Causal Narratives and Constitutional Scrutiny

JEFF LINGWALL AND MICHELLE VOS*

Abstract

Courts engaged in constitutional scrutiny analysis often consider statistical evidence to link government ends and means. Courts and litigants generally rely on shallow reasoning when applying this evidence, particularly when causal relationships are at issue. We argue that courts should increasingly adopt, and adapt, elements of causal reasoning developed in the social sciences. We propose a flexible framework for applying causal evidence in scrutiny analysis that parallels traditional tiers of constitutional scrutiny. This framework offers a Socratic-style narrative approach to causal evidence in scrutiny analysis—one based on qualitative concepts that do not require specialized statistical training to apply. This expanded legal toolkit gives needed nuance to causal reasoning and increased alignment between evidence and tiers of scrutiny. Applying this framework to ongoing and historic cases suggests new lines of evaluation and inquiry that clarify evidentiary requirements and support for constitutional analysis.

*Jeff Lingwall, JD, PhD is an Assistant Professor of Legal Studies at Boise State University; Michelle Vos, JD is a Lecturer in Legal Studies at Boise State University. We thank attendees at the Academy of Legal Studies in Business 2020 Conference, the 2021 Data Law and Ethics Workshop, and Michail Fragkias for helpful comments. Shayla Hinson and Joseph Hubble provided able research assistance. We thank the *Law Review* for careful editing. Any errors are our own.

> *It is unrealistic to expect . . . members of the judiciary . . . to be well versed in the rigors of experimental or statistical technique.*[1]
> *Craig v. Boren*

> *[C]ausality is not mystical or metaphysical. It can be understood . . . . The prerequisites are startlingly simple, the results embarrassingly straightforward*.[2]
> Judea Pearl

I. INTRODUCTION

Causality permeates the law.[3] Yet despite the centrality of causal relationships to legal analysis of nearly every sort, legal

1. 429 U.S. 190, 204 (1976).

2. JUDEA PEARL, CAUSALITY: MODELS, REASONING, AND INFERENCE xvi, 427 (2d ed. 2009).

3. Causality is central to, for example, tort law, criminal law, civil procedure, and contracts. *E.g.*, RESTATEMENT (THIRD) OF TORTS § 26 (AM. LAW. INST. 1998) ("Tortious conduct must be a factual cause of physical harm for liability to be imposed. Conduct is a factual cause of harm when the harm would not have occurred absent the conduct."); Whiteley v. Philip Morris Inc., 11 Cal. Rptr. 3d 807, 857–58 (Cal. Ct. App. 2004) ("'Causation' is an essential element of a tort action. Defendants are not liable unless their conduct . . . was a 'legal cause' of plaintiff's injury.") (quoting WILLIAM F. FLAHAVAN, ZERNE P. HANING, DORIS CHENG, & KENNETH E. WRIGHT , CALIFORNIA PRACTICE GUIDE: PERSONAL INJURY ¶ 2:979 (2003)); Mickens v. State, 881 N.Y.S.2d 854, 867–68 (N.Y. Ct. Cl. 2009) ("It is a fundamental principle of tort law, that a right to recover damages arises only when the defendant breaches a duty it owes to the injured party and that breach of duty causes the injury for which compensation is sought."); United States v. Hayes, 589 F.2d 811, 821 (5th Cir. 1979) ("A fundamental principle of criminal law is that a person is held responsible for all consequences proximately caused by his criminal conduct. Thus, where events are foreseeable and naturally result from one's criminal conduct, the chain of legal causation is considered unbroken and the perpetrator is held criminally responsible for the resulting harm."); United States v. Spinney, 795 F.2d 1410, 1415 (9th Cir. 1986) ("A basic tenet of criminal law is that the government must prove that the defendant's conduct was the legal or proximate cause of the resulting injury."); Siegel v. U.S. Dept. of Treasury, 304 F. Supp. 3d 45, 59 (D.D.C. 2018) ("Establishing standing requires a showing of three elements—injury in fact, causation, and redressability—which together constitute the 'irreducible constitutional minimum of standing.'") (quoting Lujan v. Defs. of Wildlife, 504 U.S. 555, 560 (1992)); Nat'l Mkt. Share, Inc. v. Sterling Nat'l Bank, 392 F.3d

reasoning often treats causal analysis in shallow ways relative to the sciences.[4] Particularly when causal claims are based on social scientific evidence, courts and litigators may be dismissive of the implications of causal analysis or struggle to precisely express or analyze the assumptions underlying causal claims.[5] This weakness in causal

520, 525 (2nd Cir. 2004) ("Causation is an essential element of damages in a breach of contract action; and, as in tort, a plaintiff must prove that a defendant's breach *directly and proximately caused* his or her damages.") (citing Wakeman v. Wheeler & Wilson Mfg. Co., 4 N.E. 264 (N.Y. 1886)).

