing how unfounded omissions involve a "tacit assumption [] that some important and relevant factor is unimportant and irrelevant.").

260 Hill et al., *supra* note 15, at 353.

261 *Id.* (citing the court's holding in *Bazemore v. Friday*, 478 U.S. 385 (1986) that omitted variables go to weight, not admissibility); Lopatka & Page, *supra* note 10, at 691 n.479 (citing, *inter alia, Blue Dane Simmental Corp. v. Am. Simmental Ass'n*, 178 F.3d 1035, 1040–41 (8th Cir. 1999) (finding an expert's before-and-after model too "simplistic" to be admissible because it failed to account for other independent variables).

262 *See generally,* Mäki, *supra* note 36, at 328-29 (describing idealizing assumptions).

263 Todd & Jewell, *supra* note 40, at 298.

264 Allensworth, *supra* note 9, at 862; Klock, *supra* note 22, at 196 (stating that "the goal is to abstract the salient features of reality without becoming mired in minutiae").

265 Allensworth, *supra* note 9, at 832-833, 840; Klock, *supra* note 22, at 196 (noting that, like the construction of roadmaps, assumptions involve "aesthetically pleasing trade-off[s] between reality and abstraction").

266 *See supra* Part II.B. (Musgrave-Mäki-Hindriks typology).

267 Hill et al., *supra* note 15, at 331-32 n.253 (claiming that these assumptions are the most understandable for non-expert counsel).

268 Lloyd, *supra* note 31, at 408.

269 Allensworth, *supra* note 9, at 840.

270 Todd & Jewell, *supra* note 40, at 298-300 (observing that unreasonable comparisons are a major category of assumptions that opponents frequently challenge).

a verifiable fact.[271] These types of assumptions regarding the plaintiff's position in a counterfactual past also see mixed results from courts.[272] Justifying these assumptions requires subjective judgments regarding "similarity," "salience" and "credibility," which, by their nature, cannot be expressed in mathematical or logical terms.[273] The modeler should accordingly be able to articulate these justifications and defend his choice of assumptions.[274]

Aside from these simplifying assumptions, there are also statistical assumptions that are not subject to a modeler's "idiosyncrasies."[275] Regression, for example, assumes a linear relationship between the variables, as well as normally distributed and random error terms.[276] Todd and Jewell note the importance of this distinction between statistical assumptions and artistic assumptions made by the econometrician in the construction of his model.[277] For example, they observe the often overlooked distinction between omitted variable bias as a statistical issue and the omission of relevant variables as a fundamental flaw in model construction.[278] The former limits the precision of a model and may skew the coefficient estimates, but can often be resolved with other statistical procedures.[279] A violation of a statistical assumption is therefore "not likely to be fatal on its own."[280] The latter, however, is an artistic choice.[281] An appropriately omitted variable may serve the purposes of isolation, since models cannot capture the vast complexity of the world.[282] An inappropriate omission, however, can "easily invalidate any statistical

---

271 Blair & Page, *supra* note 183, at 435-36 (emphasizing that due to the "multitude of potential influences on business conditions, a plaintiff cannot prove what would have happened with the same degree of certainty that it can prove what did occur"); Hill et al., *supra* note 15, at 335 (citing *Gross v. Comm'r*, 272 F.3d 333, 356 (6th Cir. 2001)) (calling the damages estimation process a "fiction").

272 Todd, *supra* note 28, at 236 (comparing the conflicting rulings about whether the "likelihood and extent" of business expansion (*Polymer Dynamics, Inc. v. Bayer Corp.*, 2005 WL 1041197 (E.D. Pa)) or consumer behavior (*Nebraska Plastics, Inc. v. Holland Colors America, Inc.*, 408 F.3d 410, 416 (8th Cir. 2008)) are issues for the judge or jury to decide.).

273 Sugden, *supra* note 61, at 4.

274 Todd, *supra* note 28, at 239-40.

275 Allensworth, *supra* note 9, at 839-40.

276 Sykes, *supra* note 178, at 5-6.

277 Todd & Jewell, *supra* note 40, at 290.

278 *Id.* at 290-93.

279 David L. Rubinfeld, *Reference Guide on Multiple Regression, in* FED. JUDICIAL CTR., REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 303, 314, 322 (3d ed. 2011); Sykes, *supra* note 178, at 23-27.

280 Todd & Jewell, *supra* note 40, at 292.

281 Rubinfeld, *supra* note 279, at 281 (describing how omitted variables involve "assumptions made going into the analysis, rather than conclusions that come out of the data.").

282 Uskali Mäki, *MISSing the World. Models as Isolations and Credible Surrogate Systems*, 70 ERKENNTNIS 29, 30 (2009).

results," if a factor that the record suggests is relevant is excluded from the model.[283]

Todd and Jewell also note that while opponents purport to attack the purely scientific or statistical elements of a model, most arguments invoke the lack of foundation or reasoning an expert has for his modelling choices.[284] Essentially, opponents attack the artistic rather than the scientific elements of a model.[285] As a result, the application of the *Daubert* standard to these artistic choices is improper.[286] The inconsistency arising from this improper application of standards is consequential, as the outcome of a case is frequently contingent on the admission or exclusion of expert testimony.[287] Furthermore, large class actions and antitrust suits often involve eight- or nine-figure damages; estimates in *Dover v. British Airways* ranged between $143 and $161 million.[288]

3. Proposed Solutions in Legal Scholarship

Some legal commentators have attempted to illuminate this issue through various paradigms that center around the distinction between the judge's domain of law, and the jury's domain of facts. Allensworth states that models do not meet the criteria to be categorized as "facts," and thus advocates that judges—ideally those with some basic quantitative training—are best equipped to deal with model-based testimony.[289] Casey and Simon-Kerr stand on the opposite end of the spectrum, equating expert testimony with that of lay witnesses, the evaluation of which requires no more than "run-of-the-mill fact-finding."[290] Some legal scholars see model-based testimony as a mixed issue of law and fact. For example, Kaye distinguishes between "legislative" and "adjudicative" considerations.[291] The former involves facts that are external to the case such as the general acceptability of a certain methodology; these issues should be addressed by the judge.[292] "Adjudicative"

283 Todd & Jewell, *supra* note 40, at 292 (describing how poor choices in model construction cannot be fixed by statistical tools); *see id.* at 312 (stating that failure to include a clearly relevant variable can make the entire model irrelevant).

284 *Id.* at 316, 319; *see also* Hill et al., *supra* note 15, at 352-53 (observing that the inclusion or exclusion of variables in regression models is a frequent point of contention).

285 Todd & Jewell, *supra* note 40, at 293.

286 *See* Todd, *supra* note 28, at 252.

287 Mnookin, *supra* note 14, at 1569; Todd & Jewell, *supra* note 40, at 282-83.

288 321 F.R.D. 49, 57 (E.D.N.Y. 2017).

289 Allensworth, *supra* note 9, at 852.

290 Casey & Simon-Kerr, *supra* note 11, at 1182.

291 Kaye, *supra* note 31, at 1975.

292 *Id.* at 1983-85.

considerations such as failure to account for outliers, by contrast, are internal to the case and can be resolved by the jury.[293]

Other legal scholars, however, reject this law-fact debate entirely. Allen and Pardo assert that the law-fact distinction has no epistemological or ontological grounding, and that in order to protect the domains of judge and jury, pragmatic conventions should take the place of abstract dichotomies.[294] In order to be truly useful, these conventions should "import criteria and methods from the relevant scientific community" rather than pigeonhole disparate scientific disciplines into the existing legal framework.[295]

IV. A POTENTIAL SOLUTION BACKED BY ECONOMICS METHODOLOGY

Recognizing this need for practical conventions, Todd proposes a functional framework that is based in economics methodology.[296] This approach is in line with evidence law, since comments to FRE 702 instruct judges to look to the standards of the field in evaluating expert testimony.[297] The Musgrave-Mäki-Hindriks typology of assumptions describes how assumptions should be evaluated relative to their second-order purpose, since all models have false first-order assumptions.[298] Mäki's types of realisticness illuminate how assumptions that are unrealistic in one sense (*e.g.*, false) can still be valid if they are realistic in another sense (*e.g.*, plausible).[299]

Todd synthesizes and adapts these typologies for use in the context of litigation.[300] Perhaps most importantly, judges should evaluate assumptions according to their purpose and context, rather than in isolation.[301] The second-order assumption clarifies this purpose, which may be negligibility, tractability, or applicability.[302] It is crucial that modelers be specific about the second-order purpose of their assumptions,[303] as well as be willing to defend these choices in the event they are challenged.[304] Judges should focus their admissibility decisions on whether the second-order assumption is

293 *Id.* at 2012-13.

294 Allen & Pardo, *supra* note 31, at 1806-7.

295 Allensworth, *supra* note 9, at 829.

296 Todd, *supra* note 28, at 239, 282-90.

297 *Id.* at 237 (citing FED. R. EVID. 702 Committee Notes on Rules—2000 Amendment); *see also* Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 593-94 (1993) (listing general acceptance of principles or methodology by the relevant field as one factor judges may consider in their admissibility rulings).

298 Todd, *supra* note 28, at 282; *see supra* Part II.B.

299 Todd, *supra* note 28, at 284-285; *see supra* Part II.B.

300 Todd, *supra* note 28, at 282-91.

301 *Id.* at 290.

302 *Id.* at 282.

303 *Id.* at 287.

304 *Id.* at 282.

appropriately realistic, rather than exclude models with unrealistic first-order assumptions.[305] If the second-order assumption is materially disconfirmed by the evidence, the judge can exclude the testimony or require the expert to fix the assumption.[306] For example, the judge may exercise his gatekeeping function when an expert excludes a variable for negligibility purposes, but the record shows that the factor is actually non-negligible.[307] Lack of evidence, as opposed to disconfirmation by evidence, is not necessarily fatal to an assumption's validity.[308] After all, modelers must grapple with multiple "unknowns and unobservables" in constructing models that simplify a complex reality.[309] Assumptions therefore may be justified for tractability purposes, in order to handle this lack of evidence or data.[310]

If a modeler has no explanation for an assumption, the judge may exclude the testimony.[311] In cases where the modeler can both articulate the purpose for his assumptions, and the stated purpose is not disconfirmed by evidence, the judge can admit the model to the jury.[312] Jurors will then use their collective common sense and experience to assess realisticness as plausibility, as well as the coherence of the model with the rest of the evidence.[313]

Todd's synthesis between the two typologies arms judges with the vocabulary to approach a model's assumptions with a degree of clarity that has historically been lacking.[314] The mechanics of the framework also help sketch a reasonable boundary between admissibility and weight with a threshold that does not tend to either extreme.[315] The requirement that second-order assumptions be meaningfully connected to reality prevents indiscriminate acceptance of invalid models.[316] Otherwise, judges need not exclude model-based testimony where the only realisticness issue is the question of plausibility, which can be decided by a jury.[317]

Although the framework theoretically provides clarity to admissibility rulings, it is unclear whether the framework holds in practice. Todd samples contentious assumptions from several different cases but does not go into

305 *Id.* at 292.

306 Todd, *supra* note 28, at 292.

307 *Id.* at 286-87 (citing *in re* Live Concert Antitrust Litig., 863 F. Supp. 2d 966, 971-81) (an expert did not account for the effects that the advent of downloadable music might have on concert ticket prices. If the exclusion of this variable was a negligibility assumption, it would be invalid because evidence showed that downloadable music significantly affected ticket prices).

308 Todd, *supra* note 28, at 257 (arguing that scenarios projecting a counterfactual past (and thus lacking explicit evidentiary support) are issues of plausibility that the jury can evaluate).

309 *Id.* at 281.

310 *Id.* at 287.

311 *Id.* at 282.

312 *Id.*

313 *Id.* at 284.

314 *See* Todd, *supra* note 28, at 292.

315 *Id.* at 284-85.

316 *Id.* at 285.

317 *See id.* at 287, 292.

depth on any one case.[318] *Dover v. British Airways, PLC* presents an appropriate opportunity to test the framework's utility on an entire case. First, the case involves dueling experts, a dynamic that has not previously been explored by Todd in his application of the framework.[319] The presence of two experts who reached opposite conclusions regarding the correlation between fuel prices and YQ charges places increased pressure on the judge to confront alternate choices of methodology.[320] A battle of the experts also raises unique questions about the interaction between objective scientific requirements on which experts agree versus artistic choices that may be up for debate.[321] Moreover, Judge Dearie's expert ruling provides ample detail about the individual challenges to each expert's testimony;[322] this detail yields a more in-depth analysis than would otherwise be permitted. Finally, a high-profile defendant—and the concomitant massive damages estimates—helps illustrate the relevance and importance of these admissibility determinations. Judge Dearie's Expert Ruling details the challenges to expert testimony and explains the rationale behind his admissibility decisions; this document provides the basis for this analysis.

A. *Dover v. British Airways: Background*

British Airways had a frequent flyer program called the Executive Club, in which members could accumulate points ("Avios") by flying with British Airways, renting cars, and staying in certain hotels.[323] These points could then be redeemed for reward flights with British Airways.[324] The Executive Club contract explicitly granted the airline the right to impose fuel surcharges; in exercising this right, the airline's fuel surcharge committee used its cost of fuel in 2003-2004 as the baseline for these charges (internally referred to as "YQ charges").[325] Frequent flyer members alleged that this baseline was arbitrary and yielded YQ charges that were "not substantively or temporally relevant to the actual cost or price of fuel."[326] As a result, frequent fliers claimed that British Airways breached the Executive Club contract, and moved for class certification of all members who paid the YQ charges for Avios-redeemed flights between November 9, 2006 and April 17, 2013.[327] Among this class were four representative plaintiffs: Russell Dover, Suzette

318 *See* Todd, *supra* note 28, at 282-91.
319 *See* Dover, 254 F. Supp. 3d at 459-65.
320 *See infra* Part IV.B., C; *see also* Dover, 254 F. Supp. 3d at 459-65.
321 *See infra* Part V.
322 *See* Dover, 254 F. Supp. 3d at 459-65.
323 *See* Dover, 321 F.R.D. at 52.
324 *See id.*
325 *See id.*
326 *See id.* at 52-53.
327 *See id.* at 53.

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

Perry, Cody Rank, and Henry Horsey.[328] Each party enlisted testimony from economics experts to testify regarding the correlation between YQ charges and the cost of fuel.[329] Plaintiff's expert Jonathan Arnold was an economist from the Chicago Economics Corporation, and Andrew Hildreth was an econometrician retained by the defendant, British Airways.[330] Plaintiffs also provided testimony from Robert Kokonis, an expert in the airline industry who provided relevant input to the challenges against Arnold's damages models.[331]

B. *Plaintiff's Expert Testimony*

Plaintiff's expert Jonathan Arnold asserted in his reports that the level of YQ charges bore no "close relationship" with the price of fuel over time.[332] British Airways attacked this testimony as unreliable because it was "not based on a regression or some other . . . peer-review[] or published studies."[333] Instead, Arnold's methodology involved a "quarter-by-quarter comparison of . . . YQ charge[s] and fuel prices and a comparison of the relative growth of [the two figures]" across time.[334] Judge Dearie accepted Arnold's methodology as "reasonable and sufficiently reliable," and references three cases in succession as self-evident support for this decision:

> *See Zeraga Ave. Realty Corp. v. Hornbeck Offshore Transp., LLC*, 571 F.3d 206, 213-14 (2d Cir. 2009) ("[A] trial judge should exclude expert testimony if it is speculative or conjectural or based on assumptions that are 'so unrealistic and contradictory as to suggest bad faith' or to be in essence 'an apples and oranges comparison.'" (quoting *Boucher v. U.S. Suzuki Motor Corp.*, 73 F.3d 18, 21 (2d Cir. 1996)). Any arguable weakness in this methodology, or the possibility that relevant factors were omitted, goes to weight, not admissibility. *See id.* at 214 ("'[C]ontentions that the assumptions [of an expert witness] are unfounded go to the weight, not the admissibility, of the testimony.'" (alteration omitted) (quoting *Boucher*, 73 F.3d at 21)).[335]

The inherent contradiction of current admissibility precedent becomes clear here. Judges may pull in quotes ad hoc to support whichever conclusion they decide, since support for either outcome can be found in previous rulings. Referencing *Boucher*, Judge Dearie notes that any arguable weakness and any contention that assumptions are unfounded are issues of weight, not

328 *See id.*
329 *See* Dover, 254 F. Supp. 3d at 458.
330 *See id.* at 460, 463.
331 *See id.* at 459, 462.
332 *Id.* at 460.
333 *Id.* at 461.
334 *Id.*
335 Dover, 254 F. Supp. 3d at 461.

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

admissibility.[336] Such statements seem to advocate for a threshold that is so small as to be effectively non-existent. On the other hand, *Zeraga* describes a situation in which a judge may exercise his gatekeeping function, which involves instances where testimony is "speculative or conjectural."[337]

Ambiguity is also pervasive: it is unclear at what point "unfounded" assumptions become too "unrealistic" to be admissible.[338] Furthermore, essentially synonymous terms are used to justify opposite conclusions: assumptions that are "unfounded" are issues of weight, but those that are "conjectural" raise admissibility concerns.[339] A critical look at Judge Dearie's rationale thus demonstrates just how nebulous the basis for admissibility rulings can be.

While the framework does not explicitly provide guidelines on challenges toward choice of methodology, the methodological literature would likely support admission of the testimony. So long as Arnold can justify the appropriateness of his method, Arnold's testimony should be admitted so that his story may reach its intended audience, the jury.[340]

British Airways raised similar challenges in regard to Arnold's damages model. Party experts agreed that, assuming breach of contract, "the amount of damages to the class would equal the difference between the amount class members paid in YQ charges and the amount they would have paid had British Airways adopted an alternative, commercially reasonable course of conduct in compliance with the Contract."[341]

Arnold proposed two alternatives for how British Airways could have behaved in this fictional past: either the airline could have adjusted its fuel surcharge on a quarterly basis or it could have operated without a fuel surcharge at all.[342] British Airways criticized the models as presenting an "unrealistic ex post view" with assumptions that are "speculative."[343] Judge Dearie offered a brief response to these challenges, referencing the same quote as before:

> This objection alone is not grounds for excluding his testimony. *See Zeraga*, 571 F.3d at 214 (explaining that exclusion is warranted "where an expert's opinion is speculative or conjectural or based on assumptions that are 'so unrealistic and contradictory as to suggest bad faith[.]'" (quoting *Boucher*, 73 F.3d at 21)).[344]

---

336 *See id.*

337 *Id.*

338 *C.f., e.g.,* Boucher 73 F.3d at 2.

339 *Id.*

340 *See* Todd, *supra* note 12, at 1029 (concluding that if an economist's story adequately explains their artistic choices, the "gatekeeper should [] step aside so that the factfinder can hear the story.").

341 Dover, 254 F. Supp. 3d at 462.

342 *Id.*

343 *Id.*

344 *Id.*

Both the defendant and the court see an overly unrealistic assumption as grounds for exclusion. However, the judge's dismissal of the airline's challenges indicates that the two disagree regarding what is unrealistic, if it is binary, or, if it is a continuum, when an assumption passes the threshold into being too unrealistic. An exploration of Mäki's different types of realisticness may illuminate this issue.

If by "unrealistic ex post view" defendants mean unrealistic in the sense of implausible, then the testimony should be admitted because the jury can make credibility judgements based on their common sense and collective experiences.[345]

> British Airways also argues that these proposals are inadmissible because they lack foundation . . . . Kokonis, however offers a basis for Arnold's view . . . (stating, based on his experience in the airline industry, that 'it is not commercially unreasonable to manage fuel costs without a YQ charge.').[346]

One may presume the defendants mean lacking evidentiary foundation, and thus view realisticness in the sense of empirical confirmation.[347] A lack of evidence, however, is not equivalent to disconfirmation by evidence.[348] Because it is not disconfirmed, one may not say that Arnold's assumption is unrealistic in this sense.[349] To the contrary, evidence does exist to corroborate the statement in the form of Kokonis' expert testimony.[350]

Absent this supporting testimony, however, the lack of evidence would still be unproblematic because lack of data is the very reason why the assumptions were imposed in the first place.[351] In this case, Arnold's assumptions were imposed for the purposes of empirical tractability. One can infer this second-order purpose because the data needed to calculate damages are unobservable;[352] they exist only in a counterfactual past where British Airways complied with the contract.[353] To make the analysis tractable, therefore, Arnold must incorporate a hypothetical but-for condition in order to have a

---

345 *See supra* note 13, at 884; Vidmar, *supra* note 188, at 1137–38; Mnookin, *supra* note 14, at 1540-41; Griffin, *supra* note 190, at 294; Allensworth, *supra* note 9, at 848-49; Todd, *supra* note 28, at 312- 17 (issues of plausibility can and should be weighed by the jury).

346 Dover, 254 F. Supp. 3d at 462 (citing Kokonis May report ¶ 42).

347 *See* Maki, *supra* note 82, at 242 (describing Mäki's types of realisticness).

348 *See* Todd, *supra* note 28, at 257, 281, 287.

349 *Id.*

350 Dover, 254 F. Supp. 3d at 462.

351 *See* Todd, *supra* note 28, at 257, 281, 287 (modelers frequently need to incorporate factors for which they lack evidence).

352 *See* Hindriks, *supra* note 98, at 413–14; Hindriks, *supra* note 127, at 392, 399 (listing the data conditions that may justify empirical tractability assumptions).

353 *C.f.*, Blair & Page, *supra* note 183, at 435-36 (describing how plaintiffs must "construct and support a scenario of events in the but-for world").

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

basis for comparison with the plaintiff's current position.[354] He accomplishes this through the idealization (quarterly adjustment) and omission of YQ charges.[355] Thus, the formulation becomes:

> *First-order assumption*: Commercially reasonable alternatives include British Airways adjusting its fuel surcharges on a quarterly basis, or operating without fuel surcharges entirely.
>
> *Second-order assumption*: These assumptions were imposed for empirical tractability purposes because the data needed to calculate damages are unobservable.

The framework dictates that judges evaluate assumptions relative to their second-order purpose, rather than assess their realism at face-value.[356] The proper question for Judge Dearie to consider, therefore, is not whether the proposed alternatives are commercially reasonable, but whether the model was actually intractable without these first-order assumptions.[357] The answer is relatively clear in this case because, by definition, damages estimates hypothesize about what the plaintiff's position would have been but-for the defendant's wrongful conduct.[358]

Since there is no evidence that controverts the second-order tractability assumption, Arnold's model is admissible.[359] The jury can then assess realisticness as plausibility of the first-order assumptions, that is, decide whether quarterly adjustment or operating without fuel surcharges are reasonable alternatives given the need for tractability.[360] This decision will be informed by their experiences as business-owners and consumers, as well as their common sense and knowledge of how the world works.[361]

Judge Dearie ultimately made the proper decision in admitting the damages model, but his rationale conflated admissibility with weight. By citing Kokonis' testimony as support for Arnold's assumptions, Judge Dearie muddles the basis for the model's admissibility: the model is acceptable not because evidence supports the assumption, but because the question of the reasonability of a first-order assumption is for the jury to answer, not the

---

354 *See* Allensworth, *supra* note 9, at 837; Blair & Page, *supra* note 183, at 435-36; Lopatka, *supra* note 10, at 687; Hill et al., *supra* note 15, at 335 (experts necessarily craft hypothetical scenarios in damages estimations).

355 *See* Dover, 254 F. Supp. 3d at 462.

356 *See* Todd, *supra* note 28, at 257, 282-92 (the judge should limit his analysis to whether the second-order assumption is disconfirmed by evidence).

357 *Id.*

358 *See* Allensworth, *supra* note 9, at 837; Blair & Page, *supra* note 183, at 435-36; Lopatka, *supra* note 10, at 687; Hill et al., *supra* note 15, at 335.

359 *See* Todd, *supra* note 28, at 282, 284.

360 *Id.*

361 *See* Vidmar, *supra* note 188, at 1137–38; Mnookin, *supra* note 14, at 1540-41; Griffin, *supra* note 190, at 294; Allensworth, *supra* note 9, at 848-49; Faigman et al., *supra* note 13, at 884; Todd, *supra* note 12, at 1034.

judge.[362] Supporting evidence in the form of Kokonis' testimony may strengthen the plausibility of the model in the juror's eyes, but evidentiary support for first-order assumptions is not a precondition to admissibility.[363]

Fortunately, the recognition of this distinction becomes clearer as Judge Dearie continues. Judge Dearie cites the *Daubert* case in recognizing that the "presentation of contrary evidence" in conjunction with "vigorous cross-examination" are generally sufficient means for the jury to evaluate models critically.[364] British Airways will likely raise the same arguments to the jury during cross-examination and Arnold will be forced to defend his choice of assumptions.[365]

C. Defendant's Expert Testimony

Defendant British Airways enlisted the testimony of econometrician Andrew Hildreth. The expert proffered several regression analyses that supposedly showed a "high degree of correlation (over 70 percent)" between the airline's YQ charges and the price of fuel over time.[366] The results of his model would suggest that British Airways had in fact complied with the contract, since their surcharges were reasonably related to fuel prices. However, plaintiffs challenged the reliability of Hildreth's testimony, alleging that his failure to account for the non-stationarity of the underlying data rendered the model's results "spurious."[367] Non-stationarity is a property of some time-series data, in which statistical parameters such as mean and variance change over time.[368] "Colloquially, non-stationary data is said to exhibit a 'random walk,' such that knowing its value today tells us little or nothing about its value tomorrow."[369] Experts from both sides agreed that performing regression analysis on non-stationary data can yield spurious results, such as showing false correlation between unrelated variables.[370]

In response to these critiques, however, Hildreth claimed that non-stationary data may still be valid for use in a regression, so long as the data are cointegrated.[371] Cointegration refers to two variables in a time-series (YQ charges and fuel prices, for example) that exhibit a long-run

---

362 *See* Todd, *supra* note 28, at 257, 281-92 (judges should focus their analyses on the realisticness of second-order assumptions, not first-order assumptions).

363 *See* Hindriks, *supra* note 98, at 413–14; Hindriks, *supra* note 127, at 392, 399; Todd, *supra* note 28, at 257, 281, 287 (one purpose of tractability assumptions is to account for variables that lack data).

364 Dover, 254 F. Supp. at 462 (citing Daubert, 509 U.S. at 596).

365 *See* Hill et al., *supra* note 15, at 342; Todd & Jewell, *supra* note 40, at 313.

366 Dover, 254 F. Supp. 3d at 463.

367 *Id*.

368 *Id.* at 464 (citing reports from both Arnold and Hildreth).

369 *Id.*

370 *Id.* at 463-64.

371 *Id.* at 464.

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

equilibrium.[372] Hildreth claimed to have tested the data for cointegration using the Engle-Granger method.[373] However, plaintiffs presented evidence of the contrary: Engle himself rejected the expert's finding that the variables were cointegrated, even stating that Hildreth used another method entirely.[374] As a result, Engle concurred with Arnold that the results of Hildreth's regression are "statistically meaningless."[375]

While these challenges concern alleged statistical errors rather than competing notions of realisticness, the framework may still be useful for analysis. The element of Hildreth's model under scrutiny may be phrased as an applicability assumption:

> *First order assumption*: Regression analysis is a valid method for demonstrating the correlation between fuel prices and YQ charges.
>
> *Second order assumption*: *If* the data are cointegrated, *then* regression analysis of non-stationary data (such as fuel prices and YQ charges) yields valid results.

According to the framework's specifications, Judge Dearie was proper in evaluating the realism of the second-order assumption rather than that of the first-order assumption.[376] He did not call into question whether the choice of method (regression analysis) was acceptable in a general sense, but instead considered how evidence controverted the *applicability* of that method. Plaintiffs demonstrated that the domain condition was not satisfied, i.e., the data were not cointegrated. Thus, the second-order assumption is disconfirmed by evidence, which is grounds for exclusion because the use of regression analysis outside its applicable domain renders the model invalid.[377]

Judge Dearie declined to make an admissibility decision based solely on the parties' submissions and called for an evidentiary hearing to better evaluate the validity of Hildreth's testimony.[378] However, the defense withdrew the expert before an evidentiary hearing could be held.[379] The case subsequently moved to settlement, with class members collectively receiving up to $63 million in compensatory damages.[380] This result reinforces how the exclusion of expert testimony is outcome determinative: without a model that shows correlation between fuel prices and YQ charges, British Airways

---

372 Dover, 254 F. Supp. 3d at 464.

373 *Id.*

374 *Id*. at 464-65.

375 *Id.* at 465.

376 *See* Todd, *supra* note 28, at 282-92.

377 *See* Todd, *supra* note 28, at 257, 282-92.

378 Dover, 254 F. Supp. 3d at 465 (citing Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993) and Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999)).

379 Dover v. British Airways, 2017 WL 4358726, at *2 (E.D.N.Y. Sept. 29, 2017).

380 Dover v. British Airways, 323 F. Supp. 3d 338, 346 (E.D.N.Y. 2018).

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

lacked a critical thread in their narrative of valid fuel surcharges.[381] This, in conjunction with the admission of plaintiff expert testimony, would have led to a loss at trial, which potentially explains why the airline opted for a settlement.[382]

V. FRAMEWORK ASSESSMENT AND DISCUSSION

The theoretical framework certainly provides greater analytical depth to the evaluation of plaintiff expert Arnold's testimony. By approaching the assumptions of quarterly adjustment and zero YQ charges at face-value, Judge Dearie risked making a credibility determination that usurps the jury's role. The framework prevents this outcome by distinguishing between first-and second-order assumptions, as well as describing the various ways these assumptions may be realistic or unrealistic. An understanding of how these two typologies intersect would have allowed Judge Dearie to navigate his admissibility decision with consistency and clarity. Rather than evaluate whether the first-order assumption is realistic as plausible, Judge Dearie could have limited his analysis to whether the second-order purpose was unrealistic as disconfirmed by evidence.

Furthermore, the application of Mäki's typology of realisticness clarifies the previously hazy conceptions of what is "realistic" versus "unrealistic." If defendants had been required to articulate exactly what kind of realisticness Arnold failed to meet, the judge could have made a less ambiguous determination. If British Airways had attacked the assumptions about YQ charges as unrealistic in the sense of implausible, these challenges would clearly be issues for the jury, whose job is to discern credibility. Unrealistic in the sense of confirmation and disconfirmation is a stronger claim, but lack of evidentiary support for a first-order assumption is still not grounds for exclusion because all models involve unrealistic first-order assumptions to simplify the real-world.[383] The presentation of contrary evidence regarding the second-order tractability assumption, however, may warrant exclusion.[384] If the contract had stipulated a precise method British Airways must use to set fuel surcharges—perhaps automatically calculated and directly related to the cost of fuel—the data concerning plaintiffs' but-for position might be

381 *See* Mnookin, *supra* note 14, at 1569,1577; Hill et al., *supra* note 15, at 317-18, 334 (2009); Blair & Herndon, *supra* note 17, at 802; Todd, *supra* note 12, at 996-97.