4. In many cases, this is because rational litigators recognize that the costs of performing high-level scientific causal analysis exceed the benefits. In an auto accident, whether a driver caused another's injuries after speeding through a red light and colliding with their vehicle is typically an uncomplicated inquiry. *E.g.*, Smithers v. State Farm Mut. Auto. Ins. Co., 286 So. 2d 433, 435–38 (La. Ct. App. 1973) (holding that defendant's negligence in running a red light was the sole proximate cause of the accident and that the accident caused plaintiff's lumbar back injuries). In other cases, with more complex relationships between alleged cause and effect, causal claims may be intensely litigated. In a toxic tort case, whether a drug caused long-term injuries or whether those injuries were due to other factors may be a key question subject to extended and conflicting expert analysis. *See, e.g.*, McClain v. Metabolife Int'l, Inc., 401 F.3d 1233, 1251–52 (11th Cir. 2005) (holding that in a toxic tort case, plaintiff's expert opinion on whether the substance in question had a toxic effect that caused plaintiff's injury did not comply with standards utilized by experts in the field of toxicology, and therefore admitting his testimony was an error); Smith v. Ortho Pharm. Corp., 770 F. Supp. 1561, 1569 (N.D. Ga. 1991) (holding that conflicts in expert testimony as to causality go to the jury to determine the weight rather than the admissibility of expert opinions, but that the court must determine whether the data supporting an expert opinion is trustworthy).

The potential for such debate and uncertainty about causal mechanisms is especially acute when courts consider social scientific evidence. Causal evidence in the social sciences often lacks the intuitive causal proof available in, for example, randomized controlled trials performed to consider the safety and efficacy of a prescription drug. At the same time, many advances in social scientific fields in recent decades, particularly in econometrics, have opened up vast areas to potentially persuasive causal narratives. The line one might hear that "there is no causality without experimentation" is a shallow heuristic that fails to account for potentially strong non-experimental evidence based on reasonable assumptions and well-established methods for establishing causal connections in social scientific, non-experimental data. *See infra* Section II.C (discussing proof of causality in non-experimental settings).

5. While courts may struggle with how best to consistently treat such statistical-based evidence, the *practice* of law and statistics hold fundamental parallels. Both are disciplines governed by rules (e.g., the Constitution or the Central Limit Theorem),

analysis leads to poor legal reasoning, which is especially troubling when used in constitutional scrutiny cases in which fundamental rights are balanced against strong government interests in protecting health or safety. When courts seek to balance such important competing interests, they benefit from using language and frameworks that account for statistical reasoning and causal mechanisms in precise ways.[6]

As a starting point for this discussion, in this Article we examine causal analysis in a particularly fraught area: constitutional scrutiny cases.[7] We argue that *causal* reasoning in these cases is surprisingly *casual*, often entirely divorced from developments in causal analysis in the sciences. The language of causality is often either missing or used in constitutional legal argument in shallow ways. In particular, legal argument often reflects little grasp of fundamental issues regarding how causal evidence is built and interpreted and the role of foundational assumptions in that evidence. For example, judges and attorneys may focus their analysis on the significance of p-values[8] without

but the application of those rules to solve real-world problems remains an art guided and cabined by the rules, rather than being strict expressions of them. For example, the text of the Constitution does not tell litigators how dignity or privacy interests may interact with a particular case, just as the Central Limit Theorem does not dictate to statisticians when it can be plausibly invoked in any given situation involving real-world data.

6. Courts may give short shrift to causal arguments at least in part because courts—like other boundedly rational actors—rely on heuristics when deciding cases, and few satisfactory legal heuristics exist for causal reasoning in the constitutional scrutiny context. One aim of this Article is to begin providing such heuristics, such as in *infra* Section II.C, which provides a rough alignment of types of research with the strength of causal claims, and *infra* Section IV.B, which discusses qualitative approaches to discussing the assumptions behind quantitative causal claims.

7. The implications from the discussion apply broadly across various areas of the law that frequently invoke causal principles or causal inference, such as for expert testimony in *Daubert*. *See* Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993). A focus on constitutional scrutiny cases here allows us to narrow the discussion to a particular context and strain of causal reasoning. By exploring this particular area in some depth, we hope to suggest the beginnings of a foundation that may extend to other areas of the law.

8. In frequentist statistics, a p-value is the probability of observing results as, or more extreme than, the observed value, assuming the truth of a null hypothesis (what one would assume true in the absence of further evidence). Assuming causal assumptions are met, the p-value will refer to the statistical significance of whether an observed relationship (i.e., the causal evidence offered during litigation) is due to

questioning whether causal mechanisms are needed or whether the assumptions behind the causal claims reflected in those p-values are reasonable.[9] When this evidence is misinterpreted or applied without correct and consistent frameworks, the result is that statistical evidence becomes whatever a court wishes it to be. These statistical shortcuts can be misdirection away from underlying weaknesses in evidentiary foundations for constitutional claims.