382 *See* Mnookin, *supra* note 14, at 1569,1577; Hill et al., *supra* note 15, at 317-18, 334 (2009); Blair & Herndon, *supra* note 17, at 802; Todd, *supra* note 12, at 996-97 (describing how exclusion of one party's expert can increase the relative weight of the opponent's testimony).

383 *See* Klock, *supra* note 22, at 196.

384 Hindriks, *supra* note 98, at 410 (describing how the first-order assumptions are often false or unrealistic, but second-order assumptions must be realistic).

observable and the empirical tractability assumption would no longer be valid.[385] Plaintiff's expert Arnold could have looked at past fuel prices and applied the method mentioned in the contract to calculate what plaintiffs should have paid but-for the breach of contract. The availability of data in this hypothetical scenario would have invalidated the empirical tractability assumptions since the damages model would actually be tractable without them. In reality, the contract did not specify how fuel surcharges were to be calculated, so Arnold had to impose tractability assumptions regarding how British Airways should have acted in the counterfactual past.

Application to defendant's expert Hildreth was not as straightforward but highlights an important distinction. Judge Dearie's decision to hold an evidentiary hearing for Hildreth's regression was due to a perceived fundamental difference between the challenges to each expert's testimony. "Unlike the parties' arguments respecting the other experts, the possibility that Hildreth's analysis may be spurious and statistically meaningless goes to the heart of the Supreme Court's concerns in *Daubert* and *Kumho Tire*."[386] This comment indicates Judge Dearie's subtle awareness that economic modeling is both science and art.[387] The recognition that Arnold's simplifying assumptions were subjective and artistic choices was fairly unambiguous. Recall that constructing a "but-for" world is an exercise in reasoning rather than a factual investigation.[388] The science of statistics does not prescribe how Arnold may simplify reality or project a counterfactual past.[389] Further, the use of natural language makes the jury capable of assessing the plausibility of the expert's propositions.[390] Aided by their collective common sense and experience,

385 *See* Hindriks, *supra* note 98, at 413–14; Hindriks, *supra* note 127, at 392, 399; *see also* Todd, *supra* note 28, at 297 (citing Neb. Plastics, Inc. v. Holland Colors Americas, Inc., 408 F.3d 410, 416, 416 n.2 (8th Cir. 2005)) (describing a case where an expert imposed an assumption under the guise of tractability, namely that all siding panels would fade and thus be subject to warranty claims. However, historical data for this factor existed that the expert could have used, thus disconfirming the tractability assumption).

386 *Id.*

387 *See supra* Part III.C.

388 *See* Allensworth, *supra* note 9, at 840; Todd & Jewell, *supra* note 40, at 298-300; Blair & Page, *supra* note 183, at 435-36; Hill et al., *supra* note 15, at 335; Todd, *supra* note 28, at 236; Sugden, *supra* note 61, at 4.

389 *See* Kaye & Freedman, *supra* note 82, at 272; Allensworth, *supra* note 9, at 829, 841; Hill et al., *supra* note 15, at 334; Morrison & Morgan, *supra* note 72, at 15, 31; Klock, *supra* note 22, at 198 (stating that "The art of good model-building lies in the ability to assume well"); *See* Allensworth, *supra* note 9, at 840; Todd & Jewell, *supra* note 40, at 298-300; Blair & Page, *supra* note 183, at 435-36; Hill et al., *supra* note 15, at 335; Todd, *supra* note 28, at 236, 239-40; Sugden, *supra* note 61, at 4.

390 *See* Todd, *supra* note 12, at 1018-22, 1034; Mäki, *supra* note 36, at 330-31; Morgan, *supra* note 78, at 361, 366; Gilboa et al., *supra* note 80, at F518; Lopatka & Page, *supra* note 10, at 687; Casey & Simon-Kerr, *supra* note 11, at 1178; Allensworth, *supra* note 9, at 837; Blair & Page, *supra* note 183, at 435-36 (experts use stories to communicate a model and its assumptions to a lay audience).

File: Faulkner 17.1 To Publish_Edited Created on: 6/7/2022 5:18:00 PM Last Printed: 6/14/2022 9:27:00 PM

jurors can decide whether Arnold's model depicts a world that *could be* real.[391] The expert, of course, will have to convince his audience of the similarity between his model and reality and persuade them to make the "inductive leap" required to accept his conclusions.[392] Though Judge Dearie did not articulate this distinction in these terms, his ultimate decision suggests an understanding that the plausibility of artistic assumptions are issues of weight for the jury.[393]

By contrast, challenges to Hildreth's testimony prompted a separate kind of inquiry. Judge Dearie reasoned that the technical nature of these challenges – as opposed to challenges based on logic and reasoning – warranted a more in-depth treatment.[394] The question becomes whether this assessment was correct, and consequently, to what extent the theoretical framework can apply.

Recall Todd & Jewell's distinction between omitted variable bias as a statistical issue versus omitted variables as flaws in model construction.[395] The former involves a violation of a linear regression assumption that the error term is random.[396] This is because the influence of an independent variable on a dependent variable becomes reflected in the error term, making these errors systematic rather than random.[397] Similarly, omitted variable bias results in overestimated coefficient estimates.[398] Though these problems are "scientific" in nature, they do not necessarily ruin a model's utility, and may only decrease the model's goodness of fit.[399] The latter of conception of omitted variables, by contrast, constitutes an artistic choice that may be a "fatal flaw[] in the model's construction."[400] Omitting a variable that the record suggests is relevant can make the entire model irrelevant for its purpose.[401]

It is not immediately clear under which conception Hildreth's testimony should be placed. One on hand, plaintiff's challenges may concern purely statistical issues, akin to the former conception of omitted variable bias that often can be fixed and may only weaken precision. Alternatively, Hildreth

391 *See supra* Part III.B.; *see also* Sugden, *supra* note 59, at 25, 28 (describing models as credible counterfactual worlds that could conceivably be true, given a general understanding of how the world works).

392 *See* Sugden, *supra* note 59, at 20; Mäki, *supra* note 36, at 330-31; Morgan, *supra* note 78, at 361, 366; Itzhak Gilboa et al., *supra* note 80, at F518.

393 *See supra* Part IV.A.

394 *Id.* (noting how evidentiary hearings are "highly desirable" and "commonly held in cases like this one that involve expert testimony on complex scientific or economic topics.").

395 Todd & Jewell, *supra* note 40, at 290-93.

396 Sykes, *supra* note 178, at 23-24.

397 Sykes, *supra* note 178, at 24-27; *see also* Todd & Jewell, *supra* note 40, at 291.

398 Sykes, *supra* note 178, at 25-27.

399 *Id.;* Todd & Jewell, *supra* note 40, at 290-92.

400 Todd & Jewell, *supra* note 40, at 291.

401 *Id.* at 312.

may have made poor choices in model construction that rendered the entire model invalid.

Plaintiff's challenges to the expert's model concerned his choice of methodology in relation to the type of data he used. Those challenges hinged on whether the data exhibited certain properties (stationarity at first, then cointegration) that would allow a regression analysis to yield valid results.[402] The technical complexity of these challenges may suggest that Hildreth's model exhibits statistical and thus purely "scientific" flaws, those which *Daubert* aims to prevent and which are distinct from the framework's intended use.[403] Since the alleged flaws concerned certain data properties, Hildreth's assumptions perhaps should be treated more like statistical assumptions, which often concern data properties such as linearity and normally distributed error terms.[404] Sometimes, the violation of these statistical assumptions may be fairly innocuous. After all, "owing to the nature of economic relationships and the lack of controlled experimentation, these (statistical) assumptions are seldom met."[405] One can still accept the general conclusions of the model to help answer, although with less confidence than if all statistical assumptions had been met.[406] Thus, the model may still be useful in helping the trier of fact answer a certain question.

In Hildreth's case, however, the alleged error was material to the purposes for which the model was designed. Regression analysis applied to non-stationary and non-cointegrated data may yield "spurious" results such as false correlation.[407] When the purpose of the model is to establish correlation between fuel surcharges and the price of jet fuel, and false correlation is a potential consequence of the statistical error, the model fails at achieving its purpose. In this case, the judge would be proper in exercising his gatekeeping function, either by requiring that Hildreth fix the issue or excluding his testimony altogether.

Alternatively, Hildreth's errors may be construed as poor artistic choices to which the framework can be applied. Recall that model construction decisions regarding type of methodology, data, and variables are tacit assumptions.[408] This alternative framing would more clearly fit within the assumptions' typology.

---

402 *See supra* Part IV.C.

403 *See* Todd, *supra* note 28, at 249-52 (noting how statistical assumptions can be tested empirically for validity but artistic assumptions cannot, thus creating confusion among courts).

404 Allensworth, *supra* note 9, at 844.

405 PETER KENNEDY, *A GUIDE TO ECONOMETRICS*, 1 (6th ed. 2008).

406 Rubinfeld, *supra* note 279, at 322.

407 Clive Granger & Peter Newbold, *Spurious Regressions in Econometrics*, 2 J. ECONOMETRICS 111, 112–14 (1974) (explaining the phenomenon of "spurious" regressions in time series data, that is regressions that may show a false correlation between unrelated variables).

408 *See* Todd & Jewell, *supra* note 40, at 298, 302, 318; Gavil, *supra* note 241, at 876; Todd, *supra* note 12, at 990; Lopatka & Page, *supra* note 10, at 689, 691 n.479; Hill et al., *supra* note 15, at 313, 353; Mäki, *supra* note 82, at 248.

Hildreth's use of regression analysis involved an implicit assumption regarding the appropriateness of that method for showing correlation between fuel surcharges and the cost of jet fuel. The second-order assumption restricted the application of that method to instances where the nonstationary data are cointegrated. Plaintiffs alleged that his combination of regression and non-stationary data could yield spurious results that were "meaningless as a matter of statistics."[409] Lacking the technical knowledge to make a decision based on the parties' submissions alone, Judge Dearie called for an evidentiary hearing. This decision stemmed from a fear about the jury's incapacity to parse the statistical merits of Hildreth's model.[410]

Had the theoretical framework been applied, the evidentiary hearing may not have been necessary. If Hildreth had been unable to defend his choice of methodology, this lack of a "story" would be grounds for exclusion.[411] While experts have discretion in the construction of their models, their choices must be grounded by a second-order purpose that connects their model to reality in a meaningful way.[412] Hildreth initially justified his choices through an applicability assumption that would have made regression analysis a valid method.[413] The presentation of evidence showing that Hildreth violated the domain of applicability means that the second-order assumption is unrealistic in the sense of disconfirmation, making his model invalid and excludable.[414]

Therefore, without having to know whether regression analysis can actually yield valid results with non-stationary and non-cointegrated data, the judge could have looked to how Hildreth's applicability assumption was supported or contradicted by the record and made his decision on this basis. Approaching Hildreth's issues as poor choices in model construction – as opposed to strictly scientific flaws – permits an evaluation of expert testimony that does not require perfect knowledge of statistical requirements and conventions. This type of approach comprises the most frequent basis for exclusion of expert testimony.[415] Rather than attack purely statistical problems, opponents and courts have instead seized upon the lack of justification or support for the expert's artistic choices.[416]

---

409 Dover, 254 F. Supp. 3d at 463.

410 *Id*. at 465 (considering whether Hildreth's testimony is the very kind of "unsound science" that should trigger the judge's gatekeeping role).

411 Todd, *supra* note 28, at 285-86 (stating that an economist's story necessarily involves explanations for their assumptions, therefore "courts need only admit expert testimony if the modeler has a story for the jury.").

412 *Id.* at 270; *see supra* Part II.B.

413 *See supra* Part IV.C.

414 *See supra* Part IV.C.

415 Todd & Jewell, *supra* note 40, at 319 (noting that "though courts cite to problems arising from the violation of regression assumptions like omitted variable bias, those courts excluded expert testimony not because of statistical problems but because of poor choices in model construction.").

416 Todd & Jewell, *supra* note 40, at 314-15.

The framework thus aligns with how courts have historically approached challenges to expert testimony. Though Todd and Jewell only make claim to a positive rather than normative analysis,[417] it is a strength of the framework that it would not fundamentally change how courts address contentious issues. After all, the goal is to give judges (a "lay audience" with regards to econometric principles) a set of guidelines that do not require them to be experts in statistics themselves. Though the framework's appropriateness for Hildreth's testimony is initially ambiguous, a closer look reveals that it can help resolve challenges that straddle the scientific and artistic aspects of model-building. Therefore, the framework affords judges the vocabulary needed to address a variety of complex issues from a standpoint that is familiar to litigation: justification for assumptions (specifying second-order purpose), argumentation in defending these choices (both implicit and explicit), and an eye to how the record supports or controverts these arguments.[418]

CONCLUDING THOUGHTS

While the framework has wide applicability, it could benefit from additional guidelines that help courts distinguish between attacks to statistical issues versus attacks to choices in model construction. An awareness of the dual-identity of models as both art and science would be the first step to making this distinction. Increased specificity from opponents regarding the precise objects of their attack (*e.g.*, omitted variable bias vs. omission of relevant variables) could also prevent future conflation of the two types of issues. Finally, understanding that models involve many tacit assumptions—such as choice of methodology, data, and variables—can expand the framework's applicability to more general issues. Application to Arnold's damages models confirmed the framework's utility for evaluating internal and explicit assumptions whose realisticness is at issue. Hildreth's case demonstrated that the framework is also useful when applied to choices that are external to a model's construction, such as type of methodology and variables. Once these tacit assumptions are phrased explicitly as a pair of first-and second-order assumptions, the judge may evaluate the second-order assumption in light of the available evidence. Therefore, the theoretical framework—in conjunction with a firm grasp on modeling as a science and art—can provide more precision and consistency to how courts currently approach admissibility rulings on economic models.

417 *Id.* at 319.
418 *See id.* at 283.

File: Brown 17.1 To Publish_Edited Created on: 4/19/2022 10:18:00 AM Last Printed: 6/14/2022 9:27:00 PM

A BLIND EYE: HOW THE RATIONAL BASIS TEST INCENTIVIZES REGULATORY CAPTURE IN OCCUPATIONAL LICENSING

Jack Brown

INTRODUCTION

Occupational licensing, which requires workers to obtain credentials demonstrating training and education and pay fees in order to work in their chosen fields, is often justified by its proponents on a health and safety rationale.[1] However, licensing is frequently imposed by legislatures to shield those already engaged in an occupation from competition by preventing newcomers from entering the field.[2]

The state of Oklahoma is a prime example. Oklahoma requires individuals to obtain a funeral director's license in order to sell caskets and other funeral-related paraphernalia.[3] This requirement was in place despite caskets being mere empty boxes in which to store bodies, with little risk to public health or safety involved in their sale to justify licensure. Instead, the requirement was created to freeze competitors out of the market. Licensed funeral directors recognized that most of the money to be made in their industry involved selling funeral-related merchandise, including caskets. For this reason, they lobbied legislatures, including that of Oklahoma, to give them a monopoly over casket sales by requiring individuals wishing to sell caskets to have a funeral director's license.[4] In response, Kim Powers and Dennis Bridges, two individuals who had started a business to sell caskets, filed a federal suit in 2002 to protect their right to engage in the profession of their choice in Oklahoma and compete with the licensed funeral directors.[5] The Western District of Oklahoma held that the licensing scheme was constitutional, despite its protectionist nature. The court held that under the rational

---

[1] Zach Herman, *The National Occupational Licensing Database: Executive Summary*, NAT'L CONF. OF ST. LEGISLATURES (Feb. 7, 2020), https://www.ncsl.org/research/labor-and-employment/occupational-licensing-statute-database.aspx.

[2] *See* Edward Rodrigue and Richard B. Reeves, *Four ways occupational licensing damages social mobility*, BROOKINGS INST. (Feb. 24, 2016), https://www.brookings.edu/blog/social-mobility-memos/2016/02/24/four-ways-occupational-licensing-damages-social-mobility/.

[3] OKLA. STAT. tit. 59, §§ 396.3a, 396.6(A) (2019).

[4] Dick M. Carpenter II, et al., *License to Work: A National Study of Burdens from Occupational Licensing*, INST. FOR JUSTICE 31 (2d ed. 2017), https://ij.org/wp-content/themes/ijorg/images/ltw2/License_to_Work_2nd_Edition.pdf.

[5] Jim Stafford, *Attempt to change law falls short*, OKLAHOMAN (Dec. 16, 2005), https://oklahoman.com/article/2924010/attempt-to-change-law-falls-short?.

basis test, the standard of review under which courts analyze the constitutionality of economic regulations, the actual facts underlying the legislature's decision were irrelevant, and all that mattered was that there was a reasonably conceivable state of facts to sustain a rational basis for the classification, even if it was based on speculation. The court held that the purpose offered by the government, that the restriction existed for consumer protection, was legitimate.[6] On appeal, the United States Court of Appeals for the Tenth Circuit upheld constitutionality of the licensing scheme as well, and held that intrastate economic protectionism, such as the protection of the established funeral directors from competition, was a legitimate state interest.[7]

Although it might seem unusual for the active shielding of an established political interest group from competition by those who wish to work in an industry to be a legitimate government interest, it is a state of affairs that has arisen as a result of the courts applying an extremely deferential standard of review to economic regulations, such as those governing occupational licensing. The United States Supreme Court has recognized that the freedom to work in the occupation of one's choice without undue government interference is a right.[8] In fact, it sometimes refers to occupational freedom as a "fundamental" right.[9] Occupational freedom also has a long history of support as a right in the United States.[10] Despite this, the Court reviews the constitutionality of restrictions on occupational freedom under the extremely deferential rational basis test, also called rational basis review.[11] Under rational basis review, a statute is constitutional if it is rationally related to a legitimate government interest.[12] The test presumes that the government is acting constitutionally and often requires judges to accept and invent explanations for the government activity at issue in the case, even if the explanation is purely hypothetical.[13] Because of the deferential nature of the test, critics have argued that it amounts to little more than a rubber stamp that "stack[s] the deck in favor of lawmakers."[14]

---

6 Powers v. Harris, Case No. CIV-01-445-F, 2002 U.S. Dist. LEXIS 26939, at *24-35 (W.D. Okla. Dec. 12, 2002).

7 Powers v. Harris, 379 F.3d 1208, 1217-22 (10th Cir. 2004).

8 Connecticut v. Gabbert, 526 U.S. 286, 291–92 (1999); Bd. of Regents v. Roth, 408 U.S. 564, 572 (1972); New State Ice Co. v. Liebmann, 285 U.S. 262, 278 (1932); Meyer v. Nebraska, 262 U.S. 390, 399–400 (1923).

9 Supreme Court of N.H. v. Piper, 470 U.S. 274, 280 n.9, 285 (1985).

10 TIMOTHY SANDEFUR, THE RIGHT TO EARN A LIVING 18-25 (2010).

11 Williamson v. Lee Optical of Okla., Inc., 348 U.S. 483, 487-91 (1955); United States v. Carolene Prods. Co., 304 U.S. 144, 152-54 (1938).

12 Williamson, 348 U.S. at 487-91; Carolene Prods., 304 U.S. at 146-54.

13 CLARK NEILY, TERMS OF ENGAGEMENT: HOW OUR COURTS SHOULD ENFORCE THE CONSTITUTION'S PROMISE OF LIMITED GOVERNMENT 50 (2013).

14 *See, e.g.*, NEILY, *supra* note 13, at 49-63; DAMON ROOT, OVERRULED: THE LONG WAR FOR CONTROL OF THE U.S. SUPREME COURT 135-36 (2014).

Rational basis review has come under fire on a number of grounds. Critics have argued that it violates due process by encouraging or mandating judicial bias in favor of one of the parties, that it treats facts as irrelevant and relies on hypotheticals with little basis in reality, that it creates a burden of proof requiring that plaintiffs negate every conceivable justification for the regulation at issue, that it is nearly impossible to meet as a matter of formal logic, and that it is applied inconsistently by courts to achieve desired outcomes.[15] Another major criticism leveled at the rational basis test is that it creates an environment in which regulatory capture can thrive.[16] Regulatory capture occurs when a political entity or regulatory body is acquired by the industry being regulated and the body or entity acts primarily for the benefit of that industry.[17] Because rational basis review is so deferential to the government, it leaves courts unable to serve as effective checks against politically connected interest groups that have captured governmental bodies and then use them to stop their potential competitors from operating in the marketplace. In fact, this problem with the test is so pervasive that it has been recognized by judges and scholars alike.[18]

This article will examine how the rational basis test enables regulatory capture. The test is far too deferential to the government to serve as an effective check against attempts to undermine the right of occupational freedom in order to protect politically favored actors from competition. Instead, in light of the strong foundation for the right to occupational freedom in the history and traditions of the United States, this article will argue that restrictions on occupational liberty should be reviewed under strict scrutiny. Strict scrutiny is a much more rigorous standard of review that creates much less risk of regulatory capture.[19] Failing that, review under intermediate scrutiny, a standard of review that is more rigorous than the rational basis test but less rigorous than strict scrutiny, would also help to mitigate the capture problem.[20] This paper will also examine potential legislative solutions to the problem of regulatory capture in occupational licensing that apply more thorough examination of licensing restrictions during the legislative process.

Part I of this article will examine the history and background of regulatory capture, occupational freedom as a right, and the rational basis test. Part II will examine the use of the rational basis test to promote regulatory capture in occupational licensing through its deferential nature, even when the laws in question are blatantly protectionist. Part III will argue that in light of the long history and tradition of the right to occupational freedom, occupational

---

15 Clark Neily, *No Such Thing: Litigating under the Rational Basis Test*, 1 N.Y.U. J. L. & LIBERTY 897, 902-12 (2005); Andrew Ward, *The Rational Basis Test Violates Due Process*, 8 N.Y.U. J. L. & LIBERTY 714, 715-35 (2014).

16 *See*, *e.g.*, Hettinga v. United States, 677 F.3d 471, 482-83 (D.C. Cir. 2012); NEILY, *supra* note 13, at 56-63.

17 George Stigler, *The Theory of Economic Regulation*, 2 BELL J. ECON. & MGMT. SCI. 3 (1971).

18 Hettinga, 677 F.3d at 482-83; NEILY, *supra* note 13, at 56-63; ROOT, *supra* note 14, at 135-36.

19 *See* NEILY, *supra* note 13, at 49.

20 *See* NEILY, *supra* note 13, at 49-50.

licensing restrictions should receive strict scrutiny. Finally, Part IV will examine intermediate scrutiny and legislative solutions to the problem. Although strict scrutiny would be preferable, both represent ways to avoid the problem of undue deference to protectionist laws motivated by regulatory capture.

I. BACKGROUND

Regulatory capture, occupational freedom as a right, and the rational basis test all have long histories. Regulatory capture, an economic theory developed in the 1970s, has examples that are centuries old. Occupational freedom has a history dating back to 17th century England. Finally, the origins of rational basis review date to several decades before the New Deal.

A. *Regulatory Capture*

As mentioned in the Introduction of this article, regulatory capture occurs when a political or regulatory body is acquired, or "captured," by the industry which the body is intended to regulate. Once captured, the body then acts for the benefit of that industry.[21] Capture theory was developed by economist George Stigler in the 1970s during his time at the University of Chicago, for which he was eventually awarded the Nobel Prize in Economics in 1982.[22] Regulatory capture has been cited as a part of public choice theory.[23] Public choice theory involves the application of economics to the analysis of political behavior, and holds that rather than being caring public servants motivated by a desire to promote the common good, political actors are human beings who are motivated by self-interest.[24] James Buchanan, the public choice economist and winner of the Nobel Prize in Economics in 1986, aptly described the idea as "politics without romance."[25]

Although regulatory capture theory arose from Stigler's work in the 1970s, examples of the theory in action date back centuries before Stigler was born. This is unsurprising, as the power to use one's political connections for personal gain is a powerful temptation. The Greeks and the Romans,

---

21 Stigler, *supra* note 17, at 3.

22 David R. Henderson, ed., *George J. Stigler*, CONCISE ENCYCLOPEDIA OF ECON. (2d ed. 2008), https://www.econlib.org/library/Enc/bios/Stigler.html.

23 *See*, *e.g.*, Daniel A. Crane, *Tesla, Dealer Franchise Laws, and the Politics of Crony Capitalism*, 101 IOWA L. REV. 573, 574-76 (2016).

24 William F. Shugart, *Public Choice*, CONCISE ENCYCLOPEDIA OF ECON. (2d ed. 2008), https://www.econlib.org/library/Enc/PublicChoice.html.

25 James Buchanan, *Public Choice: Politics without Romance*, POL'Y, Spring 2003, at 13, 16, https://www.cis.org.au/app/uploads/2015/04/images/stories/policy-magazine/2003-spring/2003-19-3-james-m-buchanan.pdf.

for example, were concerned with the idea of tyranny, in which a tyrant would consume the government sphere and deprive citizens of the opportunity to participate in politics.[26] Meanwhile, one of the most prominent examples of regulatory capture, as well as some of the most well-known restrictions on occupational freedom, can be found in the medieval guild system. Guilds were associations of people who shared characteristics and sought to accomplish mutual goals.[27] They initially appeared during the era of the Roman Empire, and occasionally during the Dark Ages, from 400 to 1000 A.D.[28] Guilds reached their height in prominence during the Middle Ages, from 1000 to 1500 A.D., with some surviving into the 19th century.[29] They held legal monopolies over their respective professions, meaning that in order to work in a certain field, individuals needed to be affiliated with the relevant guild.[30] They also tended to be highly exclusive, with only a small minority being able to participate.[31] Guilds proved to be quite controversial. Supporters of guilds argued that they were necessary to ensure that the only individuals working in a given field were the ones who were qualified to do so.[32] However, opponents, including customers, employees, and those attempting to compete with the guilds, argued that they harmed the economy and harmed competition.[33] For example, Adam Smith, the Scottish economist and philosopher, argued that guilds were anticompetitive and that "[p]eople of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices."[34] Modern scholarship is divided regarding the benefits and costs of guilds as well, with some arguing that they generated economic benefits such as facilitating contract enforcement, enhancing commercial security, and improving information symmetry between consumers and producers, and others arguing that guilds created monopolies, stifled innovation, engaged in rent-seeking, and harmed disfavored groups.[35] Guilds represent one of the most notable examples of regulatory capture to appear in human history. They had a legal monopoly over their respective professions, which they used to protect their members from competition. However, they are far from the only example of powerful interests obtaining protections from competition in history.

---

26 Alissa Ardito, *Regulatory Capture, Ancient and Modern*, REG. REV. (June 30, 2016), https://www.theregreview.org/2016/06/30/ardito-regulatory-capture-ancient-and-modern/.

27 Sheilagh Ogilvie, *The Economics of Guilds*, 28 J. ECON. PERSP. 169, 169 n.1 (2014).

28 *Id.* at 170.

29 *Id.*

30 SHEILAGH OGILVIE, INSTITUTIONS AND EUROPEAN TRADE: MERCHANT GUILDS, 1000-1800 19-94 (2011).

31 Ogilvie, *supra* note 27, at 172-73.

32 *Id.* at 173.

33 *Id.* at 173-74.

34 ADAM SMITH, THE WEALTH OF NATIONS 72 (Shine Classics 2014) (1776).

35 Ogilvie, *supra* note 27, at 174.

The guild system had largely become a thing of the past by the end of the 19th century.[36] Nevertheless, regulatory capture did not die with it. The phenomenon of interests capturing regulatory bodies became a prominent concern in the economy of the United States. One particularly notable example of this phenomenon occurred from 1946 to 1947.[37] In 1945, Trans World Airlines (TWA), owned by the noted aviator and filmmaker Howard Hughes, was awarded the right to fly the Atlantic by the Civil Aeronautics Board. This decision came after stiff opposition from Pan American Airways (Pan Am), which would have competition on its extremely lucrative North Atlantic route for the first time.[38] In 1946, after TWA started flights to Europe, the Chosen Instrument Bill was introduced in Congress.[39] The bill, for which Pan Am lobbied heavily, would have required all airlines in the United States that were flying abroad to surrender their overseas routes to a consolidated international airline corporation that Pan Am owned.[40] After TWA and other airlines successfully lobbied to block the bill, the bill's author, Senator Ralph Owen Brewster (R-ME), introduced another bill, the Community Airline Bill, which had the same practical effect.[41] When Brewster was elected chairman of the Special Senate Committee Investigating the National Defense Program, he announced an investigation into a contract from the Second World War that Hughes had with the federal government to build a gigantic, all-plywood flying boat called the Hercules, which had not been completed before the end of the war and had never been flown.[42] Hughes believed that the investigations were motivated by corruption on Brewster's part, and that Brewster was investigating him at the behest of Pan Am and its president, Juan Trippe, to stop Hughes from competing on the Atlantic.[43] Hughes was not alone in this perception. Drew Pearson, the noted columnist, later described Brewster as Pan Am's "kept senator."[44] At the public hearings on the subject of Hughes's military dealings, which began on July 28, 1947, Hughes accused Brewster of using the investigation to benefit Pan Am at the expense of Hughes and TWA.[45] He claimed that Brewster had implied that he would stop the investigation if TWA merged with Pan Am, that the investigation was a smear campaign aimed at sabotaging TWA to benefit Pan

---

36 *Id.* at 170-73.

37 DONALD L. BARLETT & JAMES B. STEELE, HOWARD HUGHES: HIS LIFE AND MADNESS 145-56 (Paperback ed., W. W. Norton Co. 2004) (1979).

38 *Id.* at 145.

39 *Id.*

40 *Id.*

41 *Id.* at 145-46.

42 *Id.* at 146-59.

43 DONALD L. BARLETT & JAMES B. STEELE, Howard Hughes: His Life and Madness 145-46 (Paperback ed., W. W. Norton Co. 2004) (1979).

44 *Id.* at 146 (quoting DREW PEARSON, DIARIES 1949-1959 478 (Tyler Abdell ed., 1974)).

45 *Id.* at 151.

Am.[46] By portraying Brewster as beholden to a powerful special interest, Hughes was able to portray himself as an underdog to the press and audience despite his vast wealth.[47] Hughes proved to be so successful at the hearings that Senator Homer Ferguson (R-MI), who chaired the hearings, recessed them for three months under pressure from other Republicans to cancel them entirely because they proved to be disastrous.[48] Regulatory capture had seized headlines and the public square.