Recent strict scrutiny litigation over the right to bear arms on university campuses in Missouri provides a telling example. The Missouri Constitution contains an expansive clause guaranteeing "the right of every citizen to keep and bear arms" and requiring courts to apply strict scrutiny analysis when considering a restriction on those rights.[10] The Missouri Supreme Court takes a generally dim view on using statistical evidence in that analysis, finding that, for example, "statistics do not bear on . . . constitutional analysis because they prove nothing about the law's design . . . . This is merely one example of why the ever-changing body of science and statistics is ill-suited to constitutional analysis."[11] Despite this, statistical evidence became central to whether a prohibition on firearms on university campuses within the University of Missouri System violated the Missouri Constitution.

University policy prohibited the possession or discharge of firearms on university property, with some exceptions.[12] The State

random chance (i.e., no causal effect) or an actual relationship (i.e., causation). P-values are expressed as a number between 0 and 1. Colloquially speaking, and given that causal assumptions are satisfied, the smaller the p-value, the more likely there is evidence of statistically-significant results, and hence causation. Typically, a p-value of less than 0.05 is considered strong evidence, and a p-value greater than 0.05 is considered not to be statistically significant, and hence provides little evidence of causality. If causal assumptions are *not* met, the significance of the p-value will have little bearing on the existence of a causal relationship, and instead may be *misleading*.

9. For further discussion of p-values, see *infra* Section IV.B.6 and accompanying text.

10. MO. CONST. art. I, § 23 (amended 2014). Outside Missouri, courts have applied an intermediate level of scrutiny to Second Amendment questions. *E.g.*, People v. Chairez, 104 N.E.3d 1158, 1163 (Ill. 2018).

11. State v. Merritt, 467 S.W.3d 808, 814 n.6 (Mo. 2015) (en banc).

12. *Facilities and Equipment Management: Collected Rules and Regulations*, UNIV. OF MO SYS. ch. 110.010, https://www.umsystem.edu/ums/rules/collected_rules/facilities (last visited Mar. 12, 2022).

challenged these regulations in *State v. Mun Choi* and both parties brought statistical evidence to bear.[13] The State (trying to show the *lack* of a statistical relationship between firearm restrictions and crime), employed an expert who showed evidence from four statistical models. Each failed to show the lack of a statistically significant relationship between the policy and reduced crime levels.[14] The trial court read the State's evidence in the *opposite* direction: rather than finding the *lack* of statistical significance in multiple studies as evidence that there was no relationship between government means (the policy) and ends (reducing crime), the court noted that such a relationship was "wholly unnecessary" and that—regardless—such studies could "imply causation" of the very facts which they failed to find significant.[15] On appeal, the court went even further: it took this lack of statistical significance and found it "credible, competent evidence" of a causal relationship it was introduced *to disprove*.[16]

Because of examples like this, we argue that courts and litigators should reconsider the causal paradigms they bring to causal evidence

13. Plaintiff State of Missouri's Petition for Declaratory and Injunctive Relief at 7–10, State v. Middleton, No. 16BA-CV02758 (Mo. Cir. Ct. Aug. 16, 2016) (on file with author).

14. *See* State v. Mun Choi, No. 16BA-CV03144, slip op. at 8–10 (Mo. Cir. Ct. Nov. 18, 2019) (on file with author). The first model compared violent crime levels at UC Boulder (UC) with Colorado State University (CSU), as UC allowed concealed carry beginning in 2012 and CSU in 2003. Violent crime increased at UC, but not at a statistically significant level. *Id.* at 11. The second compared statistics at UC Boulder to seven other Colorado universities with similar results. *Id.* at 12. The third compared statistics across eight states and showed an insignificant but positive correlation between allowing concealed carry on campus and number of deaths from firearms. *Id.* The fourth compared violent crime between universities in Missouri with different policies on locking firearms in vehicles. It also did not show a statistically significant relationship between the relevant policy and crime levels. *Id.* at 12–13. The University employed an expert who brought statistical evidence from a state-to-state comparison showing that violent crime increased by small amounts each year after passage of right to carry laws, and the effect became significant after seven years. *Id.* at 14–15.

15. *Id.* at 26–27.

16. *State v. Mun Choi*, No. WD834274, at 24 (Mo. Ct. App. Feb. 2, 2021) (on file with author). We examine these decisions in detail in *infra* Sections III.C (discussing the reasoning behind the courts' conclusions) and IV.C (applying our causal framework to the decisions).

in the constitutional context.[17] We first argue that causality *should matter* when considering the link between government action and outcome. The heart of science is establishing causal relationships, because understanding those relationships enables manipulation of the natural world to promote human welfare. Public policy is similar: the stronger the evidence that a policy will cause its desired effect, the sounder the justifications for its advancement, and if little evidence exists that the policy will have a causal effect (i.e., achieve its aims), one wonders at the reasons for its establishment.