The Hughes hearings were not the last time in which airlines would feature prominently in debates about regulatory capture. Regulatory capture in the airline industry also played a major role in President Jimmy Carter signing the Airline Deregulation Act into law in 1978, phasing out the Civil Aeronautics Board.[49] Although the Civil Aeronautics Board ostensibly existed to protect the public from excessively high ticket prices, the reality of the situation was quite different. It became apparent in the late 1960s and early 1970s that airfares on federally regulated interstate flights were much higher than those covering the same distance within one state, despite the fact that the intrastate flights were not subject to federal regulation.[50] This occurred because the Civil Aeronautics Board never authorized the entry of new commercial airlines into the marketplace and because federal regulations prevented airlines from setting prices below the minimums set by the Civil Aeronautics Board to compete for customers. Instead, the airlines competed for customers by adding expensive measures like high-quality meals, even for coach flights, and offering more convenient departure and arrival times, which required purchasing more aircraft and hiring more personnel.[51] Although the lack of competition created by the Civil Aeronautics Board initially created large regulatory profits for the airlines, non-price competition eventually eroded these profits, to the point that the airlines recognized that the Civil Aeronautics Board could no longer protect them and asked to be deregulated.[52] Regulation of airlines therefore initially provided them with the benefit of protection from market competition, before the lack of flexibility created by the regulation eventually caught up with them.

Regulatory capture also represents a major motivation for the occupational licensing restrictions and many other economic regulations of today. Louisiana represents a notable example. Louisiana is the only state that requires florists, individuals who create floral arrangements, to get a license in

---

46 *Id.*

47 *Id.*

48 *Id.* at 146-56.

49 William F. Shughart II, *Airline Deregulation Act of 1978*, INDEP. INST. (Oct. 24, 2014), https://blog.independent.org/2014/10/24/airline-deregulation-act-of-1978/.

50 *Id.*

51 *Id.*

52 *Id.*

order to work in that field.[53] In response to the exceptionally onerous nature of the licensing requirements in existence at the time, such as an extremely subjective licensing exam with a pass rate of below 50%, three aspiring florists sued, arguing that the law was unconstitutional.[54] The Louisiana floristry licensing statute was blatantly motivated by a desire to keep competitors out of the marketplace. The exam was graded by already-licensed florists, enabling them to stop anyone from competing with them.[55] Additionally, the Louisiana Agriculture Commissioner testified that he helped prevent a reform bill that would have eliminated the licensing requirement from being passed because of an agreement that he had with established florists to support whatever measures they desired regarding licensing.[56] Louisiana floristry licensure is also not the only relatively recent example of using occupational licensing to undermine competition for the sake of established actors.

Another notable example of regulatory capture in the realm of occupational licensing appears in the form of regulations requiring a dentistry license to perform teeth whitening services. Teeth whitening services and products have become increasingly popular with consumers as they have grown more readily available.[57] They typically involve allowing customers to apply over-the-counter teeth whitening products in clean environments while teaching them how to use the products, as well as placing a safe Light Emitting Diode (LED) light in front of customers' mouths.[58] Despite the lack of danger associated with teeth whitening services, offering those services requires a dentistry license in many states. As of 2013, fourteen states clearly defined teeth whitening as a form of dentistry and therefore required a dentistry license for the provision of those services, while an additional twelve states were unclear as to whether unlicensed teeth whitening was illegal due to ambiguities in their statutory language or because their laws predated modern teeth whitening products and practices.[59] Meanwhile, the dental boards in other states often enforce their states' dentistry practice acts against those offering teeth whitening services under a broad interpretation of their

---

53 LA. STAT. ANN. §§ 3:3801-3:3816 (2020); Shoshana Weissmann and C. Jarrett Dieterle, *Louisiana is the only state that requires occupational licenses for florists. It's absurd.*, USA TODAY (Mar. 28, 2018), https://www.usatoday.com/story/opinion/2018/03/28/louisiana-only-state-requires-occupational-licenses-florists-its-absurd-column/459619002/.

54 Cynthia Joyce, *An Illegal Arrangement*, LEGAL AFF. (May-June 2004), https://www.legalaffairs.org/issues/May-June-2004/scene_joyce_mayjun04.msp; Weissmann and Dieterle, *supra* note 53.

55 Weissmann and Dieterle, *supra* note 53.

56 Neily, *supra* note 15, at 908 (citing Deposition of Bob Odom, July 9, 2004 at 40, 56-57, Meadows v. Odom, No. CA 03-960-B-M2 (M.D. La. July 9, 2004)).

57 Angela C. Erickson, *White Out: How Dental Industry Insiders Thwart Competition From Teeth-whitening Entrepreneurs*, INST. FOR JUSTICE 2 (2013), http://www.ij.org/images/pdf_folder/other_pubs/white-out.pdf.

58 *Id.* at 4.

59 *Id.* at 6-10.

authority.[60] As with the Louisiana floristry licensing scheme, attempts to require a dentistry license to whiten teeth are driven more by a desire to prevent competition than a desire to protect the health and safety of the public. Because dentists charge two to six times as much money for teeth whitening as salons and kiosks and because they are often paid tens of thousands of dollars annually for those services, laws excluding everyone but licensed dentistry professionals from offering teeth whitening services are frequently lobbied for by dental associations and boards.[61] Meanwhile, the overwhelming majority of complaints about unlicensed teeth whitening come from dentists, hygienists, and dental boards and associations rather than from consumers and allege unlicensed practice rather than consumer harm.[62] This suggests that fear of competition is the true motive for the complaints. The restrictions were designed to stop competition with licensed dentistry professionals for the ability to provide a highly lucrative service.

Restrictions such as licensure are merely the latest in a long line of similar practices that date back centuries. From floristry licensure in Louisiana and laws requiring a dentistry license to whiten teeth to the medieval guilds, regulatory capture has long been a part of human civilization.

B. *Economic Liberty as a Right*

Although the phenomenon of regulatory capture has a history dating back to before the Middle Ages, the notion that economic liberty is a right worthy of protection also has a long history. The right to earn a living in the field of one's choice without needless state interference is well-established and deeply rooted in the history and tradition of the United States.[63]

The tradition of the right to occupational liberty predates the foundation of the United States. The recognition of the right of individuals to engage in the occupation of their choice without fear of the government intervening on behalf of favored interests occurred during the reign of Queen Elizabeth I of England.[64] Thomas Allen manufactured and sold some playing cards in London. This was against the law at the time. Edward Darcy held a patent on the manufacture and sale of playing cards, which Queen Elizabeth had previously granted to Ralph Bowes in 1576 before its purchase by Darcy.[65] Allen was prosecuted for his unauthorized playing card sale, but was acquitted by

---

60 *Id.* at 10.

61 *See id.* at 1, 11-17.

62 *See id.* at 19.

63 *See, e.g.*, Truax v. Raich, 239 U.S. 33, 41 (1915) ("It requires no argument to show that the right to work for a living in the common occupations of the community is of the very essence of the personal freedom and opportunity that it was the purpose of the Amendment to secure."); Yick Wo v. Hopkins, 118 U.S. 356 (1886).

64 SANDEFUR, *supra* note 10, at 17.

65 *Id.* at 18.

the court.[66] The court held that the patent represented a violation of the common law under the Magna Carta.[67] The monopoly on the manufacture and sale of playing cards was deemed to have violated the right to engage in a trade and to have harmed the public through reduced employment and higher prices.[68] It would be the first time that occupational freedom would find protection under the English common law, but not the last.

Conflicts between the English monarchy on one side and Parliament and the court system on the other would continue after Queen Elizabeth's death. Elizabeth was succeeded by her nephew, James I, who was a strong believer in the divine right of kings, the idea that kings received their authority to rule from God.[69] James was opposed in this by a group of political leaders called the Whigs, who argued that Parliament and English law could limit the king's power. [70] One of the Whig leaders, Sir Edward Coke, who had ironically prosecuted Thomas Allen while he was the Attorney General of England and Wales, took the lead in opposing James with respect to his desire to grant legal monopolies after he was made Lord Chief Justice of England.[71] Coke was eventually removed from his post as Chief Justice due in part to his continuous decisions to limit James' power to grant monopolies.[72] He continued to resist monopolies after he was elected to Parliament, authoring the Statute of Monopolies to generally prevent the Crown from granting monopolies and writing a series of legal treatises criticizing the practice.[73] Coke was not alone in his opposition to monopolies, as other English judges struck down monopolies as violations of the right to occupational liberty as well.[74] The Whigs opposed monopolies because they were granted by the King without the backing of Parliament, they barred workers for engaging in occupations that would benefit the English economy and the workers themselves, and they used government power to benefit well-connected individuals rather than society in general.[75] As a result of the work of Coke and other Whigs, the English common law developed a strong tradition of opposing monopolies.[76] However, this tradition would not end on the shores of the British Isles.

---

66 *Id.*

67 Darcy v. Allen (The Case of Monopolies) (1603) 77 Eng. Rep. 1260, 1262-64 (KB).

68 *Id.* at 1262-64.

69 SANDEFUR, *supra* note 10, at 18.

70 *Id.*

71 *Id.* at 18-19.

72 *See*, *e.g.*, The Case of the Tailors of Ipswich (1615) 77 Eng. Rep. 1218 (KB); SANDEFUR, *supra* note 10, at 18-19.

73 SANDEFUR, *supra* note 10, at 19-21.

74 *See*, *e.g.*, Mounson v. Lyster (1632) 82 Eng. Rep. 122 (KB); The Case of the Bricklayers (1624) 81 Eng. Rep. 871 (KB); Colgate v. Bachelor (1602) 78 Eng. Rep. 1097 (KB); SANDEFUR, *supra* note 10, at 21.

75 SANDEFUR, *supra* note 10, at 19-20.

76 *Id.* at 21-23.

The American Founders also recognized the importance of occupational liberty in the wake of the American War for Independence. However, they took a different approach from the Whigs. Where the Whigs were heavily focused on combatting monopolies because of conflicts between the Crown and Parliament and a concern over the damage that monopolies could cause to the English economy, the Founders took the view that occupational freedom was a natural right.[77] Thomas Jefferson, for example, argued that individuals have a natural right to choose the profession that they think is most likely to give them "subsistence" and that individuals have a right to "free exercise of [their] industry."[78] James Madison, meanwhile, argued that legal monopolies violated the natural right of individuals to earn a living through their chosen occupation and that they violated the principle of equality by granting favors to people with political influence.[79] Many of the most well-known documents from the Founding Era reflect this view of occupational liberty as a natural right. For example, the Declaration of Independence proclaims that among the "certain inalienable Rights" held by human beings are "Life, Liberty, and the pursuit of Happiness."[80] The "pursuit of Happiness" referenced in the Declaration is likely a reference to, among other things, the right to improve one's position in life through pursuit of a legitimate occupation in light of the belief of the Founders in the right to engage in business and keep what one earned, though it should be noted that the significance of the choice of this language by Jefferson is debated among scholars.[81] Another prominent example can be found in the Virginia Declaration of Rights, written by George Mason, which declares that among the "inherent rights" of mankind are the means of acquiring property and pursuing and obtaining happiness.[82] Occupational freedom would continue to receive robust protection for the first century of the history of the United States.[83] However, this protection would not last, and would come to be undermined as the rational basis test came to play a prominent role in American constitutional law.

---

77 *Id.* at 18-24.

78 *Id.* at 24 (quoting THOMAS JEFFERSON, THOUGHTS ON LOTTERIES (1826), *reprinted in* 17 THE WRITINGS OF THOMAS JEFFERSON 449 (Albert Ellery Bergh ed., Thomas Jefferson Memorial Association 1907); Letter from Thomas Jefferson to Joseph Milligan (Apr. 6, 1816), in 14 THE WRITINGS OF THOMAS JEFFERSON 466 (Albert Ellery Bergh ed., Thomas Jefferson Memorial Association 1907)).

79 *Id.* (quoting JAMES MADISON, *Property* (1792), *reprinted in* JAMES MADISON: WRITINGS 516 (Jack Rakove ed., Library of America 1999)).

80 THE DECLARATION OF INDEPENDENCE para. 2 (U.S. 1776).

81 SANDEFUR, *supra* note 10, at 24.

82 VIRGINIA DECLARATION OF RIGHTS para. 1 (1776); SANDEFUR, *supra* note 10, at 24.

83 NEILY, *supra* note 13, at 41.

C. *The Rise of the Rational Basis Test*

Rational basis review is the most deferential of the three standards of review by which courts analyze the constitutionality of laws. The most rigorous standard of review, strict scrutiny, requires the government to prove that the law in question is necessary to fulfil a compelling state interest, the law is narrowly tailored to achieve that interest, and uses the least restrictive means to achieve that interest.[84] Intermediate scrutiny, which is less rigorous but still provides meaningful constitutional review, requires the government to prove that the law is substantially related to an important government interest and uses means that are substantially related to that interest.[85] Finally, rational basis review requires that the plaintiff prove that the law is not rationally related to a legitimate government interest.[86] The test stems from a movement to make the judiciary more deferential to lawmakers.[87] It would prove to be highly successful in this regard.

The rational basis test has existed informally since 1877.[88] In *Munn v. Illinois*, which dealt with an Illinois law setting maximum rates that could be charged by private companies for the storage and transportation of grain, the United States Supreme Court upheld the constitutionality of the law, holding that it was permissible as a part of state police powers.[89] In what would be the precursor of rational basis review, the Court held that it would presume that the statute was valid if there was any conceivable set of circumstances to justify it.[90] This attitude of judicial deference would soon become prominent in the Progressive movement of the late 19th and early 20th centuries, which took the view that the state represented a powerful agent for social change and that majority rule should be supported.[91] Law professor James Bradley Thayer of Harvard Law School, for example, argued in an 1893 law review article that judges should almost always defer to legislators and that laws should be upheld unless their unconstitutionality is "so clear that it is not open to rational question."[92] Similarly, United States Supreme Court Associate Justice Oliver Wendell Holmes, Jr., a standard bearer of the Progressive movement in the federal judiciary, was a strong proponent of judicial

---

84 *See*, *e.g.*, Johnson v. California, 543 U.S. 499, 505 (2005).

85 *See*, *e.g.*, Turner Broad. Sys. v. FCC, 520 U.S. 180, 186 (1997).

86 *See*, *e.g.*, Williamson v. Lee Optical of Okla., Inc., 348 U.S. 483, 487-91 (1955); United States v. Carolene Prods. Co., 304 U.S. 144, 152-54 (1938).

87 ROOT, *supra* note 14, at 135-36.

88 Neily, *supra* note 15, at 898 (citing Munn v. Illinois, 94 U.S. 113, 132 (1877) ("If no state of circumstances could exist to justify such a statute, then we may declare this one void, because in excess of the legislative power of the State. But if it could, we must presume it did.")).

89 Munn, 94 U.S. at 123, 125-36.

90 *Id.* at 132.

91 ROOT, *supra* note 14, at 50-54.

92 James Bradley Thayer, *The Origin and Scope of the American Doctrine of Constitutional Law*, 7 HARV. L. REV 129, 144 (1893).

deference and not interfering with majority rule.[93] The Progressives, particularly Holmes, would come to have a massive impact on constitutional law in subsequent decades.

Despite efforts to make the courts more deferential to legislatures in the area of economic affairs, the courts remained protective of economic liberty for some time. The Supreme Court was particularly protective of economic rights during the so-called *Lochner* era, named for the case *Lochner v. New York*, in which the Court struck down a New York law setting limits on the number of hours that bakery employees could work as a violation of the right to freedom of contract, which it held was protected by the Due Process Clause of the Fourteenth Amendment.[94] During this period, from late in the 19th century until the New Deal era, the Court struck down laws limiting economic liberty, including those restricting occupational freedom.[95] The era represented a rejection of the advocacy for judicial deference that characterized the views of Oliver Wendell Holmes and his ideological allies, with the Court favoring a more engaged judiciary with respect to economic rights.[96] This era would ultimately come to an end during the administration of President Franklin Delano Roosevelt.

Shortly after taking office, Roosevelt would come into conflict with the Supreme Court over the constitutionality of his New Deal programs. At the time, the Court had four libertarian-leaning justices, melodramatically nicknamed the "Four Horsemen" by Roosevelt's allies, who favored giving economic regulations rigorous judicial review.[97] The Four Horsemen, Justices Pierce Butler, James McReynolds, George Sutherland, and Willis van Devanter, were opposed by the Progressive "Three Musketeers," Justices Louis Brandeis, Benjamin Cardozo, and Harlan Stone.[98] The remaining two justices served as swing votes, with Chief Justice Charles Evan Hughes being more inclined to vote with the Musketeers and Justice Owen Roberts tending to vote more often with the Horsemen.[99] In response to the Court's decision to strike down a number of New Deal laws in 1935, Roosevelt attempted a court-packing plan which would have given Roosevelt the power to appoint one new federal judge for every federal judge who had served at least ten

---

93 *See*, *e.g.*, Bartels v. Iowa, 262 U.S. 404, 412-13 (1923) (Holmes, J., dissenting) (Holmes argued that a law restricting the teaching of foreign languages could be a reasonable and necessary method of ensuring that United States citizens spoke a common tongue, and that the law should be upheld as constitutional); Lochner v. New York, 198 U.S. 45, 74-76 (1905) (Holmes, J., dissenting) (Holmes argued that a law limiting the amount of hours worked by workers in bakeries was constitutional based on judicial deference); ROOT, *supra* note 14, at 41-62.

94 Lochner, 198 U.S. at 53-65; ROOT, *supra* note 14, at 45-61.

95 NEILY, *supra* note 13, at 41-48; ROOT, *supra* note 14, at 45-61.

96 ROOT, *supra* note 14, at 45-61.

97 *Id.* at 73.

98 *Id.*; Daniel E. Ho and Kevin M. Quinn, *Did A Switch in Time Save Nine?*, 2 J. LEGAL ANALYSIS 93 (2010).

99 ROOT, *supra* note 14, at 73; Ho and Quinn, *supra* note 98, at 2-3.

years and who had not retired within six months of reaching age 70.[100] Although the court-packing plan failed as a result of backlash from the Court, Congress, and the American public, Roosevelt's New Deal would ultimately find much more success in the Court in 1937, when Justice Roberts' jurisprudence shifted in a direction more friendly to the New Deal in a move dubbed "the switch in time that saved nine" due to speculation that it was aimed at protecting the Court from packing efforts by Roosevelt.[101] The rational basis test would arise from this environment.

The rational basis test gained a solid place in American constitutional law in the wake of Roberts' switch. The test had previously been applied to economic regulations in *Nebbia, v. New York*, which held that a New York law setting the price of milk for dairy farmers was constitutional under the deferential standard.[102] The case *United States v. Carolene Products Company*, which upheld a federal law banning the shipping of filled milk, skim milk with a fat or oil added in place of milkfat, in interstate commerce, then applied rational basis review to economic regulations in 1938.[103] Since then, the test has enjoyed a dominant position in constitutional law with respect to economic regulations.[104] As judicial deference took a place of prominence in American jurisprudence, becoming popular among conservatives after initially rising from the Progressives of the 19th and early 20th centuries, rational basis review has faced little challenge for its position as the standard of review for economic regulations.[105] It has ensured that the courts treat economic restrictions deferentially to this day.

II. THE RATIONAL BASIS TEST INCENTIVIZES REGULATORY CAPTURE

Because the rational basis test is so deferential to government actors, it is unable to provide an effective deterrent against instances of regulatory

100 Humphrey's Executor v. United States, 295 U.S. 602, 628 (1935) (holding that President Roosevelt could not fire a member of the Federal Trade Commission without the approval of the Senate for political reasons); Louisville Joint Stock Land Bank v. Radford, 295 U.S. 555, 601-02 (1935) (holding that a federal law making it easier for farmers to buy back their farms after foreclosure was unconstitutional under the Fifth Amendment); A.L.A. Schechter Poultry Corp. v. United States, 295 U.S. 495, 552-55 (1935) (held that regulations on the poultry industry, including wage and price fixing, were invalid under the nondelegation doctrine and the Commerce Clause); ROOT, *supra* note 14, at 67-72.

101 DAMON ROOT, *supra* note 14, at 73-76 (2014); National Constitution Center Staff, *How FDR lost his brief war on the Supreme Court*, NATIONAL CONSTITUTION CENTER (Feb. 5, 2020), https://constitutioncenter.org/blog/how-fdr-lost-his-brief-war-on-the-supreme-court-2.

102 291 U.S. 502, 537-39 (1934).

103 304 U.S. 144, 146, 152-54 (1938).

104 *See*, *e.g.*, Williamson v. Lee Optical of Okla., Inc., 348 U.S. 483, 487-91 (1955); Sensational Smiles, LLC v. Mullen, 793 F.3d 281, 284 (2d Cir. 2015); Meadows v. Odom, 360 F. Supp. 2d 811, 816-25 (M.D. La. 2005), *vacated as moot*, 198 Fed. Appx. 348 (5th Cir. 2005).

105 NEILY, *supra* note 13, at 56-63, 107-13; ROOT, *supra* note 14, at 78-131.

capture. The test does not provide meaningful review to economic regulations, meaning that those regulations will be upheld even if they are clearly aimed at stopping legitimate competition rather than existing on the basis of a public health and safety rationale.[106] This enables actors seeking to engage in regulatory capture to more easily stop competitors from entering the market without fear of the regulations protecting them being struck down by the courts.

A. *The Rational Basis Test Does Not Provide Meaningful Judicial Review*

As mentioned in Part I, the rational basis test requires the law under review to be rationally related to a legitimate government interest, with the plaintiff bearing the burden to prove that it is not.[107] This standard is extremely deferential to the government, to the point that the requirement that laws be rationally related to a legitimate government interest places few, if any, substantive limits on the ability of the state to regulate the economy.[108] This leaves the test unable to provide any form of effective deterrent against regulatory capture and practically makes it a rubber stamp for even the most blatant of anticompetitive statutes.

For example, the rational basis test does not require that the "rational basis" in question be based on empirical fact. Instead, "if there is any reasonably conceivable state of facts that could provide a rational basis," the constitutionality of the law is upheld.[109] This means that the reviewing court will not look to the actual rationale for which the law was created, and will instead rely on justifications which may be purely hypothetical.[110] The actual reasons for which the law was created, whether they be for a legitimate public health or safety purpose or to protect a favored interest group from competition, are irrelevant for the court's determination.[111] All that matters is that a state of affairs might hypothetically exist that could justify the restriction. This hypothetical state of affairs could be extremely unlikely to ever occur or could be so completely absurd that it would never be contemplated or accepted as a good reason outside the confines of a courtroom in a rational basis case.[112] For example, in *Williamson v. Lee Optical of Oklahoma, Inc.*, which

---

106 *See*, *e.g.*, Sensational Smiles, 793 F.3d at 284-88; Meadows, 360 F. Supp. 2d at 816-25.

107 *See*, *e.g.*, F.C.C. v. Beach Commc'ns, Inc., 508 U.S. 307, 315 (1993); Williamson, 348 U.S. at 487-91; Carolene Prods., 304 U.S. 144, 146-54 (1938).

108 *See*, *e.g.*, NEILY, *supra* note 13, at 50-51; Timothy Sandefur, *Rational Basis Scrutiny Is Just a Stupid Rock*, CATO UNBOUND (Feb. 18, 2014), https://www.cato-unbound.org/2014/02/18/timothy-sandefur/rational-basis-scrutiny-just-stupid-rock.

109 Beach Commc'ns, 508 U.S. at 313.

110 *See*, *e.g.*, *id.* at 315.

111 *Id.*

112 *See*, *e.g.*, Meadows v. Odom, 360 F. Supp. 2d 811, 816-25 (M.D. La. 2005), *vacated as moot*, 198 Fed. Appx. 348 (5th Cir. 2006).

dealt with the constitutionality of an Oklahoma statute making it illegal for anyone besides a licensed optometrist or ophthalmologist to fit eyeglass lenses to a face, duplicate lenses, or replace older lenses or other optical appliances into frames without a written prescription from an eye doctor, the United States Supreme Court held that only hypothetical rationales are relevant in rational basis cases.[113] In upholding the constitutionality of the statute, the Court held that the Oklahoma legislature might have concluded that it was necessary to regulate lenses to regulate eye health, that an eye exam was needed often enough with new lenses to make them necessary in all cases, or that prescriptions were necessary to detect ailments.[114] The Court invented these rationales out of whole cloth and ignored the possibility that opticians had been arbitrarily licensed out of a well-established activity to benefit licensed eye doctors, despite ample evidence that this was the case.[115] The rational basis test is deferential to the point that hypothetical justifications are considered rational bases despite evidence that the real reason for which the law in question was enacted is to protect favored actors from competition.

Additionally, the plaintiffs in rational basis cases, who bear the burden of proof as mentioned above, must negate every single one of these hypothetical justifications for the laws they are challenging.[116] This presents a significant challenge for plaintiffs to overcome, to the point of near hopelessness. As a matter of formal logic, proving that something does not exist, such as a valid justification for a law in a rational basis case, is actually impossible.[117] This is because to know that no valid hypothetical justification exists would require omniscience, an ability which no human being possesses. One can prove that a state of affairs has never been encountered or is extremely unlikely to occur, but proving that it cannot occur is impossible as a practical matter without limitless knowledge.[118] Truth cannot be established simply because something has not been proven false.[119] No human possesses the knowledge necessary to conclusively declare that every single hypothetical justification for a law is invalid, so all that plaintiffs can do is show that the justifications they can find can be negated, not that no valid justifications exist. Logically speaking, the government should bear the burden of proof to prove that its laws are valid, not the individual wanting to be left alone,

---

113 Williamson v. Lee Optical of Okla., Inc., 348 U.S. 483, 485-88 (1955).

114 *Id.* at 487.

115 *Id.* at 485-91; Lee Optical of Okla., Inc. v. Williamson, 120 F. Supp. 128, 137, 140-41 (W.D. Okla. 1954), *aff'd in part and rev'd in part by* Williamson v. Lee Optical of Okla., Inc., 348 U.S. 483 (1955); NEILY, *supra* note 13, at 61-62.

116 Beach Commc'ns, 508 U.S. at 315; Lehnhausen v. Lake Shore Auto Parts Co., 410 U.S. 356, 364 (1973).

117 Sandefur, *supra* note 108.

118 Neily, *supra* note 15, at 908-09.

119 IRVING M. COPI ET AL., INTRODUCTION TO LOGIC 130-33 (Pearson International ed., 14th ed., Pearson Education Limited 2014).

since the government is making the claim that its laws are justified restrictions on individual liberty and a statement cannot be proven true simply because it has not been proven false.[120] As it is, the test stacks the deck against plaintiffs significantly, requiring them to be able to respond to hypothetical justifications that could be dreamed up not just by the government's attorneys, but by anyone else.[121] The obstacles plaintiffs face as a result of the requirement that they negate every single conceivable justification for the statutes they are challenging are practically insurmountable.

Meanwhile, in some federal circuit courts of appeals, the test has been interpreted so that its deference goes beyond even allowing hypothetical justifications and requiring plaintiffs to negate every single possible scenario. Some circuit courts have gone so far as to hold that the test not only permits, but mandates that the reviewing court come up with valid justifications for the law on behalf of the government.[122] There is a strong argument that this violates due process.[123] This is because the test actively encourages judicial bias when framed this way.[124] The promise of a neutral adjudicator is one of the most basic principles of due process of law, and the rational basis test severely undermines this principle by allowing or requiring the reviewing court to assist one of the parties in winning the case.[125] Furthermore, the ability of judges to actively help the government win the case places an additional barrier in the path of plaintiffs. In addition to negating every conceivable hypothetical justification that the government can come up with, they must also do so for every one of those justifications developed by the judge as well, at least in jurisdictions in which the rational basis test has been applied to allow for this level of bias.[126] This, in turn, reduces the deterrent effect of the rational basis test even further. Plaintiffs must accomplish the herculean task of negating all the hypothetical justifications for a law while contending with an adjudicator who is hostile rather than neutral. The judge will credulously accept any justification offered for the law's validity and may even help the government come up with justifications. This leads to courts upholding the constitutionality of laws that blatantly protect the interests of favored groups by restricting the right of their competitors to enter the market.

---

120 *See id.*; Sandefur, *supra* note 108.

121 Neily, *supra* note 15, at 908-09; Sandefur, *supra* note 108.

122 Flying J, Inc. v. City of New Haven, 549 F.3d 538, 547 n.2 (7th Cir. 2008); Powers v. Harris, 379 F.3d 1208, 1217 (10th Cir. 2004); Starlight Sugar, Inc. v. Soto, 253 F.3d 137, 146 (1st Cir. 2001); Shaw v. Or. Pub. Emps.' Ret. Bd., 887 F.2d 947, 948 (9th Cir. 1989).

123 Neily, *supra* note 15, at 907, 910-12; Ward, *supra* note 15, at 715-35.

124 Ward, *supra* note 15, at 721-24.

125 *See*, *e.g.*, Ward v. Village of Monroeville, 409 U.S. 57, 61-62 (1972); Tumey v. Ohio, 273 U.S. 510, 532 (1927).

126 Neily, *supra* note 15, at 906-09; Sandefur, *supra* note 108; Ward, *supra* note 15, at 721-26.

B. *Meadows v. Odom – Louisiana Floristry Licensure*

One notable example of the rational basis test's failure to provide any sort of deterrent effect against blatantly protectionist legislation came in *Meadows v. Odom*, a case from the United States District Court for the Middle District of Louisiana which dealt with the constitutionality of a law requiring florists to get a license to work.[127] As discussed in Part I, Louisiana was, and remains, the only state that requires florists to get a license in order to work in that field.[128] The exam to become a florist was extremely difficult and was graded by already-licensed florists, enabling them to stop anyone from competing with them.[129] In response to the burdensome licensing requirements, three individuals desiring to become florists sued, arguing that the law was unconstitutional.[130] During deposition, the Louisiana Agriculture Commissioner revealed his role in saving the licensing requirement from elimination because he had agreed to work for the benefit of already licensed florists and that he had not consulted with any consumers or consumer groups in doing so.[131] The licensing scheme was thus created at the behest of established florists and enabled them to keep potential rivals from entering the marketplace. It represents a textbook example of regulatory capture, to the point that the regulatory body in question admitted that the licensing framework was in place to benefit licensed florists.

Despite this, the district court held that the floristry licensing law was constitutional under rational basis review.[132] The court held that the right to pursue an occupation of one's choosing was "a protected liberty interest, subject to reasonable regulation."[133] Regarding reasonableness, the court held that for an occupational regulation to be reasonable, it only had to "have a rational connection" to the individual's "fitness or capacity to serve in that … profession" and "an arguably legitimate state interest in regulating that … profession."[134] When it upheld the constitutionality of the licensing statute, the court made no mention of the Louisiana Horticulture Commissioner's admission that the law had been created for the benefit of established florists. Instead, the court's rationale focused on hypothetical physical dangers that

127 *See generally*, Meadows v. Odom, 360 F. Supp. 2d 811, 822 (M.D. La. 2005), *vacated as moot*, 198 Fed. Appx. 348 (5th Cir. 2005).

128 LA. REV. STAT. §§ 3:3801-3816 (2020); Weissmann and Dieterle, *supra* note 53.

129 Weissmann and Dieterle, *supra* note 53.

130 Jacob Sallum, *Flower Power*, REASON (Jan. 9, 2004), https://reason.com/2004/01/09/flower-power-2/.

131 Neily, *supra* note 15, at 908 (citing Deposition of Bob Odom, July 9, 2004 at 40, 56-57, Meadows v. Odom, No. CA 03-960-B-M2 (M.D. La. July 9, 2004)).