In the realm of policy affecting constitutional rights, this worry is particularly acute. Without evidence that the rights-infringing action will actually cause a government purpose to advance, courts should justifiably question whether those rights should be infringed for even a compelling government purpose. Based on this reasoning, we show that common treatment of social scientific evidence in constitutional scrutiny cases is so shallowly applied as to do a disservice to the compelling need for such causal relationships. A paradigm-shifting change to greater precision in causal analysis—one that allows for a spectrum of causal evidence to be evaluated by courts based on the assumptions behind causal claims—would result in an expanded legal toolkit when discussing scientific evidence. This would provide needed nuance in how courts consider causal evidence and an increased alignment of that evidence with constitutional decision making.[18]

To facilitate this, Section II explores the idea of causality and then offers a brief history of causal thinking, ranging from philosophy to statistics and economics. It then shows how causal evidence can be evaluated with more precise language based on the assumptions behind causal claims.[19] Section III begins with reviewing the history of

17. Legal argument often draws on foundational paradigms or frameworks. These paradigms frame how cases are argued and decided, often without much thought to the validity of the paradigm itself. For example, when a court speaks of the idea of justice, one might argue justice for whom, but not that courts should be justice-seeking entities. Or when attorneys make arguments appealing to economic efficiency, one might argue that efficiency must be balanced with other interests, but not generally the paradigm that efficiency is good.

18. *See infra* Section IV.A (discussing how levels of scrutiny may correspond to levels of causal rigor).

19. *Infra* Sections II.C and III.C (discussing terminology and reasoning behind causal thinking in social scientific settings).

scrutiny analysis, how courts have treated causal evidence, and then notes how specific court decisions have shaped causal reasoning in the constitutional scrutiny context.[20] Section IV then argues that causality and causal arguments should form a foundational aspect of constitutional scrutiny analysis, even in cases in which courts ultimately conclude they do not need causal evidence to proceed. The narrative approach we advocate does not require mathematical expertise, rather, it relies on lines of Socratic-style questioning that will probe causal fallacies using reasoning familiar to courts and litigators, bolstered by a plain-language explanation of how a variety of statistical models approach causality.[21] The Section then applies elements of this framework to examine the potential causal reasoning in existing constitutional scrutiny cases, showing how precision and nuance in causal thinking could benefit legal argument.

To conclude, we argue that causal thinking is not best left to experts alone, although experts may be needed to bring and interpret causal evidence. The foundational concepts of causal thinking and their common pitfalls are well within the grasp of practicing attorneys and courts. To omit these principles when considering whether constitutional rights should be limited does those rights a disservice. Whatever the conclusion of a court in a particular constitutional issue, that conclusion is strengthened by incorporating careful causal reasoning.

II. EVALUATING CAUSAL EVIDENCE FROM SOCIAL-SCIENTIFIC SOURCES

This Section considers the idea of causal thinking and how the development of statistical evidence interplayed with developments in that thinking. It then draws on modern concepts in statistical reasoning to explore the spectrum of potential assumptions behind statistical-based causal evidence. This development is instructive to lay a

20. *Infra* Section III.C (discussing, e.g., the Supreme Court's requirement of causal evidence in *Brown v. Ent. Merchs. Ass'n*, 564 U.S. 786 (2011)).

21. We introduce a variety of these models in *infra* Section II.C, and then discuss how one might probe the reasoning behind those models in *infra* Section III.B.

foundation for discussing causal thinking in the context of constitutional reasoning.[22]

A. Theories of Causality

When one speaks of something "causing" something in everyday conversation, deep philosophical thinking or statistical terminology is not typically at issue.[23] This is generally true for use of the word "cause" in legal situations as well. For example, while "causation" is central to many legal doctrines, the term itself is often undefined.[24] As we will be considering causation in some detail, it is worth considering briefly various ways "cause" and its variants have been defined.

A colloquial definition of "cause" is "something that brings about an effect or a result."[25] Legal definitions are similar. "Cause" is

22. Albert Einstein famously noted that the "[d]evelopment of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance)." *See* Julian C. Jamison, *The Entry of Randomized Assignment into the Social Sciences*, (World Bank Grp., Working Paper No. 8062, 2017), https://openknowledge.worldbank.org/bitstream/handle/10986/26754/WPS8062.pdf?sequence=1&isAllowed=y.

23. There are, of course, exceptions, such as when popular culture explores the nature of cause and effect in the context of free will. *E.g.*, THE MATRIX RELOADED (Warner Bros. 2003) ("You see, there is only one constant, one universal, it is the only real truth: causality. Action. Reaction. Cause and effect . . . . Causality. There is no escape from it, we are forever slaves to it. Our only hope, our only peace is to understand it, to understand the why.").