132 Meadows, 360 F. Supp. 2d at 825.

133 *Id.* at 813.

134 *Id.* at 813-14.

could arise from unlicensed floristry.[135] These dangers included "not having an exposed pick, not having a broken wire, not hav[ing] a flower that has some type of infection, like, dirt that remained on it when it's inserted into something [the florists are] going to handle."[136] The court also accepted the state's claim that licensing florists was necessary to enhance the reputation of Louisiana's floristry industry, even though none of the other 49 states felt that floristry licensure was necessary to achieve this goal because they do not license florists.[137] The court even held that industry protectionism as a goal of an economic regulation does not invalidate that regulation under rational basis review.[138] The court upheld the constitutionality of a piece of legislation that was clearly aimed at protecting those already established in the field from competition without providing meaningful review as a result of the exceedingly deferential nature of the rational basis test. This lack of meaningful review is all too common in rational basis cases.

C. *Sensational Smiles, LLC v. Mullen – Connecticut Teeth Whitening*

The rational basis test has resulted in courts turning a blind eye to regulatory capture and protectionism in rational basis cases related to matters other than floristry. Another prominent instance in which rational basis review resulted in the reviewing court upholding a blatantly protectionist licensing restriction occurred in *Sensational Smiles, LLC v. Mullen*, a case from the United States Court of Appeals for the Second Circuit examining the constitutionality of a declaratory ruling by the State Dental Commission of Connecticut which held that only licensed dentists could provide teeth whitening services using an LED light.[139] As noted in Part I of this article, these types of restrictions are frequently lobbied for by dental professionals in order to prevent others from competing with them for the right to practice the lucrative business of teeth whitening, for which licensed dentists often charge significantly more than kiosks and salons.[140] The ruling by Connecticut's State Dental Commission was no exception. Six of the nine members of the Commission were required to be licensed dentists, so the members of the Commission had a clear financial incentive to keep non-dentists out of the teeth whitening market.[141] Testimony during litigation showed that there

---

135 *Id.* at 823-24.

136 *Id.* at 824.

137 *Id.* at 824-25.

138 Meadows, 360 F. Supp. 2d at 824.

139 Sensational Smiles, LLC v. Mullen, 793 F.3d 281, 283 (2d Cir. 2015).

140 Erickson, *supra* note 53, at 1-16.

141 Matthew Yglesias, *Lisa Martinez vs The Dentists' Cartel*, SLATE (Nov. 23, 2011), https://slate.com/business/2011/11/cheap-teeth-whitening-services.html.

were no safety hazards associated with using LED lights to whiten teeth.[142] Instead, the ruling was motivated by a desire to ensure that only licensed dentists could offer those services, thus putting a halt to legitimate competition.[143] As was the case with Louisiana's floristry licensing scheme, the Commission's ruling was an attempt to shield a politically favored interest group from having to compete with others for business.

Sensational Smiles, LLC, a teeth-whitening business that was not owned by dentists, sued after receiving a cease-and-desist letter from the Connecticut State Department of Public Health threatening them with legal action if they continued to offer teeth whitening services and argued that the ruling violated the Equal Protection and Due Process Clauses of the Fourteenth Amendment.[144] Unfortunately for Sensational Smiles, the United States District Court for the District of Connecticut denied its request for a declaratory judgment that the ruling was unconstitutional as applied and a permanent injunction preventing its enforcement.[145] The Second Circuit affirmed the district court's decision. Applying rational basis review, the court held that the government had a legitimate interest in protecting the oral health of the public.[146] Although dentists were not trained in the use of LED lights, so there was little basis for limiting the use of the lights to dentists, the court held that the Commission might have decided that dentists were better equipped than non-dentists to treat any hypothetical harms resulting from the use of LED lights or that customers should consult a dentist before using the lights in view of the alleged risks.[147] In response to the argument by Sensational Smiles that the ruling did not bar people from pointing the lights at their own mouths but forbade teeth-whitening professionals from doing it for them, the court stated that "[t]he law … does not require perfect tailoring of economic regulations, and the Dental Commission can only define the practice of dentistry; it has limited control over what people choose to do to their own mouths."[148] Application of the rational basis test therefore resulted in the court upholding the constitutionality of a law that had little ability to combat the alleged harm based on little more than blind deference.

The Second Circuit took its deference a step further, however. Not only did it uphold the ruling based on an illogical rationale for combatting the alleged risks associated with LED-based teeth whitening services, it also held that the prevention of legitimate market competition constituted a legitimate

142 Memorandum of Law in Support of Plaintiff's Motion for Summary Judgment at 11, Martinez v. Mullen, 11 F. Supp. 3d 149 (D. Conn. 2014) (No. 3:11-CV-01787-MPS).

143 Damon Root, *Federal Court Allows State Government to Practice 'Naked Economic Protectionism'*, REASON (July 22, 2015, 11:27 A.M.), https://reason.com/2015/07/22/federal-court-allows-state-government-to/.

144 Sensational Smiles, 793 F.3d at 283-84.

145 *Id.* at 284.

146 *Id.* at 288.

147 *Id.* at 285.

148 *Id.*

state interest.[149] Blatant favoritism in support of an entrenched interest group represented a valid interest in the eyes of the court because "[w]e call this politics."[150] The Second Circuit argued that even without an arguably consumer-friendly justification such as attempting to use higher priced teeth-whitening services to subsidize other services that can only be offered by dentists, it was still constitutionally permissible for the state to favor dentists over teeth whiteners.[151] Based on both credulous acceptance of the Commission's illogical rationales for the restriction as well as the view that pure anticompetitive favoritism constituted a legitimate form of government activity, the court upheld the constitutionality of the restriction under rational basis review.[152] The rational basis test had once again failed to provide an effective shield against a law that was clearly designed to protect a well-connected interest group from having to fairly compete with those wishing to provide similar services based on a shaky public safety justification.

The rational basis test is ultimately unable to provide an effective deterrent against regulatory capture and cronyism because it is too deferential to government actors to provide an effective constraint on their powers. It requires plaintiffs to achieve a herculean, arguably impossible task of negating every conceivable hypothetical justification for a statute or regulation, whether it be real or hypothetical, and it can also allow courts to actively help the government win the case. In some jurisdictions, it even treats pure favoritism as a legitimate government interest. This leads to patently protectionist laws such as Louisiana's floristry licensure and Connecticut's ban on anyone but licensed dentists providing LED-based teeth whitening services being held to have been rationally related to a legitimate government interest. A stricter standard of review is necessary to avoid these issues and ensure that occupational licensing restrictions are given more rigorous evaluation.

III. ECONOMIC REGULATIONS SHOULD RECEIVE STRICT SRUTINY

Instead of rational basis review, the standard of review used to analyze the constitutionality of economic regulations should be strict scrutiny. As mentioned in Part I, with strict scrutiny, the government must prove that the law in question is necessary to achieve a compelling state interest, the law must be narrowly tailored to achieve that interest, and the law must utilize the least restrictive means to achieve that interest.[153] Strict scrutiny represents a superior alternative to rational basis review for two reasons. First, it is appropriate that restrictions on occupational freedom receive strict scrutiny

149 *Id.* at 285-88.
150 Sensational Smiles, 793 F.3d at 287.
151 *Id.*
152 *Id.* at 286-88.
153 *See*, *e.g.*, Johnson v. California, 543 U.S. 499, 505 (2005).

because occupational freedom is a fundamental right. Second, strict scrutiny involves a rigorous examination of the government's motives for passing the legislation, making it much better equipped to combat regulatory capture than the rational basis test.

A. *Strict Scrutiny is Appropriate because Occupational Liberty is a Fundamental Right*

Strict scrutiny is the standard of review used to review the constitutionality of restrictions on fundamental rights.[154] It is for this reason that strict scrutiny represents the most appropriate standard of review for regulations impacting occupational freedom. Occupational freedom is a fundamental right as the term is defined in case law because it has roots in the history and tradition of the United States and is implicit in a system of ordered liberty.[155] It should therefore receive strict scrutiny.

In order for a right to be classified as "fundamental" and receive strict scrutiny, it must meet certain criteria. First, fundamental rights are deeply rooted in the history and tradition of the United States.[156] For a right to be based in history and tradition, there must be a strong historical basis supporting the right's existence. Fundamental rights have a long history, dating back to the founding of the United States or finding their basis in English common law.[157] For example, in *Washington v. Glucksburg*, the Court held that the right to have assistance in committing suicide was not fundamental.[158] This was on the grounds that the United States had laws against assisted suicide dating back to 1828 and that English common law had prohibited suicide and assisted suicide for over seven centuries, and punished the practices harshly.[159] Second, fundamental rights are imbedded in a system of ordered liberty, to the point liberty and justice would not exist if they were infringed.[160] This means that they are essential to an orderly pursuit of happiness by free individuals.[161] The purpose of this framework is to provide guideposts for the Court's decision-making and ensure that interpretation of the Due Process Clause is based on articulable legal principles rather than the

---

154 *See*, *e.g.*, Troxel v. Granville, 530 U.S. 57 (2000); Boy Scouts of Am. v. Dale, 530 U.S. 640 (2000); Zablocki v. Redhail, 434 U.S. 374 (1978); Bounds v. Smith, 430 U.S. 817, 828 (1977); Harper v. Va. State Bd. of Elections, 383 U.S. 663 (1966); United States v. Guest, 383 U.S. 745, 757 (1966).

155 *See* Washington v. Glucksberg, 521 U.S. 702, 720-21 (1997); NEILY, *supra* note 13, at 156-57; SANDEFUR, *supra* note 10, at 17-25.

156 Glucksberg, 521 U.S. at 721 (citing Moore v. E. Cleveland, 431 U.S. 494, 503 (plurality opinion)).

157 *Id.* at 710-16, 720-21.

158 *Id.*

159 *Id.*

160 *Id.* at 721 (citing Palko v. Connecticut, 302 U. S. 319, 325, 326 (1937)).

161 Loving v. Virginia, 388 U. S. 1, 12 (1967).

mere policy preferences of the justices on the Court.[162] The Court has held that the right to marry, the right to have children, the right of parents to direct the education and upbringing of their children, and the right to use contraceptives are protected rights.[163] It should afford occupational freedom with similarly rigorous protections.

Occupational liberty is a fundamental right as the Court currently defines that term, and therefore should receive strict scrutiny. The right to earn a living without fear of unwarranted government intrusion is deeply rooted in the history and tradition of the United States. Occupational liberty has a long history in English common law, dating all the way back to the Elizabethan era in the early seventeenth century and having been heavily promoted by the Whigs in the years since.[164] This tradition continued during the founding of the United States, when founders such as Thomas Jefferson, James Madison, and George Mason acknowledged that the right to earn a living in the occupation of one's choosing was a crucial natural right that was worthy of recognition and protection.[165] This protection was reflected in much of the history of the United States, with few jobs requiring an individual to get a license to enter that profession. In fact, only around five percent of occupations required a license 60 years ago, compared to around 25 percent today.[166] Occupational liberty has sufficiently deep roots in our history and tradition to warrant strict scrutiny protection. Additionally, occupational liberty is implicit in a system of ordered liberty. The ability to choose a legitimate occupation without undue interference is essential to daily living. One's job enables one to earn money to put food on the table. It also represents a major way in which people can interact with others, and people can find fulfillment in their work due to the nature of their jobs, assuming that they enjoy it. It is essential to the ability of free people to pursue happiness. Liberty and justice are lost when the ability of individuals to work in their chosen professions is infringed unjustifiably. Occupational liberty therefore meets the necessary criteria to receive strict scrutiny.

---

162 Glucksberg, 521 U.S. at 720-21.

163 Eisenstadt v. Baird, 405 U. S. 438, 453-54 (1972) (holding that the rights of individuals, single or married, to be free from government intrusion into the decision to have children, including the use of contraception, are the same); Loving, 388 U. S. at 12 (holding that the Due Process Clause protects the right to marry because it is crucial to the ability of free people to pursue happiness); Griswold v. Connecticut, 381 U. S. 479, 485-86 (1965) (holding that married individuals have the right to use contraceptives under the Due Process Clause based on the long history of the right to privacy); Skinner v. Oklahoma *ex rel.* Williamson, 316 U. S. 535, 545 (1942) (holding that the right to have children is protected by the Due Process Clause); Pierce v. Society of Sisters, 268 U. S. 510, 534-36 (1925) (holding that an Oregon law requiring public school attendance and banning private and parochial schools violated the Due Process Clause); Meyer v. Nebraska, 262 U. S. 390, 399-403 (1923) (holding that a Nebraska law prohibiting the teaching of foreign languages violated the Due Process Clause).

164 *See*, *e.g.*, Mounson v. Lyster (1632) 82 Eng. Rep. 122 (KB); The Case of the Bricklayers (1624) 81 Eng. Rep. 871 (KB); Darcy v. Allen (*The Case of Monopolies*) (1603) 77 Eng. Rep. 1260, 1262-64 (KB); Colgate v. Bachelor (1602) 78 Eng. Rep. 1097 (KB); SANDEFUR, *supra* note 10, at 17-23.

165 SANDEFUR, *supra* note 10, at 24.

166 Herman, *supra* note 1.

File: Brown 17.1 To Publish_Edited Created on: 4/19/2022 10:18:00 AM Last Printed: 6/14/2022 9:27:00 PM

Occupational freedom is a fundamental right. It is deeply rooted in the history and tradition of the United States, going back to the founding and English common law, and is implicit in a system of ordered liberty to the point that liberty and justice would be irreparably harmed if it were violated because it is crucial to the ability to pursue personal fulfillment. It should therefore receive strict scrutiny. However, consistency with the Court's definition of "fundamental right" is not the only reason for which the right to earn a living in the occupation of one's choice should receive strict scrutiny. Strict scrutiny has other advantages as well.

B. *Strict Scrutiny Involves Rigorous Review and Can Combat Regulatory Capture*

In addition to meeting the criteria to qualify as a fundamental right, meaning that it should receive strict scrutiny under the standard set by the Supreme Court, occupational liberty should also receive strict scrutiny because strict scrutiny, in contrast to the rational basis test, involves rigorous review of the motives involved in passing the law in question.[167] This enables its use to combat regulatory capture far more effectively than the more credulous standard of rational basis review.

Under strict scrutiny, the government must have a compelling interest supporting the law and the law must be narrowly tailored to fulfill that interest using the least restrictive means possible.[168] This requires courts to examine the government's motives for enacting legislation. In fact, the ability of the test to identify and strike down laws with illegitimate motives is regarded as one of the features of the test.[169] For example, strict scrutiny is well-recognized for its use in cases regarding racial discrimination, in which it has proven itself capable of identifying when that invidious purpose is the motive for the law in question.[170] In *Adarand Constructors, Inc. v. Pefia*, a case dealing with the constitutionality of financial incentives in government contracts to subcontract with "socially disadvantaged individuals," defined to include minority individuals, the Court held that classification based on race is suspect by nature.[171] This is because racial classifications are at extreme risk of illegitimate use.[172] Unlike with the rational basis test, courts reviewing racial classifications do not blindly accept the government's stated rationales for why it needs to have racial classifications. Instead, using strict

167 *See*, *e.g.*, Adarand Constructors, Inc. v. Pefia, 515 U.S. 200, 226 (1995).

168 *See*, *e.g.*, Johnson v. California, 543 U.S. 499, 505 (2005).

169 Adarand Constructors, 515 U.S. at 226; City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493 (1989).

170 Ozan O. Varol, *Strict in Theory, but Accommodating in Fact*, 75 MO. L. REV. 1243, 1246-47 (2010).

171 Adarand Constructors, 515 U.S. at 226.

172 *Id.*

scrutiny, they make a genuine effort to ensure that the government "is pursuing a goal important enough to warrant use of a highly suspect tool" like racial classifications.[173] Strict scrutiny therefore involves a legitimate analysis of the government's motives for enacting legislation.

The rigorous analysis of the government's motives for establishing the law or regulation in question makes it an ideal tool to combat the problem of regulatory capture in occupational restrictions. Strict scrutiny involves looking to the government's actual motives for enacting a law and requires that the government bear the burden of proof.[174] It does not involve credulously accepting thin justifications for a law, accepting hypothetical rationales, or requiring plaintiffs to bear the burden of disproving every single conceivable rationale.[175] This enables it to be used to combat regulatory capture by looking directly at the protectionist motives for the restrictions in question. Occupational restrictions are frequently used for a purpose that some commentators would say is illegitimate, as a way to protect politically popular individuals and businesses from competition by preventing people from entering the market and competing with them.[176] By examining whether or not such a purpose is at play, rather than simply accepting an implausible public safety rationale as the truth, courts can hold that laws that exist purely to protect favored actors from competition are unconstitutional while still upholding laws that are genuinely motivated by legitimate health and public safety concerns. A careful examination of the government's motives could help ensure that protectionist laws that do little to protect public safety but do much to protect established practitioners from competitors can no longer stop people from working in the occupations of their choosing. Occupational liberty would have more robust protection with more rigorous judicial review.

Strict scrutiny can therefore provide protection to occupational freedom by carefully examining the government's motives for enacting restrictions on that right. Anticompetitive laws can be more readily combatted when courts examine the actual reasons behind government activity, not hypothetical ones.

C. *Addressing Concerns with Strict Scrutiny in Occupational Liberty Cases*

Applying strict scrutiny to restrictions on occupational freedom and other economic regulations is not without its objectors. Critics of this

---

173 *Id.* (quoting J.A. Croson, 488 U.S. at 493).

174 Johnson v. California, 543 U.S. 499, 505 (2005); Adarand Constructors, 515 U.S. at 226.

175 *See*, *e.g.*, Johnson, 543 U.S. at 505; Adarand Constructors, 515 U.S. at 226.

176 *See*, *e.g.*, Sensational Smiles, LLC v. Mullen, 793 F.3d 281, 285-88 (2d Cir. 2015); Meadows v. Odom, 360 F. Supp. 2d 811, 816-25 (M.D. La. 2005), *vacated as moot*, 198 Fed. Appx. 348 (5th Cir. 2005).

approach argue that it effectively revives the *Lochner* era, which is unpopular among both the political left and right, albeit for different reasons. On a related note, the strict scrutiny approach also faces criticism because of a perception that it will make it impossible for political majorities to regulate the economy.

Critics of applying strict scrutiny to economic regulations, including restrictions on occupational liberty, argue that applying that standard will lead to a revival of the *Lochner* era.[177] The *Lochner* era saw the courts strike down numerous economic regulations under the Due Process Clause.[178] The *Lochner* Court's practice of striking down economic regulations under a theory of substantive due process has attracted critics from the left and the right.[179] Both have argued that the *Lochner* Court was engaging in "judicial activism" by deciding the case based on the justices' economic views rather than the legal merits of the case.[180] On the left, criticism has also been directed at the *Lochner* Court's penchant for striking down economic regulations on the ground that this represented a small minority putting a stop to progressive reforms.[181] On the right, hostility toward *Lochner* stems from hostility to its protection of an unenumerated right through substantive due process, which many conservatives believe is an illegitimate doctrine because the Due Process Clause on its face only applies to procedure.[182] Fears of a *Lochner* revival are abundant across the political spectrum, due to claims that the justices at the time made their decisions based on their own policy preferences rather than valid constitutional principles and fear of such a revival's impact on the justices' respective constitutional and policy preferences. It is a bogeyman of constitutional law.

However, the *Lochner* era does not deserve the level of vitriol it has received over the past century. First, the criticism that the justices decided the cases of that era in order to promote their policy preferences is largely unfair. Far from being a decision completely untethered from constitutional

177 *See*, *e.g.*, Nan Aron, *The Rise of "Judicial Engagement" and the Future of the Supreme Court*, MEDIUM (June 29, 2016), https://medium.com/@nanaron/the-rise-of-judicial-engagement-and-the-future-of-the-supreme-court-8744f17317ee (arguing that the desire among libertarians and some conservatives for a more engaged judiciary with more rigorous review for economic regulations represents an attempt to return to the *Lochner* era and erase progressive economic reforms).

178 NEILY, *supra* note 13, at 41-48; ROOT, *supra* note 14, at 45-61.

179 *See*, *e.g.*, ROBERT H. BORK, THE TEMPTING OF AMERICA: THE POLITICAL SEDUCTION OF THE LAW 36-49 (1990); LAURENCE H. TRIBE & MICHAEL C. DORF, ON READING THE CONSTITUTION 66 (1991).

180 *See*, *e.g.*, Lochner v. New York, 198 U.S. 45, 74-76 (1905) (Holmes, J., dissenting); BORK, *supra* note 179, at 36-49; TRIBE & DORF, *supra* note 179, at 66.

181 *See*, *e.g.*, Aron, *supra* note 177; Ian Millhiser, *If You Want To Understand What's Happened To The Supreme Court, You Need To Listen To Rand Paul*, THINKPROGRESS (Jan. 16, 2015), https://thinkprogress.org/if-you-want-to-understand-whats-happened-to-the-supreme-court-you-need-to-listen-to-rand-paul-d2f245be1706/.

182 *See*, *e.g.*, BORK, *supra* note 179, at 31; Steven G. Calabresi, *Substantive Due Process After* Gonzales v. Carhart, 106 MICH. L. REV. 1517, 1531 (2008).

principles, *Lochner* was, in reality, well within the bounds of reasonable constitutional interpretation.[183] As mentioned above, economic rights, including occupational freedom, have a long history and tradition of protection dating back to the history of the early United States and English common law.[184] Occupational liberty in particular has a strong background of protection, with it having been recognized as a natural right since the founding era.[185] Additionally, it is unlikely that the Supreme Court decided *Lochner* in the way that it did based on hostility to economic regulations because it took a broad view of state police powers in a number of cases from that period, including cases dealing with economic regulations.[186] Based on its other decisions from around the same time, it appears that the *Lochner* Court was less motivated by a hostility to economic regulations than by a reasonable perception of what constituted a right worthy of protection under the United States Constitution. Far from striking down every regulation that crossed its path, the *Lochner* Court upheld the constitutionality of several economic regulations.

Fears from the left and right regarding a more rigorous standard of review for economic regulations are also likely overblown. Regarding the view among progressives that it will stand in the way of reforms desired by progressives and stand in the way of economic regulations, strict scrutiny for occupational regulations would not be an insurmountable obstacle for them to overcome. Although there is a perception that strict scrutiny is "strict in theory, fatal in fact," this is not the case.[187] One empirical study found that 30% of cases in which strict scrutiny was applied saw the challenged law upheld as constitutional, with the rates for individual constitutional issues ranging from 22% to 59%.[188] Strict scrutiny is far from an impenetrable barrier in the areas where it is applied today, and this would likely be true were economic rights to receive strict scrutiny as well. It is entirely conceivable that an economic regulation, including one related to occupational liberty, could be motivated by genuine health and safety concerns to the point of meeting the compelling interest requirement, and could be the least restrictive means to achieve that interest. Strict scrutiny would be effective at combatting regulatory capture by preventing protectionist laws from being

---

183 DAVID E. BERNSTEIN, REHABILITATING LOCHNER: DEFENDING INDIVIDUAL RIGHTS AGAINST PROGRESSIVE REFORM 1-39 (2011).

184 BERNSTEIN, *supra* note 183, at 14-22; SANDEFUR, *supra* note 10, at 17-29.

185 SANDEFUR, *supra* note 10, at 17-29.

186 *See*, *e.g.*, New York Central R.R. Co. v. White, 243 U.S. 188, 201 (1917) (holding that a worker's compensation law was constitutional); Wisconsin, Minnesota & Pacific R.R. v. Jacobson, 179 U.S. 287, 296 (1900) (holding that a requirement that railroad intersections have track connections was constitutional); Holden v. Hardy, 169 U.S. 366, 397-98 (1898) (holding that a limitation on the number of hours worked for miners and smelters was constitutional under a broad reading of state police powers).

187 Adam Winkler, *Fatal in Theory and Strict in Fact: An Empirical Analysis of Strict Scrutiny in the Federal Courts*, 59 VAND. L. REV. 793, 869-71 (2006).

188 *Id.* at 813-15.

upheld while still upholding laws actually motivated by health and safety concerns.

Similarly, conservative concerns are also overblown. Regarding concerns about the legitimacy of substantive due process, substantive due process has a stronger basis in history than its critics claim. Starting in the early 19th century, some state courts held that enforcing invalid laws violated due process of law, and most legal theorists recognized that due process of law had a role in protecting substantive rights.[189] The abolitionist movement also frequently invoked substantive due process in opposition to the institution of slavery, arguing that it protected inalienable rights and that slavery took away people's rights without due process of law.[190] Furthermore, even if substantive due process is illegitimate, the outcome in *Lochner* and the protection of economic rights would still be defensible under the Privileges or Immunities Clause of the Fourteenth Amendment, which based on the historical record was intended to protect fundamental rights, including economic rights, against state encroachment.[191] Conservative fears of the consequences of protection of the right to earn a living are therefore also unjustified.

Occupational freedom should receive strict scrutiny. It is a fundamental right with a strong basis in the history and tradition of the United States and it is implicit in a system of ordered liberty. Liberty and justice are irreparably harmed when occupational freedom is violated. The application of strict scrutiny would also enable economic restrictions based on favoritism and regulatory capture to be more easily combatted. Finally, fears on both the left and the right regarding the application of strict scrutiny are unfounded. Occupational liberty should receive the rigorous standard of review warranted by its fundamental nature.

IV. OTHER POTENTIAL SOLUTIONS

Should implementing strict scrutiny for the review of occupational restrictions prove too difficult, there are other options to ensure that protectionist motives hold less sway in the context of those kinds of regulations. One would be to ensure that the standard of review used when determining the constitutionality of economic restrictions is intermediate scrutiny, a standard of review that is not quite as rigorous as strict scrutiny but has more teeth than the rational basis test. Another option would be for legislatures to build in protections against protectionism and regulatory capture at the front end when creating laws on occupational restrictions.

---

189 BERNSTEIN, *supra* note 183, at 9-10.

190 *Id.* at 10-11.

191 *See*, *e.g.*, Slaughter-House Cases, 83 U.S. 36, 83-111 (1872) (Field, J., dissenting); *id.* at 111-24 (Bradley, J., dissenting); *id.* at 124-30 (Swayne, J., dissenting); NEILY, *supra* note 13, at 85-90; ROOT, *supra* note 14, at 20-39.

A. *Intermediate Scrutiny Could Address the Regulatory Capture Issue*

One potential solution that would represent a viable alternative to strict scrutiny would be to apply intermediate scrutiny to economic regulations such as those governing occupational restrictions.

Intermediate scrutiny requires that the government prove that the law under review furthers an important government interest and uses means that are substantially related to that interest in order for the law to pass constitutional muster.[192] Although this standard lacks the exactitude of strict scrutiny, it is nevertheless significantly more rigorous than the rational basis test.[193] This gives it somewhat similar advantages to strict scrutiny when compared to rational basis review, albeit not to the same degree. Intermediate scrutiny, though it lacks the same exacting analysis of strict scrutiny, does involve a genuine effort to examine the government's motives in establishing a regulatory scheme.[194] Although it does not require the government to use the least restrictive means available in determining how things should be regulated, credulous acceptance of the government's rationale for why a regulation should be in place is not a characteristic of intermediate scrutiny.[195] This rigorous review, if applied to economic regulations, would likely enable courts to identify political favoritism as a motive for statutes and regulations and give them the ability to strike them down as being insufficiently related to an important government interest. It would therefore be more capable of combatting regulatory capture than the rational basis test, since it would still involve meaningful review of the relevant laws, even if it is not quite as rigorous a standard as strict scrutiny.

Of course, intermediate scrutiny does not represent as ideal of a solution to the capture problem as strict scrutiny would. Although intermediate scrutiny has more teeth than the rational basis test, it is still a less firm standard than strict scrutiny.[196] It therefore lacks the same ability to combat regulatory capture that strict scrutiny has, even if it is not as deferential as rational basis review. This could potentially lead to protectionist regulations that would be struck down under strict scrutiny being upheld under intermediate scrutiny. Another issue with intermediate scrutiny is that, under the terms set by the Supreme Court, it is not the standard of review used to review fundamental

---

192 *See*, *e.g.*, Turner Broad. Sys. v. FCC, 520 U.S. 180, 186 (1997); Wengler v. Druggists Mut. Ins. Co., 446 U.S. 142, 150 (1980).

193 *See*, *e.g.*, Craig v. Boren, 429 U.S. 190, 199-210 (1976) (holding that a law setting different ages at which men and women could legally purchase 3.2% beer violated the Equal Protection Clause under intermediate scrutiny); NEILY, *supra* note 13, at 49-50 (2013).

194 *See* Craig, 429 U.S. at 199-210.

195 *See id.* at 199-204.

196 *See*, *e.g.*, Johnson v. California, 543 U.S. 499, 505 (2005); Craig, 429 U.S. at 199-210 (1976); NEILY, *supra* note 13, at 49-50.

rights, which receive strict scrutiny.[197] Occupational liberty is a fundamental right, with a strong grounding in American history and tradition, so strict scrutiny represents a more appropriate standard of review for restrictions on that right than intermediate scrutiny.[198] As a fundamental right, occupational liberty should be afforded the protection of strict scrutiny for the sake of consistent treatment of review of fundamental rights. Intermediate scrutiny lacks the advantage of this consistency.

However, intermediate scrutiny nonetheless represents a superior alternative to the rational basis test as the standard for reviewing the constitutionality of restrictions on occupational freedom. It involves a much more thorough examination of the government's motives behind a regulation than rational basis review, making it a potential way to combat regulatory capture were it to be applied in occupational liberty cases and other cases dealing with economic rights. While it may not represent the ideal solution to the problem of regulatory capture when compared to strict scrutiny, it is a better option than blind deference.

B. *Legislatures Could Combat Regulatory Capture on the Front End*

If the courts do not use a more rigorous level of scrutiny, whether strict or intermediate, when reviewing restrictions on occupational freedom, there are other solutions available to help mitigate regulatory capture. Legislatures can find ways to attempt to combat regulatory capture in occupational regulations by structuring their licensure laws to ensure that licensing schemes are genuinely motivated by health and safety concerns.

The state of Nebraska provides a good example of how a legislature can combat regulatory capture through the structure of its licensing laws. In April 2018, Nebraska adopted Legislative Bill 299, the Occupational Board Reform Act, which contained provisions aimed at reducing the burden of arbitrary licensure.[199] The law established a procedure for reviewing occupational restrictions that involved ensuring that the regulations are combatting "present, significant, and substantiated harms" and that the regulations use the "least restrictive means" to address these harms, with legislative standing committees reviewing one-fifth of the occupational regulations per year to determine whether they should be terminated or modified based on this

197 Troxel v. Granville, 530 U.S. 57, 64-68 (2000); Boy Scouts of Am. v. Dale, 530 U.S. 640, 658-60 (2000); Zablocki v. Redhail, 434 U.S. 374, 383-88 (1978); Bounds v. Smith, 430 U.S. 817, 827-28 (1977); Harper v. Va. State Bd. of Elections, 383 U.S. 663, 667-70 (1966); United States v. Guest, 383 U.S. 745, 757 (1966).

198 *See* SANDEFUR, *supra* note 10, at 17-24.

199 Neb. Legislature, *LB299 - Adopt the Occupational Board Reform Act and change procedures for rules and regulations*, NEB. LEGISLATURE (last visited Dec. 9, 2020), https://nebraskalegislature.gov/bills/view_bill.php?DocumentID=31200.

standard.[200] The requirements that the law be aimed at actual harm and use the least restrictive means to combat those harms bears a strong resemblance to the strict scrutiny standard.[201] L.B. 299 therefore created a standard strongly resembling strict scrutiny, only it is used by legislative standing committees when reviewing occupational restrictions for potential elimination or modification, rather than by courts during a constitutional challenge.[202] It helps mitigate the issue of regulatory capture by ensuring that the least restrictive means are used and that the licensing is actually motivated by public health and safety concerns. It manages to avoid the deference of the rational basis test by providing rigorous review through legislative committees rather than the court system.

Of course, legislative solutions face obstacles as well. It may prove to be just as difficult for legislatures to effectively combat regulatory capture as it would be to persuade courts to stop using rational basis review when reviewing economic regulations. Judges are supposed to be insulated from political pressure, which is why, on the federal level, at least, they are appointed for life conditioned upon good behavior.[203] By contrast, election and reelection are primary concerns for legislators, making them highly vulnerable to political pressures.[204] It therefore may not be wise to rely on them to resist the temptations of regulatory capture. In fact, considering that it is legislatures, rather than judges, that have the power to create "the rules by which the duties and rights of every citizen are to be regulated," the threat of regulatory capture and protectionist rules are more likely to come directly from them.[205] Leaving the task of combatting regulatory capture to legislators therefore puts the task in the hands of entities that are among the mostly likely to fall prey to its temptations, making a stricter standard of review by the courts an important safeguard against legislative abuse in this area. Strict scrutiny would provide an important check against protectionist laws by the branch of government that is actually intended to review the constitutionality of legislation.[206] Nevertheless, legislatures can play an important role in combatting occupational restrictions by reforming their licensing laws to ensure that they are truly motivated by genuine public health and safety concerns, rather than political favoritism. Such reforms are particularly important when the judiciary is unable or unwilling to provide proper review to ensure that this is the case.

Legislatures can therefore also fight regulatory capture by structuring their licensing laws so that they are designed to address actual harms rather

---

200 NEB. REV. STAT. §§ 84-933 to 84-948 (2019).

201 *See id.* § 84-946 (2019); Johnson, 543 U.S. at 505.

202 *See* NEB. REV. STAT. § 84-946 (2019); Johnson, 543 U.S. at 505.

203 *See* THE FEDERALIST NO. 78 (Alexander Hamilton), https://avalon.law.yale.edu/18th_century/fed78.asp.

204 *See* EAMONN BUTLER, PUBLIC CHOICE – A PRIMER 28-29 (2012); DAVID R. MAYHEW, CONGRESS: THE ELECTORAL CONNECTION 5-27 (2d ed. 2004).

205 *See* THE FEDERALIST NO. 78, *supra* note 201.

206 *See id.*

than protect favored interest groups from competitors. Although relying on such an approach would leave the entities most likely to fall prey to regulatory capture responsible for defeating it, it does represent a way to limit the rise of protectionist laws in an era in which economic restrictions receive an extremely deferential standard of review from the courts.

CONCLUSION

Because rational basis review is so deferential to the government, it is unable to effectively stop legislatures from crafting economic regulations aimed at granting political favors to special interests rather than requiring that laws combat actual health and safety concerns. The rational basis test requires courts to accept justifications for the law based on hypotheticals rather than empirical evidence, requires plaintiffs to negate every conceivable state of facts that could justify the law, a task so herculean that it borders on impossibility, and in some federal circuit courts requires courts to come up with justifications for the government. This leads to courts acting as little more than a rubber stamp for whatever laws and regulations governments wish to pass, even if they are blatantly motivated by cronyism instead of a desire to protect the public. This leads to the fundamental right of being able to work in the occupation of one's choice without undue government interference being violated in the name of benefiting political favorites, who often do not want to face potential competitors on an even playing field in the marketplace.

To mitigate this issue, courts should instead review occupational restrictions under a strict scrutiny standard of review. This would be more appropriate for two reasons. First, fundamental rights are required to receive strict scrutiny, and occupational liberty is a fundamental right as the term is defined by the courts. It is deeply rooted in the history and tradition of the United States, dating back to before the Founding era. It is also fundamental to a system of ordered liberty, to the point liberty and justice would be lost if they were infringed, because it is essential to an ordered pursuit of happiness. A strict scrutiny standard would also involve meaningful review of the restrictions in question, meaning that blatantly protectionist laws would not be upheld as they are under the rational basis test. Failing that, an intermediate scrutiny standard or efforts by legislatures to ensure that laws are motivated by health and safety instead of protectionism would also be a way to combat regulatory capture. Whatever solution is chosen, however, it is crucial to ensure that occupational liberty receive the level of protection it deserves. People should be able to work in the field of their choice without fear of obstacles designed to reduce competition and grant political favors.

File: Carmin 17.1 To Publish_Edited Created on: 4/19/2022 10:31:00 AM Last Printed: 6/14/2022 9:28:00 PM

BETTER THAN NOAH'S ARK: FLOOD INSURANCE THAT WORKS

Jake Carmin

INTRODUCTION

In the figurative wake of Hurricanes Harvey, Irma, and Maria, Congress quietly forgave $16,000,000,000 owed to the Department of the Treasury by the National Flood Insurance Program (NFIP).[1] The move, while expensive, was made necessary by $8,700,000,000 in federal flood insurance payouts, which would have pushed the NFIP past its borrowing limits.[2] Such a scenario, to emergency managers, was hardly surprising: Craig Fugate, former Director of the Federal Emergency Management Agency under President Obama, told CNN in 2018 that keeping the NFIP financially soluble under the current structure was "damn near impossible."[3]

Soon after Hurricane Florence a year later, FEMA Administrator Brock Long called for changes in the program, telling the press that the public's failure to carry insurance and/or live in risk-free areas contributed to the high cost of recent disasters.[4] "If you want to live in these areas, you've got to do it in a more resilient fashion," Long said, adding that when "you see this enough in your career, you get ticked off."[5] The assessment behind Long's irritation is largely on the mark. States affected by Hurricane Florence had relatively low flood insurance market penetration rates; even then, most insurance policies were held on the coast, leaving inland flood victims uninsured and thus dependent on disaster relief.[6] More generally, few argue with the premise that on the national level, geographical housing choice is suboptimal with respect to disaster risk. Cities are built on fault lines, on plains famous for tornadoes, and most relevant for purposes of this paper, in

---

1 Chris Isidore, *Hurricane Florence Is The Latest Setback To Struggling Flood Insurance Program,* CNN (Sep. 14, 2018, 12:41pm), https://www.cnn.com/2018/09/14/politics/hurricane-florence-national-flood-insurance/index.html.

2 *Id.*

3 *Id.*

4 Christopher Flavelle, *FEMA Head 'Ticked Off' Over Cycle of Inadequate Storm Preparation, Insurance, Evacuations*, INSURANCE JOURNAL (Oct. 15, 2018), https://www.insurancejournal.com/news/national/2018/10/15/504455.htm.

5 *Id.*

6 Gloria Gonzalez, *Florence Drenches the Carolinas in a Test of Insurance Policies*, BUSINESS INSURANCE (Sept. 18, 2018, 7:00am), https://www.businessinsurance.com/article/20180918/NEWS06/912324043/Florence-hurricane-tropical-depression-drenches-the-Carolinas-in-test-of-flood-i.

floodplains and along coastlines. If this were not the case, the costs of catastrophes in the United States would be significantly lower for federal, state, and local governments.

The hurricanes of 2018 further highlighted what many in the emergency management community already knew: America has a flood resilience deficiency. This deficiency is not the fault of any single entity. The NFIP suffers from severe under-enrollment, premiums that do not cover payouts, and covering expensive properties at an astronomically high risk of repeat loss; Congress struggles to either budget for disaster relief or reset the federal insurance market while climate change continually increases the cost of flood disasters; and communities neither consider flood risk when developing nor retreat when a flood strikes.

The importance of reform is only increasing; climate change[7] and the ever-growing federal deficit[8] mean that the status quo is neither sustainable nor likely to remain relatively apolitical. But policymakers have struggled to find politically feasible reforms. After one legislative attempt to make the NFIP financially sustainable led to public discontent and housing market turmoil, Congress quickly repealed many of the provisions.[9]

Regulatory reform is even less popular, even in the face of obvious risk; after FEMA announced an update to Maine's Flood Insurance Risk Maps (FIRMs), which inform federal insurance premiums, one coastal homeowner angrily told a television news station that her community, named "The Village by the Sea," was not at risk of flooding.[10] During the interview, the ocean could be seen a few hundred yards from her neighborhood.[11]

Given the myriad of problems with the NFIP, one might be forgiven for concluding that the program should simply be scrapped. However, as this comment will discuss below, the NFIP was a vast improvement over the flood relief scheme of the first half of the twentieth century. Congress has intentionally and correctly preferred subsidized flood insurance to pure *ad hoc* disaster relief. Even still, federal policy is far from optimal. Improving it is possible – but doing so requires an understanding of historical disaster response policy, behavioral economics, congressional budgeting, and more.

---

7 Jennifer Wriggins, *Flood Money: The Challenge of U.S. Flood Insurance Reform in a Warming World*, 119 PENN ST. L. REV. 361, 424 (2014).

8 *See* ALEX GRAY, WORLD ECONOMIC FORUM, THIS IS HOW MUCH DEBT YOUR COUNTRY HAS PER PERSON (Oct. 2017) (citing OECD report placing per capita debt in the United States at $61,539); ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, DATA: GROSS NATIONAL INCOME (2017), https://data.oecd.org/natincome/gross-national-income.htm (measuring gross national income at $43,981).

9 Stephen G. Fier, et. al., *The State of the National Flood Insurance Program: Treading Water or Sinking Fast?*, 33 J. INS. REG. 115, 134 (2014).

10 WMTW-TV, *More Maine towns object to new FEMA flood zone*, YouTube (Jun. 2, 2014), https://www.youtube.com/watch?v=5A53a6ioHwY.

11 *Id.*

I. BACKGROUND

A. *History of Disaster Relief and the Passage of the NFIP*

Disaster relief in American history can trace its roots back to 1803, when Congress passed a bill funding fire recovery efforts in New Hampshire.[12] Well into the next century, Congress retained this 'responsive' or *ad hoc* model, granting disaster recovery funds to affected communities in the aftermath of natural disasters.[13] However, for a long while, many Americans did not view disaster relief as a federal responsibility; Grover Cleveland vetoed a drought relief bill, arguing that there was no such power in the Constitution and that "Federal aid in [cases of misfortune] encourages the expectation of paternal care on the part of the Government and weakens the sturdiness of our national character. . ."[14] When the federal government did not step in, disaster relief was largely conducted by states or by private parties like the Red Cross.[15]

In 1927, the Mississippi Flood, the result of a near 15 inches of rain over 18 hours, flooded nearly a million homes and left hundreds of thousands in camps.[16] Congress only spent $10,000,000 in relief and reconstruction efforts, letting the Red Cross shoulder most of the burden,[17] despite that the total flood damage that year totaled nearly $700,000,000.[18] The severity of this flood led insurers to withdraw almost completely from the flood insurance market.[19]

Between 1927 and 1950, Congress mostly focused spending on flood control and continued to apportion disaster relief *ad hoc*.[20] In 1950, the Disaster Relief Act allowed the President discretion in using a permanent relief fund to help disaster victims.[21] For the next decade, Presidents Truman and

---

12 *About the Agency*, FEMA, https://www.fema.gov/about-agency (last visited Nov.15, 2018). This is widely considered the first disaster relief bill in American history.

13 *Id.*

14 David Moss, *Courting Disaster? The Transformation of Federal Disaster Policy since 1803*, 1999 THE FINANCING OF CATASTROPHE RISK 307, 313 (Kenneth A. Froot ed.) (Clara Barton, founder of the Red Cross and largely responsible for its designation as the disaster relief agent of the federal government, thought local resources could handle the drought, and supported Cleveland).

15 *Id.*

16 JOHN M. BARRY, RISING TIDE: THE GREAT MISSISSIPPI FLOOD OF 1927 AND HOW IT CHANGED AMERICA, 1998.

17 Moss, *supra* note 14.

18 H.R. Doc. No. 89-465, at 4 (1966).

19 Scott Gabriel Knowles & Howard Kunreuther, *Troubled Waters: The National Flood Insurance Program in Historical Perspective*, 26 J. OF POLICY HISTORY 328, 332 (2014).

20 Moss, *supra* note 14, at 314–15.

21 *Id.* at 315. While the Disaster Relief Act was not the first general disaster relief bill—a bill in 1947 allowed excess government equipment to be given to localities to repair roads and bridges—it was

Eisenhower repeatedly called for a national flood insurance program to no avail.[22] Congress widely believed that the moral hazard attached to insurance might entice more development in floodplains.

After Hurricane Betsy in 1965, President Johnson lamented that Congress had signed six relief bills in eighteen months; in response, Congress created the Task Force on Federal Flood Policy, which recommended that the federal government undergo a massive attempt to map floodplains and account for communities with flood problems.[23] This task force concluded that extensive spending on flood control was unsustainable and inequitable:

> [Individual beneficiaries of flood control failing to bear an adequate share of the costs], combined with the bias in favor of river control alternatives, has relieved many individual flood plain occupants of responsibility, in a fiscal sense, for the consequences of their actions. Under existing policies flood plain property owners in unprotected areas may bear only a portion of the cost, their price being exacted when damage occurs. Some shoulder the full losses; others rely on public relief and assistance in rehabilitation. No matter how serious their encroachment on the watercourse, the occupants bear few of the costs resulting from encroachment. They bear a minor fraction, through payment of general taxes, of the public cost of relief and rehabilitation. The general public, by bearing all or a major part of the cost of flood protection works and lessening the individuals' damage costs, further subsidizes their use of the flood plain. Principles of economic efficiency and social equity thereby are violated.[24]

The task force reasoned that with enough effort, a flood insurance program could successfully manage floodplains.[25] However, it warned that if the program was not expertly designed, then insurance should not be used at all.[26] Specifically, it declared that "For the federal government to subsidize low premium disaster insurance or provide insurance in which premiums are not proportionate to risk would be to invite economic waste of a great magnitude."[27]

Based on this research, the NFIP was established in the 1968 National Flood Insurance Act (NFIA) with the stated goal to "provide flexibility in the program so that such flood insurance may be based on workable methods of pooling risks, minimizing costs, and distributing burdens equitably among those who will be protected by flood insurance and the general public."[28]

---

the most revolutionary. Giving the President discretion allowed for timelier disaster relief and was successful enough that the current disaster relief scheme operates fairly similarly.

22 Knowles & Kunreuther, *supra* note 19.

23 *Id.* at 332-33.

24 TASK FORCE ON FED. CONTROL POL'Y, A UNIFIED NAT'L PROGRAM FOR MANAGING FLOOD LOSSES, H.R. DOC. NO. 89-465, at 15 (2d Sess.1966). Such an argument can easily be made about general disaster relief as well; that is, floodplain homeowners only bearing a portion of the response and recovery cost amounts to a subsidy for their high-risk choices.

25 Knowles & Kunreuther, *supra* note 19, at 333.

26 *Id.*

27 H.R. DOC. NO. 89-465, at 17. That same year, President Johnson directed the federal government to identify which of their properties were located in floodplains and to take steps to mitigate flood damage.

28 42 U.S.C. § 4001(d).

Notwithstanding the Task Force's warning, the law was designed with heavy subsidies to entice individuals into the market, and previous flood history had no effect upon a homeowner's premiums.

The law measured flood risk at the community level, conditioning eligibility for NFIP policies on whether the area had been mapped by the U.S. Army Corps of Engineers, whether actuarial rates had been calculated by Housing and Urban Development (HUD),[29] and whether a locality had adopted certain zoning and building codes.[30] While a clause in the law reduced an uninsured party's disaster relief by the amount they could have insured their property for, many communities chose to forego insurance rather than curb construction and expansion.[31]

After low early enrollment, Congress passed the Flood Protection Act in 1973, which required federally-backed mortgage holders to take out a flood insurance policy.[32] While this change gave the NFIP important power, it was later revealed that the program was severely behind on its study of flood-prone communities and had difficulty ensuring localities were enacting appropriate zoning laws.[33]

Six years later, the NFIP was moved from HUD to the newly created FEMA, whose goals more closely reflected those of flood insurance generally.[34] FEMA shifted the program to a Write-Your-Own model, where private companies sold the federally-backed flood insurance policies in exchange for an expense allowance; soon after this change, the premiums paid for the costs of the NFIP for the first time since its inception.[35]

The 1988 Stafford Act not only allowed FEMA to "buy-out" floodplain properties and pay homeowners to rebuild elsewhere, but it also reformed the general disaster relief scheme created in 1950.[36] The Stafford Act created what remains the disaster relief scheme today: a state's governor may submit a request to the President on behalf of a community overwhelmed by disaster response and recovery, and if deemed necessary, the President may declare a federal disaster, allowing FEMA to use the Disaster Relief Fund (DRF) to supplement the communities' efforts.[37] Section 404 of the act also gives FEMA the authority for the Hazard Mitigation Grant Program, which allows

29 Knowles & Kunreuther, *supra* note 19, at 336.
30 Moss, *supra* note 14, at 319.
31 Knowles & Kunreuther, *supra* note 19, at 336.
32 *Id.* at 337.
33 *Id.* at 339.
34 *Id.*
35 *Id.* at 340.
36 *Id.*
37 BRUCE R. LINDSAY & JUSTIN MURRAY, SUPPLEMENTAL APPROPRIATIONS FOR DISASTER ASSISTANCE: SUMMARY DATA AND ANALYSIS, CRS REP. R43665, 1-2 (Oct. 1, 2014).

for DRF funds to be spent in mitigation during reconstruction immediately following a disaster.[38]

In 1992, Hurricane Andrew became, at that point, the costliest disaster in American history, exhausting private insurance firms' reserves and leading to a serious reduction in coverage.[39] With the NFIP now even more vital to homeowners, Congress attempted in 1994 to increase participation by denying federal disaster relief to those in high-risk areas who had not taken out a flood insurance policy.[40]

Disaster relief changed across the board following terror attacks on 9/11, when President Bush placed FEMA, formerly a cabinet agency, underneath the Department of Homeland Security.[41] The agency's focus shifted towards terrorism preparedness and response,[42] leaving serious gaps[43] in its ability to respond to Hurricane Katrina two years later.[44] Katrina's severity left the NFIP deeply in debt, but the program was bailed out without much resistance by Congress.[45]

In 2012, Congress tried to take serious steps to reduce the burden on taxpayers caused by the NFIP's financial woes. The Biggert-Waters Flood Insurance Reform Act (Biggert-Waters), if given a longer tenure, would have phased out premiums on repeat loss properties, second homes, business properties, and damaged or improved homes.[46] While the Act also required new policies to be actuarially sound, it did not do the same for homeowners who had already taken out a policy, creating a distortion in the housing market.[47] The law received much public resistance upon its implementation, as some policyholders saw premium price increases as dramatic as three-thousand percent.[48]

To address the problems created by Biggert-Waters, Congress passed the Homeowner Flood Insurance Affordability Act of 2014 (HFIA), which scaled back premium increases on properties built on property not included on FIRMs and allowed new homeowners to obtain the older subsidized

38 TANVEER ISLAM & JEFFREY RYAN, HAZARD MITIGATION IN EMERGENCY MANAGEMENT 71, Elsevier Inc., 2016.

39 Moss, *supra* note 14, at 338.

40 *Id.* at 340-41.

41 History of FEMA, FEMA https://www.fema.gov/about/history (last visited Jan. 12, 2021).

42 Knowles & Kunreuther, *supra* note 19, at 341.

43 Whether or not FEMA can succeed in an "all-hazards" approach which includes terrorism is an extensive debate which some argue precludes FEMA from effectively responding to disaster. At minimum, it is clear that this shift severely diminished FEMA's ability to react to Katrina.

44 Knowles & Kunreuther, *supra* note 19, at 341.

45 *Id*. at 346.

46 Fier et al., *supra* note 9, at 132.

47 *Id.*

48 *Id.* at 116.

File: Carmin 17.1 To Publish_Edited Created on: 4/19/2022 10:31:00 AM Last Printed: 6/14/2022 9:28:00 PM

rates.[49] The HFIA Act did, however, require that actuarially fair premiums be charged to severe repetitive loss properties, second homes, and businesses.

The effort to move the NFIP to actuarially fair premiums, while only partially successful, was a step in the right direction for both the program's financial health and for private insurers' ability to enter the market, but it did not solve the problem.[50] As discussed above, the chain of disasters experienced in 2018 not only led to yet another Congressional bailout of the NFIP, but also indicated that the United States has a clear problem with resiliency and risk management.

B. *The NFIP's Operation*

As the law stands today, the NFIP is a largely voluntary, federally subsidized insurance program. The program relies on Flood Insurance Rate Maps, which compile hydrological, geological, and limited historical data to establish risk exposures.[51] The Special Hazard Flood Areas (SFHAs) indicate that an area is located at or below what is commonly referred to as the '100-year floodplain' but is more accurately referred to as Base Flood Elevation (BFE).[52] Somewhat confusingly, SFHAs do not experience one flood per one hundred years, but instead indicate that a property at its elevation has, on average, a one percent chance of being flooded in a given year; theoretically, this might mean a property could flood once every ten years—or more.[53]

Because SFHAs indicate that a property has a substantial risk of flooding, properties with federally-backed mortgages located in these zones are contractually required to possess flood insurance.[54] Within the SFHA, the FIRM also delineates properties to more accurately individualize their risk based on a few factors: whether an area's BFE is known, whether the property is located near levees or other flood control structures, and whether the property is in danger of "velocity hazard"—essentially, vulnerable to wave surges.[55] Outside the SFHA, flood risk is measured by proximity to drainage and whether a property is inside or outside the 500-year floodplain (or, whether it has above or below a 0.02% chance of being flooded in a given year).[56]

---

49 *Id.* at 134.

50 *Id.* at 136.

51 FIRMs historical data is limited first by the fact that little meteorological or hydrological data was compiled prior to the twentieth century. It is further limited by the fact that FEMA chose not to examine some properties' flood claim histories.

52 ISLAM & RYAN, *supra* note 38, at 77.

53 *Id.* at 76-77.

54 Knowles & Kunreuther, *supra* note 19, at 337.

55 ISLAM & RYAN, *supra* note 38, at 78.

56 *Id*.

If a property-owner believes the FIRM has incorrectly designated their property either due to elevation, mitigation efforts, or proximity to waves, they may submit their own technical data to FEMA and request a Letter of Map Revision (LOMR).[57] This appeals process has created somewhat of a cottage industry for hydrologists and engineers hired by wealthy coastal residents looking for cheaper premiums.[58] The severe disparity between various flood zone categories likely has contributed to this demand; one such firm estimates that a LOMR amending a property from "Zone V" to "Zone A" (that is, at risk of wave impact) could reduce premiums by as much as ninety percent.[59]

The NFIP's most celebrated innovation is the Community Rating System (CRS). The CRS assigns a community on a point system based on the actions it takes to mitigate its flood vulnerability.[60] There are eighteen creditable activities that fall into four categories: Flood Preparedness, Flood Damage Reduction, Mapping and Regulations, and Public Information.[61] Communities using the CRS can earn their residents within the SFHA premium discounts of up to forty-five percent.[62]

C. *NFIP Deficiencies*

Under-enrollment has plagued the NFIP since its inception,[63] and is commonly cited as a rationale for leaving the subsidized premiums in place. If people will not purchase insurance at subsidized rates, the argument goes, why would they purchase it at actuarially fair (non-subsidized) rates? This problem prevents the NFIP from using the common insurance practice of offsetting high-risk properties with numerous low-risk properties.

1. FIRMs are Outdated, Unreliable, and Undermine the NFIP

While FEMA is required to "assess the need to revise and update" all flood risk zones,[64] in practice, maps dating back to the earliest years of the

---

57 44 CFR §§ 65.3, 65.12–14, 72.1.

58 *See, e.g.*, Bill Dedman, *Meet the Flood Insurance 'Robin Hood' Who Saves Condo Owners Millions*, NBC NEWS (Feb. 19, 2014), https://www.nbc.news.com/news/investigations/meet-flood-insurance-robin-hood-who-saves-condo-owners-millions-n26711.

59 *Basics of LOMRs*, FLOOD ZONE REVISIONS, LLC (accessed Nov. 14, 2018), http://floodzone-revisions.com/basics-of-lomrs/.

60 ISLAM & RYAN, *supra* note 38, at 79.

61 *Id*. at 78

62 *Id*. at 79 (non-SFHA residents may earn between five and ten percent reductions).

63 *See* Knowles & Kunreuther, *supra* note 19, at 343.

64 42 USC § 4101(e).

NFIP still inform premiums—often inaccurately.[65] Nearly fifteen percent of counties participating in the NFIP are informed by maps that are between fifteen and forty years old.[66] But even newer maps imperfectly represent risk, neglecting to account for factors such as population growth and rapid rain accumulation.[67] FEMA often prioritizes remapping in counties that have recently experienced severe flooding events, sometimes grading next-door neighbors on opposite sides of county lines at radically different flood risk levels.[68]

A Department of Homeland Security Inspector General report estimated that merely forty-two percent of flood maps met FEMA's own standards for evaluating flood risk.[69] This failure is at least partly the fault of problems with FEMA's management of the FIRMs. FEMA's subdivision, the Federal Insurance and Mitigation Administration (FIMA) has struggled to track the allocation and status of funds directed towards map updates and has defects in its data verification procedures.[70]

A potentially more severe flaw than mismanagement is flood maps' reliance on the Base Flood Elevation benchmark. First, the hydrological science behind identifying floodplains must extrapolate so much aggregated data that it can, at best, be considered an estimate.[71] Floodplain boundary estimates are, by and large, under-inclusive; one study showed nearly 25% of flood insurance claims from 1999 to 2009 were filed outside floodplains.[72] The somewhat binary classification of properties as either in or out of a floodplain likely creates a barrier to homeowners' awareness of their own flood risk.[73] Worsening this effect is the confusing nature of the "100-year Flood" label—which in no way means that a flood event is a once-in-a-century occurrence—that may impute a false sense of security.[74]

However, Congress is also in part responsible for the maps' inadequacy. First, FEMA's claims that its funding is insufficient to meet its mapping goals[75] are likely true given that map funding was halved between 2010 and

65 Michael Keller et al., *Outdated and Unreliable: FEMA's Faulty Flood Maps Put Homeowners at Risk*, BLOOMBERG (Oct. 6, 2017), https://www.bloomberg.com/graphics/2017-fema-faulty-flood-maps/.

66 *Id*.

67 *Id*.

68 *Id*.

69 DEPARTMENT OF HOMELAND SECURITY, OFFICE OF THE INSPECTOR GENERAL, OIG-17-110, FEMA NEEDS TO IMPROVE MANAGEMENT OF ITS FLOOD MAPPING PROGRAMS, (2017).

70 *Id*.

71 Maggie Koerth-Baker, *It's Time to Ditch the Concept of 100-Year Floods*, FIVETHIRTYEIGHT (Aug. 30, 2017, 4:33 PM), https://fivethirtyeight.com/features/its-time-to-ditch-the-concept-of-100-year-floods/.

72 Wesley E. Highfield et al., *Examining the 100-Year Floodplain as a Metric of Risk, Loss, and Household Adjustment*, 33 RISK ANALYSIS 186, 189 (2013).

73 *Id.* at 187.

74 Koerth-Baker, *supra* note 71.

75 Keller, *supra* note 65.

2013.[76] Congress' recent practice of cycling short-term reauthorizations and NFIP bailouts leaves mapping funds untouched[77] and projects like building future development into flood models on the back burner.[78] Individual members of Congress further frustrate map modernization efforts by intervening on behalf of constituents who are dissatisfied with premium price increases.[79] While this political practice may ease constituent communities' financial burdens, it often leads to the reuse of old maps rather than a compromise incorporating new, more accurate risk data. [80]

Likewise, communities and residents are often unhelpful to the map modernization process. While more detailed and accurate maps would give communities a publicly available tool for more efficient and safe land development, local governments may choose to prioritize limiting premium increases and development obstacles.[81] While in many cases, communities must fight FEMA because the maps are inaccurate,[82] many local residents and governments fight new maps even in the face of obvious flood risk.[83] And the high cost of appealing these maps means that regardless of map accuracy, wealthy locales can fight to keep cheaper rates while poorer ones are largely at the mercy of FEMA.[84]

II. ANALYSIS

A. *Insurance Coverage is the Preferable Way to Manage Risk*

The preceding sections detailed the NFIP's history in the context of disaster relief, the mechanics of the program itself, and the obstacles it faces if

76 Theodoric Meyer, *Using Outdated Data, FEMA is Wrongly Placing Homeowners in Flood Zones*, PROPUBLICA (July 18, 2013, 1:07 PM), https://www.propublica.org/article/using-outdated-data-fema-is-wrongly-placing-homeowners-in-flood-zones.

77 *Id*.

78 *See* FEDERAL EMERGENCY MANAGEMENT AGENCY, FEMA REPORT TO CONGRESS ON THE TECHNICAL MAPPING ADVISORY COUNCIL RECOMMENDATIONS FROM 2015 (2016).

79 *See* Sydney Glenn, *Woman Fighting Proposed FEMA Flood Maps Devastated Same Company Will Be Used Again*, ABC NEWS (Apr. 4, 2018), https://wpde.com/news/local/woman-fighting-proposed-fema-flood-maps-devastated-same-company-will-be-used-again; SARAH PRALLE, *Drawing Lines: FEMA and the Politics of Mapping Flood Zones* 21 (2017), https://www.maxwell.syr.edu/uploadedFiles/faculty/psc/Pralle_Drawing%20Lines_APSA2017.pdf (describing Senators Chuck Schumer and Kirstin Gillibrand's role in halting map updates in Syracuse).

80 PRALLE, *supra* note 79.

81 Knowles & Kunreuther, *supra* note 19, at 337.

82 Meyer, *supra* note 76.

83 WMTW-TV, *supra* note 10.

84 *See* Dedman, *supra* note 58; Miranda Leitsinger, *For Average Joes, Fighting FEMA Flood Maps Isn't Easy or Cheap*, NBC NEWS (Feb. 19, 2014), https://www.nbcnews.com/news/us-news/average-joes-fighting-fema-flood-maps-isnt-easy-or-cheap-n23871.

it is to achieve its goals. This section will instead make the case that, despite the many financial failures of the NFIP, flood insurance has been an improvement over the previous emergency management paradigm, and is vital to Congress and emergency managers' shared goal: reducing the costs of disaster recovery in the United States.