24. Many legal doctrines will *break down* causality into various components, such as the difference between but for and proximate causation in tort law. *See, e.g.*, Powers v. Hamilton Cnty. Pub. Def. Comm'n, 501 F.3d 592, 608 (6th Cir. 2007) ("Cause in fact is typically assessed using the 'but for' test, which requires us to imagine whether the harm would have occurred if the defendant had behaved other than it did."); *id.* at 609 ("Proximate-cause analysis is a kind of line-drawing exercise in which we ask whether there are any policy or practical reasons that militate against holding a defendant liable even though that defendant is a but-for cause of the plaintiff's injury.").

25. *Cause*, MERRIAM-WEBSTER DICTIONARY, https://www.merriam-webster.com/dictionary/cause (last visited Mar. 26, 2022).

"[t]hat which produces an effect; whatever moves, impels, or leads."[26] These simple definitions capture the essential idea of cause as something that produces or brings about something else. For example, when a batter hits a pitched baseball, resulting in the ball changing directions and moving towards the outfield, we say the batter caused the ball to change direction. Or, when a billiard-ball is struck by a cue stick, we say the action by the player caused the ball to move. For situations more complex than "billiard-ball causality," the law provides additional nuance, such as distinguishing between *legal* cause and *proximate* cause, or including caveats such as "caused or contributed to," "last or nearest cause," and so on.[27]

In the sciences, the need for rigorous mathematical definitions for different circumstances leads to even more variation. In one view, most "causes" should generally be referred to as "inus conditions," where an "inus condition [is] 'an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition.'"[28] Consider whether a match has caused a forest fire in this context. The match requires oxygen, and so is "insufficient" on its own to begin a fire. The match is "non-redundant" if there is not another source of flame to trigger the fire. The match, combined with oxygen and fuel, is "sufficient" to begin the fire, but as other sources of flame exist, is "unnecessary" to do so.[29] The concept of an inus condition highlights that events may have multiple causes, like the legal differentiation of concepts such as "intervening cause" and "superseding cause."

To complicate matters further, causality is often not as deterministic as hitting a baseball, striking a billiard ball, or causing a forest fire. Often, we speak of causality in terms of probabilities, and when we say that something causes something else, we actually mean that it

26. *What is Cause?*, THE LAW DICTIONARY, https://thelawdictionary.org/cause/ (last visited Mar. 12, 2022); *see also Cause*, CORNELL L. SCH.: LEGAL INFO. INST., https://www.law.cornell.edu/wex/cause (last visited Mar. 12, 2022) ("Usually describes the reason something happens.").

27. *See, e.g.*, CSX Transp., Inc. v. McBride, 564 U.S. 685, 689 (2011) (noting or discussing each of these variants on the idea of causality).

28. WILLIAM R. SHADISH, THOMAS D. COOK, & DONALD T. CAMPBELL, EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGNS FOR GENERALIZED CAUSAL INFERENCE 4 (2002) (quoting J.L. MACKIE, THE CEMENT OF THE UNIVERSE: A STUDY OF CAUSATION 62 (1974)).

29. *See id.*

merely increases the probability of an event occurring.[30] For example, turbulence on an airline flight is a probabilistic event, with many potential inus conditions related to atmospheric phenomena. We might say a storm has "caused" turbulence, when more precisely we mean that the storm has increased the probability of turbulence occurring, and then the passengers suffered through the results of that probabilistic event.

Finally, scientists often find it useful to define causality in precise mathematical terms. This is not to say that causality itself is always a mathematically precise concept, but that scientific models, as they make simplifying assumptions about the world, rely on precise definitions of causality in their formulation. For example, one widely used model referred to as the "Rubin causal model" considers the effect of potential, if unobserved, outcomes. When studying the outcome of an emergency room (ER) visit, we only observe a single outcome for each individual—what occurred to them in reality—rather than the multiple possible outcomes that *could have* occurred for that patient. If the researcher can model the *potential* outcomes in a mathematical fashion, this framework can help when trying to disentangle the effect of selection bias on outcomes. For example, when studying health outcomes after ER visits, statistical evidence is complicated by the fact that the sicker one is, the more likely one is to seek treatment at the ER, and hence the more likely they are to have a negative outcome.[31] By mathematically grappling with unobserved potential outcomes (such as what would have happened had a sick individual who went to the ER *not* gone), researchers can more clearly lay out the assumptions on which causal claims can be made.[32]

30. *Id.* at 5.

31. This discussion is based on JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE, MOSTLY HARMLESS ECONOMETRICS 13–15 (2009).

32. For instance, in the difference-in-difference (DID) model examined in *infra* note 78 and accompanying text, specifying the potential outcomes mathematically shows *why*, with correct assumptions, statistical calculations on observed data may yield a plausible estimate of a causal effect. *See* ANGRIST & PISCHKE, *supra* note 31, at 229–31 (proceeding from a model of potential outcomes to the key practical assumption in DID modeling, expressed in their Figure 5.2.1).