1. *Ad hoc* Disaster Relief Fails to Achieve Flood Response Goals

As discussed above, disaster relief in the United States has largely been dispersed on an *ad hoc* basis. The Stafford Act governs how such payments are given to states today,[85] though they are funded by Congress via the Disaster Relief Fund.[86]

The Disaster Relief Fund is an annually funded account that draws partial funding from unused funds from the previous year.[87] When a disaster strikes, states whose response and recovery efforts are overwhelmed can request federal aid; this aid is not available unless the President issues a federal disaster declaration.[88] If the DRF balance is depleted by a disaster, Congress replenishes the funds with supplemental appropriations, which can happen with or without prompting from the President.[89]

There are a few reasons why supplemental appropriations are inherent in *ad hoc* disaster relief payments. First, because flood hazards' frequency, timetable, and cost are less certain than many other federal expenditures, budgeting for such costs is no precise science. And Congress, lacking expertise in meteorology, hydrology, and actuarial science, is unlikely to be successful even if disaster budgeting were its first priority.

Second, emergency supplemental appropriations have serious institutional benefits to members of Congress. While a responsible Congress might attempt to budget as best as possible for anticipated disaster relief costs, in practice Congress frees up funds for other spending projects by allowing the DRF to dip dangerously low. Individual members of Congress can then claim more credit in a time of crisis by supporting the appropriation than they could by supporting DRF funding prescriptively.

Since supplemental appropriations are, by definition, not budgeted for, such expenditures must either be offset by spending cuts elsewhere or add directly to the federal deficit.[90] While this may not be as pressing a problem as disaster relief, a deficit that is too large can prevent the government from

85 42 U.S.C. § 5121 *et seq*.
86 LINDSAY & MURRAY, *supra* note 37.
87 *Id*.
88 *Id*.
89 *Id*.
90 *Id*. at 3.

effectively combating economic recessions or dealing with the rising costs of an aging population.[91]

2. Insurance is the correct tool and is preferable to DRF spending.

Insurance is "one of humanity's greatest inventions."[92] When it works well, it protects people, their livelihoods, and the fruits of their labor, while creating a value-creating enterprise.[93] Insurance, with its ability to coordinate risk sharing, protect investments, and incentivize safe behavior, can be more effective in promoting self-sustainability than foreign aid in developing countries.[94] As discussed above, its power to incentivize mitigation in the disaster context leads to less societal loss and thus safer, cheaper societies.

Insurance draws its power in part from statistical probability and multiplicative power; as an insurance pool adds more policyholders, the probability that the pool will experience maximum possible loss decreases relative to the probability that each individual experiences maximum probable loss.[95] In perfect conditions, i.e. when consumers and providers know policyholders' risk exposure and when losses are independent of each other, insurance companies can charge premiums equal to the probability of losses multiplied by the value of those losses.[96] In this scenario, insurance companies can afford to reimburse all policyholders without incurring losses.[97] Consumers who are risk averse, and thus willing to pay more than the expected value of their policy, are willing to foot the administrative costs of managing an insurance fund or the costs of building up a pool's reserve funds.[98]

---

91 Heather Long, *Why America's Return to Trillion Dollar Deficits is a Big Problem for You*, WASHINGTON POST: WONKBLOG (Apr. 9, 2018), https://www.washingtonpost.com/news/wonk/wp/2018/04/09/why-americas-return-to-1-trillion-deficits-is-a-big-problem-for-you/?utm_term=.f2e4f5d84f60.

92 HOWARD C. KUNREUTHER ET AL., INSURANCE AND BEHAVIORAL ECONOMICS: IMPROVING DECISIONS IN THE MOST MISUNDERSTOOD INDUSTRY 13 (2013); *see also* Orin S. Kerr, *A Theory of Law*, 16 GREEN BAG 2D 111 (2012).

93 HOWARD C. KUNREUTHER ET AL., *supra* note 92.

94 *See generally* Innovations for Poverty Action, Examining Underinvestment in Agriculture: Returns to Capital and Insurance Among Farmers in Ghana (2012), https://www.poverty-action.org/study/examining-underinvestment-agriculture-returns-capital-and-insurance-among-farmers-ghana; *Planet Money: Episode 723: The Risk Farmers*, NATIONAL PUBLIC RADIO (Sept. 7, 2016), available at https://www.npr.org/templates/transcript/transcript.php?storyId=492988779.

95 HOWARD C. KUNREUTHER ET AL., *supra* note 92, at 21-22 (posing a hypothetical in which two people each have a fifty percent chance of $100 loss. The probability that they individually will experience their worst outcome is fifty percent. In contrast, when risks are pooled, the probability that the pool will experience the worst outcome – both policyholders losing fifty dollars – is only twenty-five percent).

96 *Id*. at 21.

97 *Id*.

98 *Id*. at 27.

File: Carmin 17.1 To Publish_Edited Created on: 4/19/2022 10:31:00 AM Last Printed: 6/14/2022 9:28:00 PM

Without serious barriers to entry, insurance costs should be close to actuarially fair premiums (that is, premiums that only reflect risk and do not include administrative and reserve costs) because buyers with perfect information, regardless of risk-aversion, will choose the cheapest option that covers their risk.[99]

3. Insurance can Promote Effective Mitigation

Of all emergency management functions, mitigation is considered to be the most overlooked and underutilized—and the function most capable of preventing loss of life and property. One study of twenty-three years of federal mitigation grants estimated that for every one dollar spent on federal mitigation grants, society saves six dollars in recovery costs.[100] The implementation of certain disaster mitigation plans at the federal level is estimated to prevent up to six hundred fatalities and nearly a million injuries.[101]

While it is widely accepted that such plans are highly cost-effective, mitigation is under-implemented. This in part is because unlike preparatory or response efforts, mitigation is not event-driven, but instead requires sustained attempts from diverse stakeholders to make a community better able to withstand disasters.[102] As a result, incentivizing homeowners and communities to mitigate is rarely politically feasible; emergency managers have to sell a policy whose return on investment may never be realized and may not be perceived anyway.[103] One state emergency manager is said to have remarked that "I will never lose my job if I fail to do mitigation, but I could lose my job if I mess up a response."[104]

Nevertheless, mitigation has substantial benefits both to individuals and society. Disaster relief is expensive to federal taxpayers and often inadequate at fully compensating home and business owners. Mitigation is an ideal solution because it both reduces the overall cost of disaster relief and leaves individuals with less severe damage.

While disaster insurance might be considered a recovery tool because its payouts may be used for rebuilding, some emergency management

---

99 *Id*. at 26.

100 NATIONAL INSTITUTE OF BUILDING SCIENCES, *Natural Hazard Mitigation Saves: 2017 Report* (2017).

101 *Id.*

102 GEORGE D. HADDOW ET AL., INTRODUCTION TO EMERGENCY MANAGEMENT, 76-82, (6th ed. 2017).

103 Mitigation and preparation are two distinct functions in the emergency management cycle. For example, placing sandbags around a house in anticipation of an oncoming hurricane would be a preparatory emergency management function, while raising a home's foundation above the 100-year floodplain would be a mitigatory one.

104 HADDOW, *supra* note 102.

scholars argue it is more aptly characterized as a mitigatory one.[105] Mitigation's primary focus is on improving community resilience, and insurance payouts post-disaster help reduce the number of people reliant on government aid. Such arrangements mean that communities are not as easily overwhelmed and individuals are less likely to find themselves with no serious recovery option.

Insurance is not just a mitigatory tool—it's the ideal method for incentivizing mitigation itself.[106] In the same way the threat of higher health or automobile insurance premiums might encourage someone to speed less or to quit smoking, in an ideal world, disaster insurance incentivizes risk avoidance in location choice or construction method.

Insurers, who seek to reduce their own risk exposure, have an interest in advising their insureds on mitigation practices as well.[107] In the early nineteenth century, factory mutuals (companies owned jointly by the factories they insured) circumvented moral hazard by establishing safety standards and conditioning insurance on compliance with those standards.[108]

Thus, the NFIP, in theory if not in reality, is an incredible opportunity for incentivizing the exact type of mitigation that is under-implemented and key to combating the rising pressure on fiscal constraints caused by disasters. However, its failure to accurately evaluate risk, and therefore to accurately set prices on premiums, fails to take advantage of such an opportunity.

B. *How to Optimize Flood Insurance*

1. Calls for Private Insurance

Congress has recognized that the NFIP's subsidized premiums do more than leave taxpayers liable for flood damage—they also crowd out private insurers who wish to enter the market.[109] The House Committee on Financial Services said in 2016 that the lack of a private insurance option "expos[es] taxpayers to virtually all of the Nation's flood risk . . . . [T]he Committee believes the biggest impediment to the development of a private insurance

105 *Id*. at 76-77.

106 Omri Ben-Shahar & Kyle D. Logue, *The Perverse Effects of Subsidized Weather Insurance*, 68 STAN. L. REV. 571, 587 (2016).

107 John Rappaport, *How Private Insurers Regulate Public* Police, 130 HARV. L. REV. 1573. One notable example of insurers seeking to reduce their own risk exposure is police insurers. Rather than simply spending resources fighting legal battles, insurers help municipalities develop response procedures, train officers, and create "watch lists" for high-risk departments, allowing them to reduce the frequency of police misconduct and negligence.

108 HOWARD C. KUNREUTHER ET AL., *supra* note 92, at 231.

109 H.R. Report No. 144-470, at 173-174.

market is the subsidized monopoly of the NFIP."[110] The Committee resolved to explore legislative initiatives for the promotion of a private market.[111]

In March of 2018, Representative Royce (R-CA) introduced the GRATER Act, which would in part require the federal government to transfer "to the maximum extent possible" its insurance risk to the private sector, including reinsurance of already assumed risk.[112] As of January 2019, the bill remained in the Committee of Oversight and Government Reform, and its future is unclear given new leadership in the House.[113]

Other groups have similarly called for an increased role of private insurers in the flood insurance market. Executives from global insurer Lloyd's said the company would be interested in writing flood policies if they could competitively charge actuarially fair rates.[114] A number of think tanks and lobbying groups have expressed their support not just for the GRATER Act[115] but also for privatized flood insurance more generally.[116] Professors Ben-Shahar and Logue argue that the private market would incentivize more prudent development and disaster mitigation from homeowners than the NFIP could.[117]

There are a few key questions for policymakers who wish to explore increased privatization in the flood insurance market. First, could private insurers coexist with the NFIP, and if so, what would be the appropriate share of the market? If not, which is better prepared to cover flood risk? Are there advantages to keeping the government as the primary insurer in this market?

Perhaps unsurprisingly, both private insurers and NFIP administrators believe the others' presence in the market would be disruptive to their respective goals. There is a view that private insurers would "cherry-pick" low-risk properties which would ruin the cost-sharing and cross-subsidization that the NFIP relies on for stability.[118] However, a Wharton study of Texas counties with both coastal and riverine flood risk found that while on average, low-risk residents were over-charged while high-risk residents were

---

110 H.R. Report No. 144-470, at 173-74.

111 *Id*.

112 H.R. 5381, 115th Cong. § 2 (2018).

113 *All Actions, H.R. 5381*, CONGRESS.GOV (last visited Jan. 23, 2022), https://www.congress.gov/bill/115th-congress/house-bill/5381/all-actions.

114 Andrea Leinfielder, *Global Insurer Lloyd's calls for revamp of federal flood insurance*, HOUSTON CHRONICLE (Oct 11, 2017), https://www.houstonchronicle.com/business/article/Global-insurer-Lloyd-s-calls-for-revamp-of-12269612.php.

115 R.J. Lehmann, *Coalition Calls on U.S. House to Pass the GRATER Act*, R STREET INSTITUTE (Oct. 25, 2018), https://www.rstreet.org/2018/10/25/coalition-calls-on-u-s-house-to-pass-the-grater-act/.

116 *See, e.g.*, R STREET INSTITUTE, https://www.rstreet.org/publications/page/1/?issue=catastrophe-insurance.

117 Ben-Shahar & Logue, *supra* note 106.

118 Ray Lehmann, *Private Insurance is Getting Bigger, More Competitive, Less Profitable*, INSURANCE JOURNAL (Mar. 18, 2018), https://www.insurancejournal.com/blogs/right-street/2018/03/18/483689.htm.

under-charged, there existed areas where the reverse was true.[119] Thus, a private flood insurer might instead "cherry pick" high-risk properties from the NFIP.

The NFIP's disruption on the private market is more convincing. In addition to the previously discussed obstacle created by subsidized, artificially low premiums, the NFIP also has a series of unique advantages that private insurers do not share. First, as a federal program, the NFIP is not subject to state regulation, minimum capital requirements, and does not have to take solvency measures nor invest its reserves.[120] More obviously, the NFIP can be bailed out when its expenses exceed its costs, leaving little incentive to correct its pricing.[121] These factors in tandem make it difficult for private insurers to compete against the NFIP. Nevertheless, the private insurance market has been steadily growing, showing an increase in coverage in forty-nine states from 2016 to 2017.[122] In Washington, D.C., private insurers underwrite more than half of all coverage.[123]

There is evidence that it is private insurance, and not the NFIP, is more capable of handling the increasing uncertainty surrounding climate change. Unlike plenty of policymakers and pundits, insurance companies are serious about accurately evaluating and responding appropriately to the risks posed by climate change.[124] In drastic contrast to FEMA's struggles to bring its own maps into this century, discussed above, are the insurance industry's efforts to build climate science into their risk models.[125]

Insurers and their actuaries often build ambiguity into their premium pricing, which has the theoretical effect of somewhat suppressing sales.[126] Thus, insurers generally wish to reduce ambiguity to charge the lowest possible prices; the insurer with the least ambiguity usually wins in this arena.[127] In private markets like car insurance, factors like daily mileage, vehicle safety ratings, and desirability of a vehicle to thieves all impact a consumer's premium price.[128] This sort of detailed risk assessment gives companies a

119 Erwann Michel-Kerjan et al., *Could Flood Insurance Be Privatised in the United States?*, 40 THE GENEVA PAPERS ON RISK AND INSURANCE 179, (2014) (using commercially developed probabilistic models to compare what a private insurer might charge with the NFIP's premium prices).

120 SAM FRIEDMAN, THE POTENTIAL FOR FLOOD INSURANCE PRIVATIZATION IN THE U.S. 5 (2014) (accessible at https://www2.deloitte.com/us/en/pages/financial-services/articles/flood-insurance-privatization-in-the-united-states.html).

121 Isidore, *supra* note 1.

122 Knowles and Kunreuther, *supra* note 19.

123 *Id.*

124 Eugene Linden, *How the Insurance Industry Sees Climate Change*, L.A. TIMES (Jun. 16, 2014), https://www.latimes.com/opinion/op-ed/la-oe-linden-insurance-climate-change-20140617-story.html.

125 *Id*.

126 HOWARD C. KUNREUTHER ET AL., *supra* note 92, at 150.

127 *Id*.

128 *See,e.g.*, State Farm, *These 7 Factors Determine Car Insurance Premiums*, https://www.statefarm.com/simple-insights/saving/these-7-factors-determine-car-insurance-premiums (last visited Jan. 23, 2022).

competitive edge in reducing ambiguity—but absent any serious competitive threats to its existence, the NFIP has not used similar strategies, to the detriment of consumers.[129] It is thus clear that the private market has untapped ability to solve many of the problems in the nation's flood resiliency.

2. Utilizing the Private Market to Solve America's Resiliency Problem

While allowing flood insurance to be covered privately has a host of advantages, it is an emerging market, and likely not prepared to take on the entirety of the nation's flood risk. Though FEMA's maps are rife with issues, they have more than forty-year head start with a de facto monopoly. If Congress truly wants to utilize the private market, a scheme that keeps the NFIP intact yet removes the institutional obstacles it creates for private insurers is ideal.

Though not to the extent that the proposed GRATER Act would require, FEMA has already begun utilizing the private industry for reinsurance of its own liability.[130] In 2018, pursuant to the HFIAA, FEMA began reinsuring its own risk through private insurers, totaling nearly $2,000,000,000.[131] While this is certainly a positive step for purposes of deficit reduction and reducing burdens on taxpayers, the underlying risk is still poorly evaluated, potentially leaving FEMA at a disadvantage when entering such agreements.[132]

A more effective solution might be found in the creation of a government corporation. A government corporation generally is "a government agency that is established by Congress to provide a market-oriented public service and to produce revenues that meet or approximate its expenditures."[133] Harry Truman explained in a budget message that

> Experience indicates that the corporate form of organization is peculiarly adapted to the administration of government programs which are predominately of a commercial character—those which are revenue producing, are at least potentially self-sustaining and involve a large number of business-type transactions with the public. In their business operations such

---

129 *See* Kate Ho & Robin Lee, *Health Insurance Competition: Effects on Premiums, Hospital Rates, and Welfare*, MICROECONOMIC INSIGHTS (Jun. 22, 2017) ("[C]onsumers are harmed when an insurer is removed in our setting - even when premiums are predicted to fall - because restricted choice sets and the consequent reduction in product variety can have substantial effects on consumer welfare."), http://microeconomicinsights.org/health-insurance-competition-effects-premiums-hospital-rates-welfare/.

130 FEMA, *National Flood Insurance Program's Reinsurance Program*, https://www.fema.gov/nfip-reinsurance-program (last visited Dec. 31, 2018).

131 *Id*.

132 It is no stretch to say that if private insurers are better at evaluating homeowner risk, they are also likely better at evaluating reinsurance risk.

133 "Federal Government Corporations: An Overview," Congressional Research Service (June 08, 2011).

> programs require greater flexibility than the customary type of appropriations budget ordinarily permits.[134]

Exactly why it is that Congress has not removed the program from federal agencies and converted it to a corporation is unclear, especially given that the Truman characteristics are unquestionably present in the current NFIP model.

One important characteristic of public corporations that distinguishes them from a federal agency is their budgeting and finance process.[135] Crucially, instead of Congressional appropriations, public corporations receive most of their funding from users. Congress has only limited time to address flood insurance prices, and as evidenced by the HFIAA, is more focused on satisfied constituents than with the solubility or effectiveness of the NFIP.[136] This is hardly Congress' fault; responsiveness and accountability are part of their institutional design. But when flood risk can be objectively quantified and priced, and residents are dismayed, lobbying Congress is a rent-seeking behavior that undermines the mitigatory incentives of insurance.[137] A for-profit corporation, on the other hand, would make premium prices its primary concern, and would be less willing to discount premiums for political capital.

Second, a for-profit corporation would likely be more attentive to flood map accuracy. Where FEMA's map efforts are dependent on Congressional funding and prioritized poorly,[138] private insurers' risk assessments are their competitive weapon, allowing them to charge lower prices than competitors. It would be unlikely that a mixed-ownership corporation would follow FEMA's example by letting maps expire or basing their ratings on but a few criteria. While this certainly would raise premium prices for many homeowners, it might also lower them for others by using more granular, industry-standard risk models.

Lastly, using this private model would both allay the crowding-out effect caused by subsidized premiums, allowing the private market to grow, and allowing Congress and the Department of the Treasury to act as the insurer of last resort. Preventing a private insurer retreat like that seen in the 1927 Mississippi Flood has always been a function of government-led insurance. For that reason, a federally provided option would provide an invaluable backstop in crisis situations. Yet more private options could allay the

---

134 KEVIN KOSAR, CONG. RSCH. SERV., FEDERAL GOVERNMENT CORPORATIONS: AN OVERVIEW 5 (2011) (quoting U.S. Congress, House, *Document No. 19*, 80th Congress, 2d session (Washington: GPO, 1948), pp. M57-M62), https://sgp.fas.org/crs/misc/RL30365.pdf.

135 *Id*. at 6.

136 Fier et al., *supra* note 9, at 134.

137 Chris Rosato, *Lawmakers plan to reauthorize the National Flood Insurance Program*, WAFB (Nov. 4, 2021, 6:19 PM) (current Congressional support for legislative action curbing FEMA's efforts to base prices upon a more accurate risk assessment is an example of how a corporate form might more effectively administer a flood insurance program), https://www.wafb.com/2021/11/04/lawmakers-plan-reauthorize-national-flood-insurance-program/.

138 *Id*.

dismal take-up rates and broaden access—or even interest in taking out a flood insurance policy. Further, it could decrease the reliance of wealthy communities on an inequitable SFHA appeals process. If a coastal property owner feels that federal maps are inaccurate, they may instead seek a competitor's product; if no such product exists, it can be assumed that the rate is close to fair, and paying a lower rate would be detrimental to the risk pool as a whole.

CONCLUSION

In sum, the NFIP represents what can only be called a policy victory. Faced with private insurer retreat, Congress found a way to pay for flood risk in a way preferable to reactive, ad hoc disaster relief declarations. However, with the increasing unpredictability in flood risk due to climate change, and with the federal deficit becoming an increasingly political issue, outdated maps, under-enrollment, and premiums that fail to reflect flood risk (and therefore, do not incentivize sustainable development and mitigation), it is clear that reform is more than necessary. A public corporation would be a way for Congress to ensure that flood insurance be available to Americans without wasting the considerable benefits created by a competitive and focused private market.

File: Mendenhall 17.1 To Publish_Edited Created on: 4/19/2022 10:37:00 AM Last Printed: 6/14/2022 9:28:00 PM

WAS WASHINGTON'S FIRST TERM LEGITIMATE?: *TEXAS V. WHITE* AND THE CONSTITUTIONAL CONVENTION

Slade Mendenhall[*]

INTRODUCTION

When scholars and professors of constitutional law recount the origins of the Union, two conflicting theoretical accounts are often presented: one in which the Constitutional Convention of 1787 was convened, as the Founders presented at the time, under the authority of the Articles of Confederation and the Constitution was created under the authority to amend the Articles (we shall call this the Formal Theory); another in which the Articles were scrapped altogether and

the Convention proceeded of its own will, with the Constitution emerging as an artifact of political realism (the Realist Theory). The standard view today embraces the latter account, with the Founding Fathers taking it upon themselves to abandon the Articles in what one could alternately characterize as a surreptitious seizure of power, or a daring act of political entrepreneurship assumed at great risk to themselves had the states rejected it.

In a landmark Reconstruction Era case, *Texas v. White*,[1] however, Justice Salmon P. Chase adopted the former view, in which the Articles are seen as preserved, forming a backbone of continuity stretching forth from the time of their adoption.[2] In the Articles' assertion of the Union as "perpetual," the Chief Justice found sufficient cause for seeing the Confederate States as having never truly seceded but as having been in a state of revolt.[3] Joseph Story, too, writing many years earlier in Commentaries on the Constitution of the United States, pointed to the term "perpetual" as making the Union indissoluble.[4] Where the two men differed was in interpreting the effect of the Constitution on the force of the Articles.[5] Story, pointing to the Philadelphia Convention's letter to state ratifying conventions stating that it was effecting a "consolidation of the Union," took the view that the Constitution

---

[*] Attorney, Solicitor General's Office, Georgia Department of Law. Contact: smendenhall@law.ga.gov. I would like to thank Francis Buckley for inspiring the pursuit of these questions and for his valued input throughout the writing. Usual disclaimers apply.

[1] 74 U.S. 700 (1869).

[2] *Id.* at 724-26.

[3] *Id.* at 725-26.

[4] Brion McClanahan, *Is Secession Legal?*, THE AMERICAN CONSERVATIVE (Dec. 7, 2012), https://www.theamericanconservative.com/articles/is-secession-legal/.

[5] *Id.*

superseded the Articles without disturbing the continuity of the Union.[6] Chase took the view that the Constitution merely made the Union "more perfect" without dispensing with the Articles.[7]

If, however, the Articles remained in effect and the Constitution was merely their refinement, then let us assume, in the traditions of legal interpretation, that the Constitution replaced only those provisions which the two documents contradict while leaving uncontradicted provisions intact. Then we are adopting the Formal Theory, contending that the Constitution was, after all, erected within the procedures and authority of the Articles. If so, the very procedural provision from which the word "perpetual" was drawn would entail a sizable wrinkle. To offer the fuller passage, Article XIII of the Articles reads, in relevant part, that "the Union shall be *perpetual*; nor shall any alteration at any time hereafter be made in any of them; unless such alteration be agreed to in a Congress of the United States, and be afterwards confirmed by the Legislatures *of every State*."[8]

The unanimity mandated by the Articles is a high bar and one in keeping with the model on which it was constructed, of a looser confederation of independent polities. Unanimity rules are, of course, forever plagued by the holdout problem of one actor's ability to obstruct the benefits of a change to all others by withholding consent.[9] So the Founders learned in dealing with Rhode Island, which, after declining to send a delegate to the Constitutional Convention, did not ratify the Constitution until May 29, 1790.[10] Fortunately for their efforts, the Founders had abandoned the unanimity principle of the Articles for a supermajority threshold of nine states to ratify.[11]

This article reasons that the Articles of Confederation's unanimity requirement combined with Rhode Island's late ratification of the Constitution would, under a Formal Theory approach, render George Washington's first term as president, and any legislation it produced, a nullity. If, as the Articles of Confederation demanded, every state's approval was required for the Constitution qua amendment to be brought into effect, and Rhode Island did not ratify until nearly sixteen months after George Washington was elected as president, then the newly created federal government was illegitimate until that date, along with Washington's election to his first term. Alternatively, the Constitution truly is a product of political realism wholly divorced from

---

6 *Id.*

7 *See* White, 74 U.S. at 725.

8 ARTICLES OF CONFEDERATION OF 1781, art. XIII (emphasis added).

9 *See generally* Knut Wicksell, *A New Principle of Just Taxation in* CLASSICS IN THE THEORY OF PUBLIC FINANCE 72 (Palgrave Macmillan 1967); JAMES M. BUCHANAN & GORDON TULLOCK, THE CALCULUS OF CONSENT 87–88 (1st ed. 1965).

10 Samantha Payne, *"Rogue Island": The last state to ratify the Constitution*, PIECES OF HISTORY (May 18, 2015 6:01 AM), https://prologue.blogs.archives.gov/2015/05/18/rogue-island-the-last-state-to-ratify-the-constitution/.

11 The Avalon Project, NOTES ON THE DEBATES IN THE FEDERAL CONVENTION – MADISON DEBATES CONTENTS August 31 (2008), https://avalon.law.yale.edu/18th_century/debates_529.asp [hereinafter *Madison's Notes*].

the Articles' authority, in which case Chief Justice Chase's argument in *Texas v. White* must be wrong.

Ultimately, this article concludes, through carefully detailed reasoning, that the Realist Theory is far less troubling in its implications. I make the case here that we must either accept, once and for all, that (i) the Constitutional Convention was a revolutionary usurpation of the Articles of Confederation, (ii) all portent of the Framers adherence to the Articles authority in crafting the Constitution must be abandoned, and (iii) cases such as *Texas v. White* that ground their conclusions in some remaining "background" force of the Articles are wrong, or we must accept that President George Washington's first term was unconstitutional and that all legislation issued during it was void on arrival. There is, it seems, no third way about it.

The following section presents a brief background of the Constitutional Convention, with special regard for the Founders' comments on the Articles and how their actions related to the Articles' terms and authority. Section Three summarizes how theories of the Convention have variously held it to be lawful or unlawful and within or beyond the authority of the Articles. Section Four examines, in detail, two occasions on which the Supreme Court has leaned on the Articles as a valid authority, at least once suggesting that they continue to run concurrent with the Articles: *Texas v. White*[12] and *United States v. Wheeler*.[13] Finally, Section Five relates these questions to the facts of Rhode Island's ratification and what it would mean for the legal and political legacy of the Washington administration. Section Six concludes.

I. BACKGROUND

The origins of the Constitutional Convention have been so often recited elsewhere and in greater detail by trained historians that it will little profit to revisit them at length here. I will therefore only pursue such a summary as is needed to tee up our questions and bring attention to the points of controversy raised below, namely whether the Constitution was adopted and ratified within the authority of the Articles of Confederation in such a way as to render it logical that the Articles would in some sense have continued to run as a concurrent source of legitimacy and of law.

The first attempt at a meeting meant to discuss the widely perceived shortcomings of the States and their national affiliation under the Articles was, of course, the September 11, 1786 meeting in Annapolis.[14] It was called for the purpose of "considering how far an uniform system in the commercial regulations may be necessary to their common interests, and their permanent

---

12 74 U.S. 700 (1869).

13 254 U.S. 281 (1920).

14 ST. GEORGE TUCKER, VIEW OF THE CONSTITUTION OF THE UNITED STATES 109 (Clyde N.Wilson ed., 1999).

File: Mendenhall 17.1 To Publish_Edited Created on: 4/19/2022 10:37:00 AM Last Printed: 6/14/2022 9:28:00 PM

harmony; and to report to the several states such an act, relative to that object, as when unanimously ratified by them, would enable congress effectually to provide for the same."[15] Finding commissioners from only five of thirteen states among them and deeming their small number too few to meaningfully proceed, those gathered wrote a letter to their constituents, recommending the appointment of representatives to meet in Philadelphia the following May.[16] Virginia's legislature passed an act which approved the appointment of commissioners to take into consideration the situation of the United States; to devise such further provisions as shall appear to them necessary to render the Constitution of the federal government adequate to the exigencies of the Union; and to report such an act for that purpose, to the United States in Congress assembled, as when agreed to by them, and afterwards confirmed by the legislature of every State, will effectually provide for the same.[17]

All other States followed suit, save for Rhode Island.[18] Similarly, Congress passed a resolution in February 1787 reading, in relevant part,

> WHEREAS, There is provision in the articles of Confederation and perpetual Union, for making alterations therein, by the assent of a Congress of the United States, and of the legislatures of the several States; and whereas experience hath evinced, that there are defects in the present Confederation; as a mean to remedy which, several of the States, and particularly the State of New York, by express instructions to their delegates in Congress, have suggested a convention for the purposes expressed in the following resolution; and such convention appearing to be the most probable mean of establishing in these States a firm national government:
>
> Resolved, That in the opinion of Congress it is expedient, that on the second Monday of May next a convention of delegates, who shall have been appointed by the several States, be held at Philadelphia, for the sole and express purpose of revising the Articles of Confederation, and reporting to Congress and the several legislatures such alterations and provisions therein, as shall, when agreed to in Congress, and confirmed by the States, render the federal Constitution adequate to the exigencies of government and the preservation of the Union.[19]

Planned to begin May 14, the resulting convention found itself again without a quorum when that day arrived.[20] By May 25, however, representatives of seven states had arrived and others gradually came as they could, and the plan for the States' more adequate national government unfolded with prodigious speed.[21]

On May 29, the first day of real, substantive (as opposed to procedural) discussion of priorities, Edmund Randolph of Virginia proposed fifteen

---

15 *Id.*

16 THE AVALON PROJECT, PROCEEDINGS OF THE COMMISSIONERS TO REMEDY DEFECTS OF THE FEDERAL GOVERNMENT, Sept. 11, 1786 (2008), https://avalon.law.yale.edu/18th_century/annapoli.asp.

17 THE FEDERALIST NO. 40 (James Madison).

18 Payne, *supra* note 10.

19 *Id.*

20 EDWARD S. CORWIN, THE CONSTITUTION OF THE UNITED STATES OF AMERICA ANALYSIS AND INTERPRETATION, S. DOC. NO. 170, at 12 (1952).

21 *Id.*

resolutions to set the Convention's overall purpose and aims.[22] These came to be known as the Virginia Plan.[23] First among the resolutions: "that the Articles of Confederation ought to be so corrected & enlarged as to accomplish the objects proposed by their institution; namely, 'common defence, security of liberty and general welfare.'"[24] The very next day, however, Randolph moved, on the suggestion of Gouverneur Morris, to postpone consideration of that resolution in favor of a set of three others:

> 1. that a Union of the States merely federal will not accomplish the objects proposed by the articles of Confederation, namely common defence, security of liberty, & genl. welfare.
>
> 2. that no treaty or treaties among the whole or part of the States, as individual Sovereignties, would be sufficient.
>
> 3. that a national Government ought to be established consisting of a supreme Legislative, Executive & Judiciary.

The motion for postponing was seconded by Mr. Govr. MORRIS and unanimously agreed to.[25]

The delegates thus delicately evaded the question of whether this was indeed a process to amend or to replace the Articles, and after the May 30 dispensation with the Virginia Plan's first resolution, it was set aside in favor of debate over substantive provisions.[26] On June 18, Alexander Hamilton roused the subject again before the Convention and addressed it head-on, not to resolve concerns over whether the Convention operated within the legitimacy of the Articles but to dispel them as too fastidious in light of the great consequences of their success or failure.[27] Madison described Hamilton's address thusly:

> He agreed moreover with the Honble gentleman from Va. [Mr. R.] that we owed it to our Country, to do on this emergency whatever we should deem essential to its happiness. The States sent us here to provide for the exigences of the Union. To rely on & propose any plan not adequate to these exigences, merely because it was not clearly within our powers, would be to sacrifice the means to the end. It may be said that the States can not ratify a plan not within the purview of the article of Confederation providing for alterations & amendments. But may not the States themselves in which no constitutional authority equal to this purpose exists in the Legislatures, have had in view a reference to the people at large.[28]

---

22 *Id.*

23 *Id.*

24 *Madison's Notes*, *supra* note 11, at May 29.

25 *Id.* at May 30.

26 CORWIN, *supra* note 20, at 12.

27 *Id.*

28 *Madison's Notes*, *supra* note 11, at June 18.

Two days later, on June 20, Oliver Ellsworth of Connecticut was first to speak, professing that "[h]e could not admit the doctrine that a breach of any of the federal articles could dissolve the whole," and warning his fellow delegates that "[i]t would be highly dangerous not to consider the Confederation as still subsisting."[29] Further,

> [h]e wished also the plan of the Convention to go forth as an amendment to the articles of Confederation, since under this idea the authority of the Legislatures could ratify it. If they are unwilling, the people will be so too. If the plan goes forth to the people for ratification several succeeding Conventions within the States would be unavoidable. He did not like these conventions. They were better fitted to pull down than to build up Constitutions.[30]

In response, John Lansing of New York contended "that the true question here was, whether the Convention would adhere to or depart from the foundation of the present Confederacy."[31] The remainder of Lansing's speech departs from the subject, however, and no other delegates saw fit to take up Ellsworth's contentions in favor of the Articles.[32]

So, it seems, is the extent of discussion of the relationship between the Articles and the Constitution at the Convention. Even in Madison's general descriptions of the delegates' discourse, lacking verbatim wording, a tenor of avoidance comes through in which most of the delegates seem implicitly committed to avoiding sticky questions about how their present meeting relates to the Articles. David Kyvig notes as much, writing that "[a]s discussion of the Virginia resolutions proceeded, the delegates vacillated as to whether they were amending the Articles or doing something other."[33]

Whatever the preference for silence on the subject, in one sense, the debate over legitimacy within the Articles was transmuted into one about ratification. The fifteenth resolution of the Virginia Plan read,

> Resd. That the amendments which shall be offered to the Confederation, by the Convention ought at a proper time, or times, after the approbation of Congress to be submitted to an assembly or assemblies of Representatives, recommended by the several Legislatures to be expressly chosen by the people, to consider and decide thereon.[34]

But as Professor Carlos Gonzales notes, the resolution was clear as to the most important issue: ratification would be secured through special

---

29 *Id.* at June 20.

30 *Id*.

31 *Id.*

32 *Id.*

33 DAVID E. KYVIG, EXPLICIT AND AUTHENTIC ACTS: AMENDING THE U.S. CONSTITUTION, 1776-1995 46 n. 23 (1996).

34 *Madison's Notes*, *supra* note 11, at July 19.

ratification assemblies or conventions, not by standing state legislatures.[35] The matter was debated on June 5, with delegates commending the ratification convention approach as a more direct form of ratification by the people in a way that ratification by legislatures was not.[36] Roger Sherman, speaking against it, pointed out that the Articles already provided for ratification of amendments by "the assent of Congs. and ratification of the State Legislatures."[37] Madison, speaking next, called the Articles "defective" for "resting in many of the States on Legislative sanction only.[38] Others joined in the debate, with Rufus King taking Madison's side of the issue and Elbridge Gerry joining with Sherman.[39] A week later, on June 12, the fifteenth resolution was adopted.[40]

The perception quickly emerged that the Virginia Plan was a plan to replace the Articles and, upon its introduction, the New Jersey Plan could be read as still within the context of amending them. The New Jersey Plan was described by John Lansing as being introduced "on the basis of amending the federal government, and the other [the Virginia Plan] to be reported as a national government, on propositions which exclude the propriety of amendment."[41] And on June 16, James Wilson described the Virginia Plan as to be ratified "by the people themselves," whereas the New Jersey Plan would be ratified "by legislative authorities according to the 13 art: of the confederation."[42] The two plans, so often characterized by their designs for the future government, were, it seems, similarly distinguished by differing views over how the delegates could enact that government.

On July 23, Ellsworth again opened discussion, this time moving that the Constitution "be referred to the Legislatures of the States for ratification."[43] George Mason of Virginia objected, drawing from his own state's experience to note that those state constitutions were "established by an assumed authority."[44] In doing so, Professor Carlos Gonzales observes, Mason "unmistakably implies that ratification by ordinary legislatures had failed to confer a popular sovereignty pedigree on these state governments."[45] Ellsworth, in a final push for observing the amendment procedures of the

---

35 Carlos E. Gonzalez, *Representational Structures Through Which We the People Ratify Constitutions: The Troubling Original Understanding of the Constitution's Ratification Clauses*, 38 U.C. DAVIS L. REV. 1373, 1415-18 (2005).

36 *Madison's Notes*, *supra* note 11, at June 5.

37 *Id.*

38 *Id.*

39 *Id.*

40 *Id.* at June 12.

41 MAX FARRAND, 1 RECORDS OF THE FEDERAL CONVENTION OF 1787 246 (Max Farrand ed., Yale, rev. ed. 1937).

42 *Madison's Notes*, *supra* note 11, at June 16.

43 *Id.* at July 23.

44 *Id.*

45 Gonzalez, *supra* note 35, at 1408-09.

File: Mendenhall 17.1 To Publish_Edited Created on: 4/19/2022 10:37:00 AM Last Printed: 6/14/2022 9:28:00 PM

Articles, contends, "[t]he fact is that we exist at present, and we need not enquire how, as a federal Society, united by a charter one article of which is that alterations therein may be made by the Legislative authority of the States."[46] Ellsworth, however, knew that he was in the minority and was fighting a losing battle. As such, he retreated to a fallback position of supporting ratification by a non-unanimous threshold of state legislatures.[47] That, too, would fail.

Ultimately, on August 6, the Committee of Detail reported back to the Convention an Article XXI of the proposed draft, reading "[t]he ratifications of the Conventions of ___ States shall be sufficient to organize this Constitution."[48] Thus, whereas observation of the Articles' prescribed unanimity of legislatures approach had been abandoned, unanimity of state conventions was still a possibility. From the time of its proposal, that blank in Article XXI would linger for weeks. On August 31, Rufus King moved to amend Article XXI to append the words "between the said States," thereby limiting the Constitution to only those states which ratified.[49] James Madison answered with a compromise: that the blank be filled with "any seven or more States entitled to thirty three members at least in the House of Representatives according to the allotment made in the 3 Sect: of art: 4."[50] This would set the threshold requirement at a majority of both people and states.[51] In characteristic fashion, Madison's offer was diplomatic for its consideration of the differences between large and small states but purposeful in its effect, as it would still, after that threshold was met, bind all thirteen states.

Roger Sherman again doubted the propriety of adopting the Constitution with less than unanimity, "considering the nature of the existing Confederation."[52] Some debate was then had about letting each State choose its method of ratification, as delegates from Maryland emphasized that the Constitution of their state would not allow for adoption by any means except those that it prescribed.[53] This, however, was not long entertained nor taken up by many others. Daniel Carroll and Luther Martin moved, in support of unanimity, to fill the blank with "thirteen."[54] All but Maryland rejected the motion.[55] Sherman and Jonathan Dayton then moved to fill the blank with "ten," which tallied seven opposed and four in favor.[56] George Mason then motioned for the nine state threshold, contending that "[n]ine States had been required in

46 *Madison's Notes*, *supra* note 11, at July 23.
47 *Id.*
48 *Id.* at August 6.
49 *Id.* at August 31.
50 *Id.*
51 *Id*.
52 *Madison's Notes*, *supra* note 11, at August 31.
53 *Id.*
54 *Id.*
55 *Id.*
56 *Id.*

all great cases under the Confederation & that number was on that account preferable."[57] It carried eight-to-three.[58]

The answer would be affirmed when the Convention met on September 10 and resolved, upon James Madison's motion with Hamilton seconding, that

> The Legislature of the U.S. whenever two thirds of both Houses shall deem necessary, or on the application of two thirds of the Legislatures of the several States, shall propose amendments to this Constitution, which shall be valid to all intents and purposes as part thereof, when the same shall have been ratified by three fourths at least of the Legislatures of the several States, or by Conventions in three fourths thereof, as one or the other mode of ratification may be proposed by the Legislature of the U.S.[59]

The Framers who supported the measure clearly saw the Constitution's threshold for future amendments as inextricable from the decision rule that would govern their own present efforts. Alexander Hamilton discussed the proposition interchangeably with the idea of "allow[ing] nine States . . . to institute a new Government on the ruins of the existing one" and predicted that "[n]o convention convinced of the necessity of the plan will refuse to give it effect on the adoption by nine States."[60] Rather clearly, they saw their vote on August 31 and votes on future amendments as being one of a kind. Roger Sherman and Rufus King spoke in support of the nine-state rule, but support was not unanimous.[61] Elbridge Gerry notably objected to "the indecency and pernicious tendency of dissolving in so slight a manner, the solemn obligations of the articles of confederation" observing that "[i]f nine out of thirteen can dissolve the compact, Six out of nine will be just as able to dissolve the new one hereafter."[62] Ultimately (and fittingly), nine states voted to adopt Madison's proposition for a nine-state threshold, with only Delaware dissenting, New Hampshire divided, and New York's delegation lacking quorum.[63]

---

57 *Id.*

58 *Madison's Notes*, *supra* note 11, at August 31.

59 *Madison's Notes*, *supra* note 11, at September 10.

60 *Id.* Hamilton's choice of word, "ruins," as recorded by Madison, presuming that the summary is exact on this point, can be counted as a point for what we will explore below as the "Outside the Articles, Lawful" theory. On the other hand, the equivalence implied in September 10's discussion between the threshold to amend the Constitution and the threshold to create it in the first place could be interpreted as further evidence that the Framers were engaged in a process of amendment. Neither side of the debate seems to gain ground from that day's transcripts.

61 *Id.*

62 *Id.* Unfortunately, no delegates saw fit to answer Gerry's challenge. Had they, their words might have been a valuable contribution to the secession debates several generations hence, at the end of which one-third of the states (eleven out of thirty-three) attempted to secede.

63 *Id.*

II. THEORIES OF THE CONVENTION

The question of whether the Convention (and therefore the Constitution) were within the authority of the Articles is not the same question as whether the Articles continued to run concurrently with the Constitution and to have legal effect, but they are deeply related. If the Convention was outside of the authority of the Articles, then the political realist approach would seem to be validated, and it is difficult to imagine how the Articles were not entirely abandoned, even if the Convention and resulting Constitution were otherwise lawfully adopted. If, however, the Formal Theory is correct and the Convention was both within the Articles' authority and lawful, then it is at least possible that the Articles could continue to run.

These questions present two dimensions of analysis: whether the Articles emerged within the authority of the Constitution and whether it emerged lawfully. One does not necessarily imply or negate the other. Here, I consider three of the four possibilities that this leaves, both of the "outside of the Articles" options and the theory that the Constitution emerged within the Articles and lawfully. These, of course, imply the possibility of a theory holding the Constitution to be within the authority of the Articles but somehow otherwise unlawful. Knowing of no such argument in the literature nor how such an argument might run, I reserve that category. None of the summaries of arguments as to various positions should be taken as nearly exhaustive but merely as indicative samplings of divergent thought on a complex question.

A. *Outside of the Articles, Lawful*

The standard interpretation is that the Convention (and therefore the Constitution) emerged outside of the Articles but lawfully. As Professors Ackerman and Katyal put it, "[i]ndeed, there are remarkably few public efforts by Federalists to disguise the revolutionary character of their enterprise with legalistic argument. By their words and deeds, leaders like Madison and Wilson repeatedly indicated their belief that revolutionary, rather than legalistic, arguments provided their best defense."[64] Nonetheless, legalistic arguments were attempted. On July 23, Madison addressed the Convention to argue that "[t]he doctrine laid down by the law of Nations in the case of treaties is that a breach of any one article by any of the parties, frees the other parties from their engagements," but that "[i]n the case of a union of people under one Constitution, the nature of the pact has always been understood to exclude such an interpretation," thereby nullifying the concerns of some that

64 Bruce Ackerman & Neal Katyal, *Our Unconventional Founding*, 62 U. CHI. L. REV. 475, 488 (1995).

their departure from the Articles' terms for amendment might be illegitimate.[65]

In Federalist 40, published as a New York newspaper article on January 18, 1788, Madison claims that those who had quarreled with the legality of their procedure had by then "waived" their objection and that we could therefore "dismiss it without further consideration."[66] As Ackerman and Katyal note, however, critics were waging this very argument in the New York General Assembly and Senate, unaware, along with many likeminded compatriots, that they had apparently waived it.[67] Madison surely knew as much and, despite trying to dismiss these objections, nonetheless seemingly felt the need to address them.

In No. 40, Madison further grapples with challenges to the Convention's authority and instructs us to begin with an inspection of the delegates' commissions.[68] All of these, he notes, had reference either to the recommendations of the Annapolis meeting of September 1786, or of Congress in February 1787.[69] In interpreting the guidance given by these calls, Madison asks us,

> Suppose, then, that the expressions defining the authority of the convention were irreconcilably at variance with each other; that a national and adequate government could not possibly, in the judgment of the convention, be affected by alterations and provisions in the Articles of Confederation; which part of the definition ought to have been embraced, and which rejected? Which was the more important, which the less important part? Which the end; which the means?[70]

In this, he seems set on a course of justifying abandonment of the Articles, arguing that the dual mandate of the Convention, (i) to provide an adequate government (ii) by amending the Articles, was irreconcilable and that one of the two had to give.[71] Nonetheless, he contends, the Constitution that emerges from them is legitimate for several reasons.

First, he argues, "the great principles of the Constitution proposed by the convention may be considered less as absolutely new, than as the expansion of principles which are found in the articles of Confederation."[72] Therefore, the principles underlying the national government are preserved even if their institutional manifestations are amended. No challenge, one might say, to the natural law has been affected by their drafting nor would be by their ratification.

---

65 *Madison's Notes*, *supra* note 11, at July 23.
66 Ackerman & Katyal, *supra* note 64, at 546; THE FEDERALIST NO. 40 (James Madison).
67 Ackerman & Katyal, *supra* note 64.
68 THE FEDERALIST NO. 40 (James Madison).
69 *Id.*
70 *Id.* at 2.
71 *Id.*
72 *Id.* at 3.

Second, he writes with evident frustration, despite the "powers of the convention hav[ing] been analyzed and tried with the same rigor, and by the same rules, as if they had been real and final powers for the establishment of a Constitution for the United States," they "were merely advisory and recommendatory. . . . [T]hey were so meant by the States, and so understood by the convention; and . . . the latter have accordingly planned and proposed a Constitution which is to be of no more consequence than the paper on which it is written, unless it be stamped with the approbation of those to whom it is addressed."[73] Thus, Madison says, our preoccupation with the authority of the Convention is misplaced; with no power to enact but only to recommend, the Convention is legitimate even on the weakest view of its powers because the real power comes from the States through ratification.[74]

Finally, he argues that "[t]he prudent inquiry, in all cases, ought surely to be, not so much from whom the advice comes, as whether the advice be good" or, as he later refines it, "if [the delegates to the Convention] had violated both their powers and their obligations, in proposing a Constitution, this ought nevertheless to be embraced, if it be calculated to accomplish the views and happiness of the people of America."[75] An even more explicit and full-throated natural law argument, this one speaks for itself.

Professor Akhil Amar has argued that "inconsistency is not illegality" and that Federalists' disregard of the Articles' clear terms requiring approval by Congress and all state legislatures did not place their project outside of the law.[76] The Articles, in his view, had become voidable by any state choosing to renounce them. They were never, he contends, more than a "tight treaty among thirteen otherwise independent states" that nowhere described itself as a "Government" or "legislature," nor "its pronouncements as 'law.'"[77] Under well-established legal principles, he therefore argues, "these material breaches freed each compacting party each state to disregard the pact, if it so chose."[78]

Another, more nuanced argument by Professor Robert Natelson situates the Constitutional Convention within the longstanding tradition of interstate conventions stretching back decades.[79] Referencing the commonality of conventions of states up to that time, this view notes that such conventions were

---

73 Id.

74 THE FEDERALIST NO. 40 (James Madison) at 3.

75 Id. at 4.

76 Akhil Reed Amar, *The Consent of the Governed: Constitutional Amendment Outside Article V*, 94 COLUM. L. REV. 457, 465 (1994) [hereinafter *Consent of the Governed*].

77 *Id.*

78 *Id.*

79 *See* Robert Natelson, *Yes, the Constitution was Adopted Legally*, THE HILL (April 11, 2017 7:00 AM), https://thehill.com/blogs/pundits-blog/uncategorized/328112-yes-the-constitution-was-adopted-legally [hereinafter *The Constitution was Adopted Legally*]; *see generally* Robert G. Natelson, *Founding Era Conventions and the Meaning of the Constitution's "Conventions for Proposing Amendments,"* 65 FLA. L. REV. 615 (2013) (discussing the long tradition of state conventions).

not held pursuant to the Articles but under sovereign powers reserved to the states.[80] Professor Natelson argues that it is a common mistake to see the Convention as called by the February 21, 1787 resolution of the Confederation Congress.[81] He points to the events of late 1786, when, in September, delegates from five states to the Annapolis Convention recommended to their states another convention in Philadelphia.[82] By the end of November, New Jersey had appointed commissioners to it.[83] And on December 1, the Virginia legislature approved the Convention and directed their governor to communicate their resolution to all other states.[84] This, Natelson argues, was the formal call to Philadelphia, and no state expressly limited its delegates to the task of amending the Articles.[85] It was not until New York and Massachusetts, concerned about the breadth of the call, recommended some limitations that Congress stated, by Resolution, an opinion (meaning not a call or an order) that the Convention should be held to amend the Articles.[86] Ten states out of the attending twelve, Natelson notes, gave their delegates sweeping proposal powers.[87]

This argument, entirely plausible, gets us to an explanation of how the Convention was permissible under decades of established practice and was well within existing customs of interstate relations at the time. It does not, however, get us an explanation of how it was not in direct contradiction to the Articles. It is an account of an otherwise legal method being used to circumvent the channels prescribed by the Articles, rendering Article XIII mere surplusage. To the extent that the Articles, in prescribing their own method of amendment, meant to supplant and prohibit extraneous methods of fundamentally altering relations between the states, the commonality of state conventions and the use of such a convention even to amend the Articles (if not to abrogate them altogether) would seem to be an usurpation of the Articles. The inclusion of Article XIII seems to preclude such outside measures, and any who would argue otherwise may find themselves hard-pressed to explain how Article V of the Constitution would not be equally vacuous should a sufficient number of political leaders today decide to abandon that document without adhering to its prescribed procedures. Presented as an argument of lawfulness based in emerged norms, this theory would seem to inevitably wind up at a conclusion of political realism that it labored to avoid.

---

80 *The Constitution was Adopted Legally*, *supra* note 79.
81 *Id.*
82 *Id.*
83 *Id.*
84 *Id.*
85 *Id.*
86 *The Constitution was Adopted Legally*, *supra* note 79.
87 *Id.*

Whatever the rationale behind it, the notion that the Convention and resulting Constitution were at once lawful and an abandonment of the Articles seems to have been the implicit understanding of the Federalist delegates to the Convention and has since become the standard view in historical and legal scholarship on the Constitution and its origins.[88]

B. *Outside of the Articles, Unlawful*

Antifederalists, more broadly imbued with a concern for the loss of state autonomy that the new Federalist plan might mean, were nonetheless also animated by a concern for the procedure and legitimacy of the Convention. Patrick Henry firmly denounced the nine-state decision rule at the Virginia ratifying convention.[89] In New York's General Assembly, Cornelius Schoonmaker introduced a resolution denouncing the convention for its illegality, losing by a vote of twenty-seven to twenty-five, and Robert Yates' similar motion would lose twelve to seven.[90] A young John Quincy Adams, writing during the ratification period, contended that to crown the whole the 7th: article, is an open and bare-faced violation of the most sacred engagements which can be formed by human beings. It violates the *Confederation*, the 13th: article of which I wish you would turn to, for a complete demonstration of what I affirm; and it violates the Constitution of [Massachusetts], which was the only crime of our Berkshire & Hampshire insurgents (in Shays's Rebellion).[91]

Elsewhere, other writers have contended, whatever they think of its content and effects, that the Constitution's origins are plainly illegal. "Illegality was a leitmotif at the convention from its first days to its last," write Professors Ackerman and Katyal.[92] They point to the fact that many Americans at the time did not see the Articles as a mere treaty that could be treated as abandoned,[93] noting "an enormous body of evidence expressing legalistic objections to the Federalists' unconventional activities" and Federalists' evident need to respond to these objections by "making the revolutionary assertion that the times required breaking the rules laid down by Article XIII."[94] Indeed, they note, Madison responded to the characterization of the Articles

---

88 *See, e.g.*, JOSEPH STORY, COMMENTARIES ON THE CONSTITUTION (1833); AKHIL REED AMAR, AMERICA'S CONSTITUTION: A BIOGRAPHY (2006) [hereinafter *America's Constitution*]; *Consent of the Governed*, *supra* note 76.

89 Ackerman & Katyal, *supra* note 64, at 548.

90 *Id.* at 546.

91 *Id.* at 487.

92 *Id.* at 506. It should be noted here that these authors ultimately conclude that the Convention, though unlawful, nonetheless produced a governing document that was legitimate, so this quote should not be construed out of context to cast their views otherwise.

93 *Id.* at 542.

94 *Id.* at 540.

as a mere treaty by describing "the federal union as anal[o]gous to the fundamental compact by which individuals compose one Society, and which must in its theoretic origin at least, have been the unanimous act of the component members."[95] He would then go on to make qualified comparisons to a treaty structure but never abandon the notion, key to the Virginia Plan, of one nation composed of one people.[96]

And objections to end-around approaches to reform were not limited to Anti-Federalists, at least if one is willing to look to the years prior to the Convention. Formalists had stressed the necessity of adhering to the Articles' substantive and procedural terms since before the Convention was ever proposed. John Jay, the prominent Federalist and representative of New York at the Convention, and Thomas Burke, the prominent North Carolina governor who died several years before its meeting, are singled out by David C. Hendrickson as having strongly maintained that "the authority of the congress rested on the prior acts of the several states, to which the states gave their voluntary consent, and until those obligations were fulfilled, neither nullification of the authority of congress, exercising its due powers, nor secession from the compact itself was consistent with the terms of their original pledges."[97]

Two soon-to-be Federalists from Massachusetts, Rufus King and Nathan Dane, responded to the Annapolis meeting's call for a convention in Philadelphia by denouncing it as unconstitutional:

> The Confederation was the act of the people. No part could be altered but by consent of Congress and confirmation of the several Legislatures. Congress therefore ought to make the examination first, because, if it was done by a convention, no Legislature could have a right to confirm it . . . . Besides, if Congress should not agree upon a report of a convention, the most fatal consequences might follow. Congress therefore were the proper body to propose alterations.[98]

In a surprising change of heart, however, both men would not only attend the convention but withdraw their concerns and come to vote for ratification.[99] Indeed, Rufus King would join Madison in advocating for ratification via convention rather than via state legislatures, further departing from the Articles' clear requirements.[100]

As the delegates left Philadelphia and the ratification debate spread across the country, a major objection of the Antifederalists was to the replacement of Article XIII's unanimity requirement with the nine-out-of-

95 Ackerman & Katyal, *supra* note 64, at 542–43.

96 *Id.*

97 DAVID C. HENDRICKSON, PEACE PACT: THE LOST WORLD OF THE AMERICAN FOUNDING 153–54 (2003).

98 Ackerman & Katyal, *supra* note 64, at 501.

99 *Id.* at 503.

100 Gonzales, *supra* note 35, at 1403–04.

File: Mendenhall 17.1 To Publish_Edited Created on: 4/19/2022 10:37:00 AM Last Printed: 6/14/2022 9:28:00 PM

thirteen threshold for approval adopted at the Convention. Many did not initially go directly to calling the enterprise therefore illegal.[101] The statement published on December 18 by the minority delegates at the Convention made a wide array of objections, including questioning whether those delegates commissioned by their states to "amend" the Articles had exceeded their powers, but despite the time spent on the question at the Convention, they neglected to mention the unanimity issue.[102] Others evinced no inclination to skirt the issue of such a fundamental rule change. As the Antifederalist writer Portius asked his audience of Massachusetts readers,

> [H]ow can Nine States dissolve a System of Government, which Thirteen had instituted, and which the whole Thirteen pledged their faith to each other should not receive any alterations without the consent and approbation of the whole Thirteen? . . . Or, in another point of view, what right has this State either at their own instance, or at the recommendation of any body of men whatever, to break through the established Constitution of the United States and openly set at defiance that System of Federal Government, for the support of which, they had pledged their most solemn engagements and sacred honour?[103]

His conclusion: that if the concern for ratification procedure "is not obviated, [it] cannot fail of over-throwing the whole structure, and reduce it to the situation of a baseless fabrick of nocturnal reverees."[104]

And what would be the status of those states who chose not to join? If the rule was not to be unanimity, then what would be the status of relations with those states that declined to ratify? Would they be forced into the new union? Made foreign states? Favored nations? The Framers present at the Convention did not say. As Amar notes,

> Generally, [the friends of the Constitution] seemed to concede that governance under the Constitution would be incompatible with continuation of the Articles of Confederation, and maintained a prudent silence on the precise nature of the relationship the new union would work out with any nonratifying states. See, e.g., *Federalist* No. 43.[105]

Portius was not so silent. Indicating his own conviction as to how nonmember states should respond, he wrote,

> Supposing Nine States should ratify and confirm the proposed Federal Government, and Four States should reject the same, Would not those Four States, still adhereing (*sic*) to the Articles of Confederation, have an undoubted right, both in the sight of God and man, to accuse the

---

101 *The Address and Reasons of Dissent of the Minority of the Convention of Pennsylvania to Their Constituents*, PA. PACKET & DAILY ADVERTISER, Dec. 18, 1787.

102 *Id.*

103 Portius, *To the People of Massachusetts*, AMERICAN HERALD, Nov. 12, 1787.

104 *Id.*

105 *Consent of the Governed*, *supra* note 76, at 465.

> Nine approbating States with the most unequivocal breach of public faith, point-blank national infidelity, and I will add, of open rebellion against the National Constitution![106]

While a legally defensible contention, it still does not address how signatory states to the new Constitution should view continued relations with those states that were not persuaded to join. James Wilson commented on August 30, in the course of debates over how many states were necessary to ratify the Constitution, that "the States only which ratify can be bound."[107] Daniel Carroll of Maryland responded by insisting that thirteen states' assent be required, "unanimity being necessary to dissolve the existing confederacy which had been unanimously established."[108] Rufus King endorsed Carroll's measure, "otherwise as the Constitution now stands it will operate on the whole though ratified by a part only."[109] That comment coming at the end of the day's transcript, it would have seemed an opportune moment for those proponing the nine-out-of-thirteen threshold to challenge the assumption that the Constitution would be expected to operate on all thirteen states even if they did not ratify it, but no one spoke up, again leaving it unstated but implied that this was the Founders' understanding.[110] As we shall further explore below, in the months prior to ratification by all thirteen, the chosen strategy of the ratifying states was a combination of presumption, gentility, and hard diplomacy.

The variation in opinions among the Framers and their contemporaries, both across individuals and, often, within the same individual over time is understandable and not easily attributed to mere political convenience or fleeting whims. These were difficult questions, with good arguments on both sides being made by some of the world's most sophisticated political minds. Thus, common and easy as it may be to dismiss those who argue today that there was legal impropriety in the transition from the Articles to the Constitution, it is a case with which to be reckoned and to be addressed directly, as many notable figures saw fit to do at that time.

C. *Within the Articles, Lawful*

The idea that the Articles and Constitution may not be in conflict is a view that has been either advocated or implied in scholarship and Supreme Court reasoning more than once, not least by James Madison in the very same writing discussed above.[111] Professor Gregory Maggs writes, "The theory is

---

106 Portius, *supra* note 103.

107 *Madison's Notes*, *supra* note 11, at August 30.

108 Id.

109 *Id.*

110 *Id.*

111 THE FEDERALIST NO. 40 (James Madison).

controversial because it goes against the generally accepted idea that the Constitution replaced the Articles of Confederation, and that the Articles, like a repealed statute or rescinded treaty, have no continuing validity. Yet, despite being somewhat counterintuitive, the theory has a strong pedigree and appears to be correct at least in some instances."[112] His analysis points to a number of considerable arguments for the view.