B. The Development of Correlation and Causal Analysis

These varied and complex definitional understandings of causality stem from the thousands of years in which philosophers and scientists wrestled with its implications. In the beginning, causal thinking likely originated in attempts to understand natural phenomena by applying religious worldviews. In the ancient world, many events were considered pre-determined, which left little motivation for attempting to understand underlying scientific mechanisms.[33] Gods or godlike powers controlled the weather, disasters, the outcome of wars, and so on.[34] These powers resided in the sky, earth, sea, trees, and rivers, and could be addressed through rituals such as animal sacrifice. Within this worldview, there was little benefit to scientific experimentation or analysis, as effort was better spent attempting to bend the will of the gods—"[w]hen the gods have made up their minds they do not change them lightly."[35] When humans acted within this paradigm, their free will (if it existed at all) invoked blessings or cursing from divine powers.

33. PEARL, *supra* note 2, at 332.

34. *See id.*

35. Homer, *The Odyssey: Book 3*, A HOMER COMMENT. IN PROGRESS, https://homer.chs.harvard.edu/read/urn:cts:greekLit:tlg0012.tlg002.perseus-eng4:3.1-4.20 (last visited Mar. 13, 2022). In one short passage, the mortals contemplate sacrifices to appease one god, worry about "mischief" from another, and offer sacrifice for safe ocean passage, which goes well because of divine intervention:

> [T]his displeased Agamemnon, who thought that we should wait till we had offered hecatombs to appease the anger of Minerva. Fool that he was, he might have known that he would not prevail with her, for when the gods have made up their minds they do not change them lightly. . . .
>
> That night we rested and nursed our anger, for Jove was hatching mischief against us. But in the morning some of us drew our ships into the water . . . while the rest, about half in number, stayed behind with Agamemnon. We—the other half—embarked and sailed; and the ships went well, for heaven had smoothed the sea. . . .

Id. The advent of engineering began to change this perspective, as causal plans could be carried out through observable physical mechanisms. PEARL, *supra* note 2, at 333.

This reasoning slowly changed with scientific advancement. Greater understanding of physical principles behind events meant increased ability to manipulate nature and greater incentive to understand the mechanisms that chained natural events together. As early engineers showed the ability to create complex mechanisms that gave physical aspect to causal thinking, causality began to be attributed to objects, rather than just people and the gods they worshipped.[36] Philosophy began to develop specific notions of cause and effect which would later develop into modern mathematical conceptions of causality.[37] For instance, in the 1680s, John Locke attempted to bring definitional sense to the idea, noting that a "[*c*]*ause* is that which makes any other thing, either simple [*i*]*dea*, substance, or mode, begin to be; and an [*e*]*ffect* is that, which had its [b]eginning from some other thing."[38]

In the 1700s, David Hume provided a significant development in applying formal philosophical structure to causal theory in a way that incorporated human analytics.[39] In Hume's theory, causality derived from the power of human observation:

> Thus we remember to have seen that species of object we call *flame*, and to have felt that species of sensation we call *heat*. We likewise call to mind their constant conjunction in all past instances. Without any farther ceremony, we call the one *cause* and the other *effect*, and we infer the existence of the one from that of the other.[40]

In Hume's reasoning, all that can be said about a causal connection between events is what has been observed about the conjunction between them: we repeatedly observe flame, and then heat, and so flame

36. "When a [complex] system broke down, it was futile to blame God or the operator—instead, a broken rope or a rusty pulley were more useful explanations, simply because they could be replaced easily and make the system work." PEARL, *supra* note 2, at 333.

37. For example, Galileo helped provoke a scientific revolution in applying the language of mathematics to explain natural phenomena. *Id.* at 334.

38. JOHN LOCKE, AN ESSAY CONCERNING HUMAN UNDERSTANDING 325 (Peter H. Nidditch ed., Clarendon ed. 1975).

39. In Pearl's account, Hume "shook up causation so thoroughly that it has not recovered to this day." PEARL, *supra* note 2, at 336.

40. 1 DAVID HUME, TREATISE OF HUMAN NATURE 264 (T.H. Green & T.H. Grose eds., 1878).

causes heat. In one of Hume's definitions, *cause* means "[a]n object precedent and contiguous to another, and where all the objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter."[41] To paraphrase, causation in this view means observable, repeated, and subsequent correlation. First comes rain, then comes mud, and it happens so every time. We might thus say that rain *causes* mud.