First and, as he rightly notes, "with little practical consequence," is the origin of the name "United States of America" being in Article I of the Articles of Confederation.[113] The Constitution, so that argument goes, only presumed but did not declare the name of the Union and its government, merely taking the point as established.[114] Naming provisions not going to the heart of what it means to be a nation, this is not a very considerable argument but nonetheless one worth noting. It arguably speaks in favor of some form of continuity in the Union and a theory of the Constitution qua amendment. Second, and more significant, is Article XIII's declaration of the Union as "perpetual," which the Constitution did not address, much to Unionists dismay some seventy years later.[115] Abraham Lincoln, in arguing for the impossibility of secession, stated,

> [W]e find the proposition that, in legal contemplation, the Union is perpetual, confirmed by the history of the Union itself. The Union is much older than the Constitution. It was formed in fact, by the Articles of Association in 1774. It was matured and continued by the Declaration of Independence in 1776. It was further matured and the faith of all the then thirteen States expressly plighted and engaged that it should be perpetual, by the Articles of Confederation in 1778. And finally, in 1787, one of the declared objects for ordaining and establishing the Constitution, was 'to form a more perfect union.'[116]

By Lincoln's account, the Union as it stood in 1861 was not born in 1787. The Constitution, in this view, is a furtherance of a Union formed by the Articles of Association in 1774 and made perpetual by the Articles of Confederation.[117] For his part, in Federalist No. 40, Madison began to make the case that the Convention was a necessary abandonment of the Articles in order to preserve their guarantee of a "national and adequate government," but in the very next paragraph, he argues that the Convention's actions can be viewed as merely an extensive amendment process within the Articles' authority:

---

112 Gregory E. Maggs, *A Concise Guide to the Articles of Confederation as a Source for Determining the Original Meaning of the Constitution*, 85 GEO. WASH. L. REV. 397, 428-29 (2018).

113 *Id.*

114 *Id.* at 429.

115 ARTICLES OF CONFEDERATION OF 1781, art. XIII.

116 *Id.*

117 Lincoln's locating the origin of the Union in the Articles of Association is an intriguing alternative to pursue and at least has the benefit of being more specific than what we shall see is Chief Justice Chase's view: that the Union formed culturally, organically over time.

> But is it necessary to suppose that these expressions are absolutely irreconcilable to each other; that no alterations or provisions in the Articles of Confederation could possibly mould them into a national and adequate government; into such a government as has been proposed by the convention? No stress, it is presumed, will, in this case, be laid on the title; a change of that could never be deemed an exercise of ungranted power. Alterations in the body of the instrument are expressly authorized. New provisions therein are also expressly authorized. Here then is a power to change the title; to insert new articles; to alter old ones. Must it of necessity be admitted that this power is infringed, so long as a part of the old articles remain? Those who maintain the affirmative ought at least to mark the boundary between authorized and usurped innovations; between that degree of change which lies within the compass of alterations and further provisions, and that which amounts to a transmutation of the government. Will it be said that the alterations ought not to have touched the substance of the Confederation? The States would never have appointed a convention with so much solemnity, nor described its objects with so much latitude, if some substantial reform had not been in contemplation.[118]

Madison may have been, to some extent, covering his bases, but he seems to genuinely argue for the latter view, in which the Constitution was within the amendatory powers of the Convention. This could be conceived of, as Madison held it, as a set of recommended amendments by an unofficial body never claimed to be vested with any authority but merely making friendly suggestions to the States.[119] Alternatively, as Akhil Amar has proposed, it could be conceived of as within the Articles' formal provisions for side deals.[120]

Amar offers an intriguing account of how this would work by contending that the Constitution might be seen as a concurrent side deal to the Articles until ratified by all states, thereby resolving any perceived conflict between the two documents.[121] In a footnote in his book, *America's Constitution: A Biography*, Amar, citing Art. VI, para. 2 of the Articles, he writes that although there seems to be no evidence that the Constitution's advocates ever advanced the argument,

> [i]t might be suggested that the proposed Constitution would merely amount to a new side alliance among nine or more of the thirteen states, and that such alliances were permissible so long as (1) the allying states lived up to all the rules of the Articles of Confederation when dealing with the remaining states, and (2) the allying states secured the blessing of the Congress under the Articles (which, presumably, they would have been able to do by so instructing their confederate delegates).[122]

Such a "side alliance" theory could hold the side alliance as lasting from the start of the convention until the ratification by thirteen states and the Articles as wholly abrogated on that date or, alternatively, hold any provisions of the Articles not in direct conflict with and not field preempted by the

118 THE FEDERALIST NO. 40 (James Madison).
119 *Id.*
120 *America's Constitution*, *supra* note 88, at 517 n.65.
121 *Id.*
122 *Id.*

Constitution to be preserved, though given the similarity of the two documents' scopes, such provisions would be scant.

Importantly for our purposes, the "within the articles and lawful" category leaves open to us a consideration that has emerged at least once explicitly but arguably one or more times implicitly in Supreme Court: whether the Articles might still be drawn upon as a basis of union and a source of law.

III. *WHITE*, *WHEELER*, AND THE ARTICLES AS LAW

At least two Supreme Court cases have treated the Articles as having legal effect, implying that they in some sense run concurrent to the Constitution: *Texas v. White* and *United States v. Wheeler*. *White*, being the more prominent, important decision and offering a clearer explication of the justices' view of the Articles' status, will be the primary focus here, but *Wheeler* also seems to take the Articles as giving substance to the Constitution's terms on a key issue, implying that they carry sufficient force to conclusively decide a constitutional question and might be in a sense "good law."

A. *Texas v. White*

The facts of *Texas v. White* are somewhat exciting for their relation to major historical events, political intrigue, and implied themes of con-artistry. In 1851, in satisfaction of the terms of the Compromise of 1850, in which Texas ceded its boundary claims north of New Mexico to be free territory under control of the federal government, the United States issued ten thousand $1,000 bonds, thereby issuing $10 million in public debt dated January 1, 1851.[123] They were coupon bonds payable to the State of Texas or bearer with interest set at five percent paid semi-annually, redeemable after December 31, 1864.[124] The terms on the face of the bonds required that no bond should be available in the hands of any holder until endorsed by the governor of Texas.[125]

At the onset of the American Civil War, on January 11, 1862, after Texas had declared itself seceded from the Union, the Texas legislature repealed the requirement of a governor's endorsement on the bonds and provided for a military board to use up to $1,000,000 of the bonds in its treasury in defense of the State.[126] In February 1862, G.W. Paschal, a Union loyalist from Texas notified the U.S. Treasury before any of Texas' bonds were sold,

123 White, 74 U.S. at 717-718.
124 *Id.*
125 *Id.* at 718.
126 *Id.* at 717-18.

and the Treasury ran a legal notice in the New York Tribune refusing to honor any bonds not endorsed by the pre-war governor, Sam Houston.[127]

Nonetheless, towards the war's end, on January 12, 1865, the military board of Texas, operating under its relaxed rules, sold one hundred and thirty-five of these bonds to George W. White and John Chiles in one transaction and seventy-six more to be deposited for them in England, for which White and Chiles contracted to deliver supplies of cotton cards and medicines in support of the war effort.[128] In 1865 and 1866, these bonds were exchanged by purchase or as security with other defendants who were party to the case.[129] Incidentally, there was also a cloud of uncertainty in the case as to whether White's and Chiles' transaction with the State may have been disingenuous. The goods for which the State had contracted were never delivered, and White and Chiles claimed that they had been destroyed in transit by disbanded troops roaming Texas after the war had ended.[130] Their loss, they claimed, was unavoidable.[131]

At the Confederacy's defeat, President Andrew Johnson appointed Union General Andrew J. Hamilton as provisional governor on June 17, 1865, and U.S. Army soldiers arrived in Texas two days later to take possession of the State and restore order under federal authority.[132] In its efforts to rebuild, a State convention in 1866 passed an ordinance to recover the bonds and authorizing the governor to take necessary measures to either recover them or compromise with their holders.[133] J.W. Throckmorton, elected governor of Texas under Texas' new constitution of 1866, authorized agents of the state to file suit directly in the Supreme Court under Art. III, 2, cl. 1, the State being a party to the suit.[134] The State's bill contended that the bonds were seized in armed hostility to the United States and sold in support of an effort to overthrow the federal government; that the recipients, White and Chiles, had failed to perform in that agreement; that the subsequent transfers to others were not in good faith and were executed despite express notice in the newspapers; that the bonds were overdue at the date of transfer; and that they had never been endorsed by any governor of Texas.[135] White contended that Texas lacked evidence, claiming that the unnumbered bonds had been destroyed by soldiers and that proof of the transaction and its terms was absent.[136]

---

127 *Id.* at 706.
128 *Id.* at 718.
129 White, 74 U.S. at 718.
130 *Id.* at 700.
131 *Id.*
132 *Id.* at 719.
133 *Id.* at 711.
134 *Id.* at 708.
135 White, 74 U.S. at 709.
136 *Id.* at 710.

Precedent to the Supreme Court's determination of whether Texas could reclaim the bonds was the standing question of whether Texas could lawfully file directly in the Supreme Court as a State. Texas filed in 1867,[137] and the Court issued its decision in 1868,[138] but Texas would not be formally readmitted to the Union by Congress until March 30, 1870.[139] White thus argued that Texas, having seceded and being at that time under military administration by the federal government, had no standing to bring the suit.[140] The Supreme Court, with Chief Justice Chase writing for a 5-3 majority, held that Texas did have standing.[141] In doing so, Chase offered a novel view on the nature of the Union and its establishment.

Chase begins by considering what it means to be a State, even apart from a union or confederation:

> It describes sometimes a people or community of individuals united more or less closely in political relations, inhabiting temporarily or permanently the same country; often it denotes only the country or territorial region, inhabited by such a community; not unfrequently it is applied to the government under which the people live; at other times, it represents the combined idea of people, territory, and government.
>
> It is not difficult to see that, in all these senses, the primary conception is that of a people or community. The people, in whatever territory dwelling, either temporarily or permanently, and whether organized under a regular government, or united by looser and less definite relations, constitute the state.
>
> This is undoubtedly the fundamental idea upon which the republican institutions of our own country are established.[142]

The State as conceived of in the Constitution, Chase thus reasoned, is a "political community of free citizens, occupying a territory of defined boundaries, and organized under a government sanctioned and limited by a written constitution, and established by the consent of the governed."[143] It is in the broader sense of a political community, Chase reasons, that the Constitution uses the term "State" in the Guarantee Clause and in its promises to protect the States against invasion.[144] In this, "a plain distinction is made between a State and the government of a State."[145]

---

137 ROBERT BRUCE MURRAY, LEGAL CASES OF THE CIVIL WAR 151 (2003).

138 Texas v. White, 74 U.S. 700 (1868).

139 Act of March 30, 1870, ch. 39, 16 Stat. 80 (admitting the State of Texas to Representation in the Congress of the United States).

140 White, 74 U.S. at 718-19.

141 *Id.* at 726.

142 *Id.* at 720.

143 *Id.* at 721.

144 *Id.; See* U.S. CONST. art. IV, § 4.

145 White, 74 U.S. at 721.

Texas, Chase noted, was admitted to the Union as a State in 1845, an act which invested the new State and its people with all of the responsibilities and duties of membership in the Union, as truly and fully as if they had been among the first thirteen at the Constitutional Convention.[146] Chase contended that the Union was not an "artificial" relation but an emergent one that grew out of common origin, mutual sympathies, kindred principles, similar interest, and geographical relations.[147] Most poignantly for our purposes, the Chief Justice went on to describe the Union as having "received definite form and character and sanction from the Articles of Confederation. By these, the Union was solemnly declared to 'be perpetual.'"[148] And by the Constitution's enactment it was made a "more perfect Union."[149] The final product: "a perpetual Union, made more perfect."[150]

From this, Chase deduced that when Texas joined the United States, it entered an indissoluble relation to all other states and was bound to guarantee its citizens a republican form of government.[151] When, on February 1, 1861, the Texas secession convention drafted an Ordinance of Secession for approval by the state legislature and a statewide referendum, Chase determined, it violated the Guarantee Clause and the Ordinance was therefore null.[152] Texas at all times remained a State within the Union, and its war against the Union was a war of rebellion, not of "conquest and subjugation.[153] Texas therefore had standing.

Pursuant to this conclusion, the Chief Justice followed an orthogonal but interesting discussion of the constitutionality of Reconstruction governments,[154] ultimately reaching the question of Texas' claim to the bonds sold by the Confederate military board of Texas.[155] As to the legal acts of Confederate governments, he concluded that those legal acts and decisions "necessary to peace and good order among citizens," including acts sanctioning and protecting marriage and domestic relations and others relating to property, wills, injuries to persons and estates, etc., that would have been valid if emanating from a lawful government were still legal but that those acts in aid of rebellion against the United States were invalid.[156] The question then was whether the sale of the bonds was in aid of rebellion. The Court held that it plainly was in aid of rebellion and that White, Chiles, and those to whom

146 *Id*. at 722.
147 *Id.* at 724.
148 *Id.*
149 *Id.*
150 *Id.*
151 White, 74 U.S. 728-30.
152 *Id.*
153 *Id.* at 724.
154 *Id.* at 727-32.
155 *Id.* at 732.
156 *Id.* at 733.

they transferred the bonds had sufficient notice of the instruments' repudiation to vindicate Texas on all counts.[157]

B. United States v. Wheeler

Wheeler is the stuff of movies. It arises from the harsh tenor of labor relations that had emerged in America by the 1910s. Nearly two-thousand miners, members of the Industrial Workers of the World (IWW), contracted by the Phelps Dodge Company and other mining operations to work in Bisbee, Arizona, struck on June 26, 1917.[158] In response, Phelps Dodge executives met with Bisbee Sherriff Harry Wheeler in conspiracy to forcibly seize all two-thousand workers (out of a town of roughly eight thousand) and deport them hundreds of miles away and abandon them there, in the desert, without food, water, or money.[159] Early on the morning of July 12, twenty-two-hundred deputies arrested every man on the list whom they could find, along with any other men who refused to work in the mines, totaling roughly two thousand persons.[160] The detainees were taken at gunpoint to a baseball stadium, where some were released in exchange for denouncing the IWW.[161] The others were loaded onto twenty-three cattle cars and transported two hundred miles over sixteen hours without food or water and unloaded at Hermanas, New Mexico at three o'clock in the morning on July 13.[162]

The local sheriff in New Mexico and the state's governor contacted President Woodrow Wilson for federal support with the relocated men, and Wilson sent U.S. Army troops to take the men to Columbus, New Mexico, where they were furnished tents until September.[163] Back in Bisbee, Sheriff Wheeler established a perimeter around the town and the neighboring town of Douglas, requiring a "passport" issued by the Douglas Chamber of Commerce and Mines to enter or exit the town and trying them before a secret sheriff's court, deporting hundreds and threatening them with lynching if they returned.[164]

In May 1918, the Department of Justice brought suit against twenty-one mining company executives along with Wheeler and other Cochise County officials, alleging conspiracy to violate § 19 of the United States Criminal Code, which prohibited injuring, oppressing, threatening, or intimidating citizens of the United States in the rights and privileges secured to them by the federal Constitution, namely to reside and remain in a state where they are

157 White, 74 U.S. at 736.
158 Philip Taft, *The Bisbee Deportation*, 13 LABOR HIST. 3, 7 (1972).
159 *Id.* at 4, 13-16.
160 *Id.*
161 *Id.*; W. Lane Rogers, *The Bisbee Deportation* 19-20 (2007).
162 Rogers, *supra* note 161.
163 Taft, *supra* note 158, at 24.
164 *Id.* at 23.

citizens and to be immune from unlawful deportation to another state.[165] The indictments mentioned no federal law, as there was no federal offense of kidnapping until the Federal Kidnapping Act of 1932.[166] The government thus relied upon the claim of an implied federal power to forbid and punish those violating § 19.[167] The defense, in turn, contended that the federal Constitution left the rights implicated "to the protection of the several states having jurisdiction."[168] The case invited notable representation, with W.C. Herron, brother-in-law of President William Howard Taft, representing the Justice Department and former Associate Justice, future Chief Justice Charles Evans Hughes arguing for the defense.[169]

Chief Justice White, writing for the majority, held that the Articles of Confederation established a uniformity of the right of citizens to peaceably dwell within their respective states and to have free ingress and egress among states.[170] States were thereby invested with an authority to forbid and punish violations of those rights. Article IV, White contended, did not assign protection of this right to Congress but instead placed direct limitations on state power to prohibit discriminatory behavior, its text stating clearly that

> The better to secure and perpetuate mutual friendship and intercourse among the people of the different states in this Union, the free inhabitants of each of these states, paupers, vagabonds and fugitives from justice excepted, shall be entitled to all privileges and immunities of free citizens in the several states, and the people of each state shall have free ingress and egress to and from any other state. . .[171]

The Constitution, in Art. IV, § 2, the Chief Justice reasoned, intended

> to preserve and enforce the limitation as to discrimination imposed upon the states by Article IV of the Confederation, and thus necessarily assumed the continued possession by the states of the reserved power to deal with free residence, ingress, and egress, cannot be denied for the following reasons: (1) because the text of Article IV, § 2, of the Constitution makes manifest that it was drawn with reference to the corresponding clause of the Articles of Confederation, and was intended to perpetuate its limitations, and (2) because that view has been so conclusively settled as to leave no room for controversy.[172]

---

165 *Id.* at 30; Wheeler, 254 U.S. at 292. Section 19 is incorporated today as 18 U.S.C. § 241, which defines the offense as "two or more persons conspir[ing] to injure, oppress, threaten, or intimidate any person in any State, Territory, Commonwealth, Possession, or District in the free exercise or enjoyment of any right or privilege secured to him by the Constitution or laws of the United States."

166 18 U.S.C. § 1201.

167 Wheeler, 254 U.S. at 292.

168 *Id.*

169 *Id.* at 281.

170 *Id.* at 293-94.

171 ARTICLES OF CONFEDERATION OF 1781, art. IV.

172 Wheeler, 254 U.S. at 294.

White goes on to discuss how a description of the original Article IV's limitation being "preserve[d] and enforce[d]" by the Constitution can easily be read as implying the preservation of the Articles' force as a constitutional background.[173] White seems to be saying more than many judges and justices before and since who have cited the Articles as interpretive references that can give Constitutional interpreters a convenient picture of possible original meanings. Indeed, in a line of reasoning harkening to Chief Justice Chase, he seems to conceive of statehood, union, and fundamental rights as tracing to some undefined point predating both founding documents:

> In all the states, from the beginning down to the adoption of the Articles of Confederation, the citizens thereof possessed the fundamental right, inherent in citizens of all free governments, peacefully to dwell within the limits of their respective states, to move at will from place to place therein, and to have free ingress thereto and egress therefrom, with a consequent authority in the states to forbid and punish violations of this fundamental right.[174]

White cites *Corfield v. Coryell*[175] and *The Slaughterhouse Cases*,[176] and in both of these precedents there is interpretive use of the Articles, but in *Wheeler*, as in *White*, there is the shade of something more than interpretive reference; there is the indication of a constitutional backbone that does more than clarify language but introduces forceful points of its own that, when applied, are capable of binding the power that public officials can wield over private rights.

IV. UNANIMITY, HOLDOUTS, AND WASHINGTON'S FIRST TERM

One implicit defense of the Convention's authority seems to have been that whatever the nature of the Convention, as long as the final product was ratified unanimously by the States, nothing had been usurped. The point at which any shade of that argument was abandoned came when the threshold for adoption of the Constitution was lowered from unanimity to a mere nine out of thirteen states.[177] This question is all the more poignant when operating under an assumption that the Convention was within the authority of the Articles. A theory holding that the Convention abandoned the Articles easily avoids the issue of their unanimity requirement,[178] but one premised on the view that the Convention was within the Articles' authority, even one as plausible as the "side alliance" theory, cannot get around the conclusion that the

---

173 *Id.*

174 *Id.* at 293.

175 4 Wash. C.C. 371, 380-81 (1823).

176 16 Wall. 36, 76 (1873).

177 Gary Lawson & Guy Seidman, *When Did the Constitution Become Law?*, 77 NOTRE DAME L. REV. 1 (2001).

178 *Id.* at 6; *see also* Vasan Kesavan, *When Did the Articles of Confederation Cease to Be Law?*, 78 NOTRE DAME L. REV. 35 (2002).

Constitution would not take effect until it had been ratified by all thirteen states.

The reluctance of Rhode Island and North Carolina to ratify the Constitution and join the Union for many months after the Convention's end makes these questions more than mere abstract brain teasers. I argue here that they are made tangible by the fact that the new federal government would commence operation before these two states had ratified. As a result, one's preference among theories of the Convention leads to different answers about the legality of federal action before unanimity was secured.

A. *Rhode Island and North Carolina's Late Ratifications*

Of the thirteen original states, eleven ratified the Constitution prior to the elections of 1788.[179] Federalists in the two holdout states, North Carolina and Rhode Island, would require more time and repeated attempts before securing a favorable vote.[180] North Carolina's is the less storied effort, with most of the evident controversy there turning on the guaranteed inclusion of a Bill of Rights before they would ratify.[181] Rhode Island, where ratification was not only very late but by the smallest margin, was a different story. Rhode Island was, from the start, the wayward state, long seen as frustratingly prone to dissent.[182] As a result, it gathered a collection of nicknames and epithets: "the perverse sister," "an evil genius," the "quintessence of villainy," and, of course, "Rogue Island."[183] Its local Country Party had won a sweeping victory in 1786, opposing the expansion of the national government for fear of a national tax, meanwhile advocating for greater reliance upon inflationary monetary policy as a tool of public finance.[184] The state legislature printed one-hundred-thousand pounds worth of paper currency in a month, generating rampant inflation and making it a cautionary tale to other states.[185]

As the Convention debates neared a close, a letter was received from Governor Collins of Rhode Island, which had never sent a delegate to represent it, presenting the state's various points of contention with both the Convention's purpose and structural propriety:

> [A]s a legislative body, we could not appoint delegates to do that which only the people at large are entitled to do. By a law of our state, the delegates in Congress are chosen by the suffrages of all the freemen therein, and are appointed to represent them in Congress; and for

179 *Id.* at 36 n.9.
180 *Id.*
181 *Id.*
182 Payne, *supra* note 10.
183 *Id.*
184 *Id.*
185 *Id.*

> the legislative body to have appointed delegates to represent them in convention, when they cannot appoint delegates in Congress (unless upon their death or other incidental matter,) must be absurd; as that delegation in convention is for the express purpose of altering a constitution, which the people at large are only capable of appointing the members.[186]

The people of Rhode Island, Collins wrote, had "not separated themselves from the principles" of the Articles, and they would need further guarantees of limitations on federal powers before they could ratify what the Convention was producing.[187] Between 1787 and 1790, Rhode Island would make eleven attempts at ratification without success.[188]

In its responses, Congress finally answered that lingering question of what its policy would be toward those states that did not ratify. Its approach entailed both gentle and hard diplomacy alike. On the gentle side, Congress still allowed Rhode Island's delegates and those of North Carolina, which was similarly reticent to ratify, to take their seats.[189] Their voting powers were not nullified by their failure to ratify.[190] And a number of Rhode Island's listed grievances with the Constitution, though not uniquely its own, would be addressed by the passage of the Bill of Rights.[191] On the hard diplomacy side, it threatened to treat Rhode Island as a foreign nation and impose tariffs on its exports.[192] In January 1790, Rhode Island persuaded Congress to extend the then-expiring deadline Congress had given to it until March.[193] But in March, another convention came and went without a vote.[194] On May 11, the U.S. Senate debated a bill to not only prohibit commerce with Rhode Island but authorize the President to demand restitution from Rhode Island for its $27,000 share in the national debt from the Revolutionary War.[195] Finally, on May 18, 1790, the Senate passed the bill prohibiting any commercial intercourse with Rhode Island.[196] Before the bill could pass the House (where it had considerable support and was sure to do so), Rhode

---

186 *Letter from the Governor of Rhode Island to the President of Congress (Sept 15, 1787) in* 10 RECORDS OF THE COLONY OF RHODE ISLAND AND PROVIDENCE PLANTATIONS IN NEW ENGLAND, 1784-1792, 259 (John Russell Bartlett ed., 1865).

187 Payne, *supra* note 10.

188 *Id.*

189 Ackerman & Katyal, *supra* note 64, at 524.

190 *Id.*

191 *See generally Amendments Proposed By the Rhode Island Constitution (Mar. 6, 1790) in* THE ANTI-FEDERALIST PAPERS AND THE CONSTITUTIONAL CONVENTION DEBATES, 225–26 (Ralph Ketcham ed., 1986).

192 Payne, *supra* note 10.

193 *Id.*

194 *Id.*

195 *See* Jonathan White, *North Carolina and Rhode Island: The 'Wayward Sisters' and the Constitution*, THE IMAGINATIVE CONSERVATIVE (Feb. 15, 2015), https://theimaginativeconservative.org/2015/02/north-carolina-rhode-island-wayward-sisters-constitution-jonathan-white.html.

196 Payne, *supra* note 10.

Island succumbed to the pressure of its merchants and ratified the Constitution on May 29, with a vote of 34-32.[197]

B. *Implications for Washington's First Term*

The concerning implication of these late ratifications arises from the unanimity principle in the Articles. Even if one takes the view that the Constitution was adopted within the authority of the Articles and that the eventual unanimity of its ratification by those thirteen states party to the Articles assures us of the Constitution's validity under law, we are still left with questions as to its validity prior to the unanimous ratification. More specifically: if the Constitution is within the Articles and lawful, it could only take effect once all thirteen states had ratified. Rhode Island's ratification not coming until May 29, 1790, means that the Constitution would not have properly come into effect until nearly halfway through George Washington's first term as president, not only rendering his election void but making any federal legislation signed prior to that date (and, arguably, any legislation signed in Washington's first term) likewise void. The only means of maintaining the "within the Articles and lawful" position and evading this conclusion would be to imply that there was something implicitly retroactive in Rhode Island's and North Carolina's ratifying instruments, but that would seem to be a stretch. Retroactivity is generally not to be presumed unless stated, and nothing in the Framers' debates would suggest the Constitution being so.

To be fair, other popular accounts of the Constitution's effectiveness date have problems of their own. The nine-out-of-twelve approach taken in Article VII would place effectiveness in the summer of 1788, but as discussed above, it merely assumed away the unanimity requirement of the Articles, taking a revolutionary approach rather than a legalistic one. In contrast, Justice Marshall's opinion in *Owings v. Speed* suggested that the Constitution became law when Congress first sat on March 4, 1789.[198] This account has attained status as the conventional wisdom on the subject but suffers from a glaring flaw in its inability to account for how Congress got there on March 4 from the elections of 1788. If the Constitution was not effective until then, there would presumably have been no constitutional law governing the federal elections of 1788, which, though the issue has never been raised in court, seemingly cannot be true.

The consequences of the "within the Articles and lawful" theory being correct, however, are much more serious. If true, it would require us to hold as void all legislation emerging from Washington's first term.[199] That

197 *Id.*

198 18 U.S. (5 Wheat.) 420, 422 (1820).

199 As noted above, there is also a weaker version by which one could say that any legislation Washington signed after Rhode Island ratified would be valid, leaving open any legislation between May 29,

includes, in notable part, the creation of the First Bank of the United States, the United States Mint, the U.S. dollar, the Tariffs of 1789 and 1790 (the latter of which spurred the Whiskey Rebellion), and, of course, the Judiciary Act of 1789. It is not difficult to imagine the consequential invalidations of all manner of government activities that would flow from those acts being rendered nullities. Thus, we find ourselves at a fascinating impasse: either we accept, once and for all, that the Articles were usurped in revolutionary fashion at the Convention and we set aside all possibility of reconciling the actions of the Framers with the Articles' authority, up to and including any more recently extolled views holding that the Articles form a backbone of continuity underlying the Constitution,[200] or we accept the unconstitutionality of all legislation issued by the first two Congresses and let the dominoes fall from there. The Court, as discussed above, has at least twice suggested that the Articles might have some continuing force.[201] This conclusion suggests that to the extent that those cases' resolutions depended upon a view of the Articles as still carrying legal force, they must be incorrect.

CONCLUSION

In the end, it seems that in the duel between the Formal Theory and the Realist Theory, the Realist Theory leads to a better, less troubling explanation of the nature of the Constitution and the origins of federalism. The Formal Theory, in which the Constitution truly arose from the Articles' procedures to amend, is, of course, not wrong in the sense of being self-contradictory or clearly refuted by some immutable points of fact. Rather, it is "wrong" only in the legalistic sense of leading to implications so great and disruptive that prudence would seem to demand that we avoid that path lest we find ourselves in a deep constitutional quagmire. For those more extremist friends willing to accept the bitter pill of George Washington's first term and all legislation signed during those four years being illegitimate, we can only say, "go boldly!" There are certainly points to be scored there for those still passionate about the Jacksonian struggle against national banking and surely for others. This article has hopefully persuaded its reader, however, that in holding the Articles as having a continued background force of law, we must accept the bitter with the sweet.

1790, and Washington's second inauguration on March 4, 1793. It seems difficult, however, to argue convincingly that an unconstitutionally elected president could gain constitutional legitimacy halfway through his first term despite the election that brought him there having been unconstitutional. I thus set aside this option.

200 *See generally* Texas v. White, 74 U.S. 700 (1868); United States v. Wheeler, 254 U.S. 281 (1920); *see, e.g.*, Abraham Lincoln, *First Inaugural Address (Mar. 4, 1861) in* THE AVALON PROJECT, https://avalon.law.yale.edu/19th_century/lincoln1.asp; Maggs, *supra* note 112, at 429 (finding at least partial merit to the argument); *America's Constitution*, *supra* note 88, at 517 n.65 (neither endorsing the view nor rejecting it; merely considering its merits).

201 *See generally* Texas v. White, 74 U.S. 700 (1868); United States v. Wheeler, 254 U.S. 281 (1920).

Further, incorrect though we find them to be, we must be grateful for cases like *White* for highlighting the issue and forcing us, as well as authors before us, to consider these questions more deeply and for, in the course of their reasoning, expounding upon the question, fleshing out the issue into refutable propositions, and casting greater light on how we should understand the founding and the nature of the Union. Ultimately, we find the concerns of some Framers that they were exceeding the powers granted by the Articles to be not only respectable but valid; they were indeed acting ultra vires as far as the Articles went. On the other hand, the case, as argued by Madison, that their actions were still wholly in accord with natural law and the law of nations and that to do otherwise, to persist under the Articles, would be an abandonment of the duties required of any legitimate government, is perfectly plausible. James Madison's careful balancing of the question, never outright declaring his intent to usurp the Articles but never denying the revolutionary nature of their project was probably just the right balance of diplomacy and vision needed to make it succeed. One can make a variety of arguments, as have already been made, that the Articles were already nullified by non-observance or ineffectuality or that alternative traditions in international law justified the Articles' neglect. However, as revealed by the Convention's records, in the Framers' own views, they were proceeding on unsure footing. Nonetheless, they proceeded. In that act, we might say, the Founders rebelled twice: the first time against a faraway king, and the second time against themselves.