While this notion has intuitive appeal and goes further even than the typical modern definitions of "cause" discussed above, Hume's ideas failed to gain traction with the developing discipline of statistics, which would become central to modern conceptions of proving cause and effect. Instead, the discipline of statistics focused on co-relation, or correlation, which could be established with precise mathematical rules, while relegating cause to a secondary role.[42]

The ability to measure the "closeness of co-relation" or correlation between two variables allowed precision in stating the strength of linear relationships between observed phenomena.[43] Perhaps because early statisticians found correlation quantitatively pleasing and because

41. C. M. Lorkowski, *David Hume: Causation*, THE INTERNET ENCYCLOPEDIA OF PHIL., https://iep.utm.edu/hume-cau/ (last visited Mar. 13, 2022). Hume offered a parallel definition of causality more directly centered on the human impression from events, and there is debate about the contiguity of these two concepts. *See id.* Hume himself apparently found any definition of cause somewhat lacking. *Id.* (noting Hume's statement that "it is impossible to give any just definition of *cause*").

42. For example, in 1888, Francis Galton, a Victorian-era founder of the discipline, famously published a study measuring head size and forearm length, which focused on the mathematical calculations of this relationship, while noting in passing that "[i]t is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common cause." Quoted in PEARL, *supra* note 2, at 339. Galton may have used the term "co-relation" to emphasize this new mathematical terminology rather than the less formal sense the term "correlation" connoted in his day—prior to Galton, correlation was understood in terms of dichotomous relationships, in which the presence of one thing implied the presence of another. STEPHEN M. STIGLER, THE HISTORY OF STATISTICS: THE MEASUREMENT OF UNCERTAINTY BEFORE 1900, at 297–98 (1986) (quoting a contemporary definition of correlation as: "Things are correlated . . . when they are so related or bound to each other that where one is the other is, and where one is not the other is not"). Galton observed the mathematical relationships between variables, and then correctly identified that one did not necessarily *cause* the other, in effect discerning between correlation and causation as mathematical correlation was established.

43. STIGLER, *supra* note 42, at 296.

the issues behind causality are fundamentally qualitative,[44] the study of causation was relegated to experimental design (where causal claims are often established through simple assumptions based on randomization) rather than permeating statistics as a discipline.[45] Statistics developed precise language to describe ideas related to correlation, such as the interpretation of regression coefficients or testing correlations to determine their significance, but failed to develop in probability theory precise language or formal models to describe causal thinking. Even a common-sense idea such as "rain causes mud, and not the other way around" is impossible to state in the traditional language of probability.[46] At best, one can say they are correlated: the more likely one is to observe mud, the more likely one is to observe rain.[47] That likelihood can be made incredibly precise, all without acknowledging the directionality of cause.

For statisticians in experimental settings, causal thinking was more prevalent.[48] An experiment meant deliberately changing something and then observing a result, which action brought causality closer to the forefront of theory.[49] Experimental techniques were developed beginning in the sixteenth century which allowed direct manipulation of inputs and increased the likelihood that the experimenter was

44. SHADISH, COOK & CAMPBELL, *supra* note 28, at 6. ("[C]ausal inference, even in experiments, is fundamentally qualitative."). That is, the reasoning behind concluding that something causes something else is non-numeric reasoning involving actions, their effect, assumptions, counterfactual evidence, and so on. These are qualitative concepts which control our interpretation of quantitative results.

45. PEARL, *supra* note 2, at 340–42. Indeed, in Stigler's monumental *The History of Statistics*, causality does not even merit an index entry, thus ranking behind, for example, psychology and psychophysics in its importance to the science. STIGLER, *supra* note 42, at 400, 408. Stigler does briefly mention causality while quoting Udny Yule on the difficulty of measuring causality in the economic realm.

46. PEARL, *supra* note 2, at 342.

47. *Id.*

48. Even though social scientific work in the law will perhaps rarely be based on evidence from laboratory experiments, understanding the language and theory developed by statisticians in experimental settings will be instructive for what *can* be said about causality in non-experimental settings, which are more likely to be encountered in legal analysis.

49. SHADISH, COOK, & CAMPBELL, *supra* note 28, at 2.

causing change rather than observing results of a spurious correlation.[50] Early experimenters also began to develop the idea of "controlling" for outside influences that might affect results. For example, if the air around a telescope was smoky, moving to a higher altitude could remove, or control for, the effect of smoke.[51]

These outside factors in the realm of the hard sciences such as physics, chemistry, and astronomy were relatively straightforward to identify, if not always to control. As scientists moved to analyze biology or social situations, the number of important outside factors that could confound results multiplied, in ways that could not be controlled for by, e.g., moving a telescope to higher elevation. For example, in social settings, selection bias could easily plague results by making those most likely to benefit from treatment more likely to seek it, thus confounding the analysis of what would happen should the treatment be applied to someone else less likely to benefit.[52] If only the very able seek education and let us observe what education produces, or if only the very sick seek treatment and let us observe what occurs, we will not know the effect on the less able or the less sick.

In these situations, the idea of randomization was developed to cut through possible extraneous influences on the experiment.[53] By randomizing allocation to a treatment, statistical theory suggested that the effect of confounding factors on the experiment could be dramatically reduced.[54] We will address this concept at some length below.[55] Even though it may be exceptionally rare to see direct, randomized experimental evidence brought in the constitutional scrutiny context, it is important to understand the vocabulary that developed to describe this kind of experimentation, as those concepts will be foundational when

50. *Id.* at 1. An example might be assuming head size causes forearm length in Galton's framework. *See supra* note 42.

51. SHADISH, COOK, & CAMPBELL, *supra* note 28, at 2.

52. By treatment, we mean application of an experimental cause, whether or not that constitutes potential "treatment" of a health issue.

53. SHADISH, COOK, & CAMPBELL, *supra* note 28, at 2–3.

54. *See* Jamison, *supra* note 22, at 3 ("The value of randomized assignment is that it implies that the measured status of the condition which is randomized is uncorrelated (in expectation) with any of the other conditions, and hence that any variation in outcomes as a function of that status must be due to the influence of that randomized condition, at least within the set of potential conditions examined.").

55. *See infra* note 64 and accompanying text.

understanding what can be said of causality outside experimental settings.

Three concepts are central to this vocabulary: "observable" variables, "unobservable" variables, and "confounding" variables.[56] An observable variable is one that can be measured by the researcher either numerically or categorically.[57] An individual's height, weight, education level, and other demographic characteristics are all observable variables in this sense.[58] In contrast, some variables are "unobservable" either by their intrinsic nature or the nature of the research design. One's innate propensity for hard work, for instance, is not an observable characteristic by its nature. This variable might have proxies, such as performance on certain kinds of exams, but exam performance will reflect a host of factors other than the variable of interest, such as education, training, one's breakfast the morning of the test, whether the exam is taken shortly after daylight savings disrupts sleep schedules, and so on.[59] Other variables might be observable in some settings, but not others, such as when a study is insufficiently funded to measure everything of interest before running an experiment.[60]

Some observed or unobserved variables matter more than others. An unobserved variable matters when, without it, the results of the study would be incorrect. We refer to this as a "confounding" variable. Consider a classic example: identifying the effect of education on earnings.[61] The causal effect of education on earnings (such as how much

56. We might add the term "variable" itself, by which we mean any set of information collected by a researcher for use in analysis.

57. *See* Jamison, *supra* note 22, at 2 (discussing observable and unobservable characteristics).

58. This does not mean that, for example, the researcher is able to tell one's ethnicity by looking at them. It merely means ethnicity can be included in an analysis after, e.g., self-identification.

59. *E.g.*, Nathan Collins, *ScienceShot: Daylight Savings Hurts Test Scores*, SCIENCE (Nov. 22, 2010), https://www.sciencemag.org/news/2010/11/scienceshot-daylight-savings-hurts-test-scores (discussing differences in exam scores due to daylight savings clock adjustments).

60. The idea of observable and unobservable variables is closely related to the difference between manipulable and non-manipulable variables or causes. *See* SHADISH, COOK, & CAMPBELL, *supra* note 28, at 8.

61. *See* Karen Clay, Jeff Lingwall, & Melvin Stephens, Jr., *Laws, Educational Outcomes, and Returns to Schooling: Evidence from the First Wave of U.S. State*

additional income one could be expected to make if required to attend another year of school) is of great interest to those designing education policy.[62] For instance, the question of how many years of compulsory attendance to require of children hinges on the effect those years of education are expected to have in their lives. Yet despite an abundance of data on education and earnings, such as in the United States censuses, identifying the causal relationship is difficult. The problem is that both education and earnings are impacted by innate characteristics which cannot be observed, like one's innate attitude towards hard work. This attitude is likely to affect both education *and* earnings. People who like hard work are more likely to press through many years of education and are also likely to earn more over their lifetimes because of their work ethic regardless of educational attainment.[63]

Because of the problem with many potential confounding variables in experimental settings, randomization was developed as a key tool to control for these effects.[64] Consider if one could run an enormous, randomized, education experiment. If one could take a large number of students and randomly assign them to eleven or twelve years of education, then one would expect those with more or less propensity to work hard to be divided at least somewhat evenly among the two

Compulsory Attendance Laws, 68 LABOUR ECON. 101935 (2021) (collecting literature examining returns to education in the econometric context).

62. *See id.* at n.1.

63. Even if one studies the average earnings of those with eleven versus twelve years of education, the difference between those averages is unlikely to reflect what will occur if one makes an eleventh-grade student attend another year of school, such as through a change in compulsory attendance or child labor laws. The unobservable variable (work ethic) has confounded the causal conclusions of the observational study. An observed, or unobserved, variable does not matter when it has no effect on the relationship of interest. For instance, a researcher might not observe someone's opinion of the television show *Jersey Shore*, but that opinion is relatively unimportant to the effect of education on that person's earnings. This is not to say that one's opinion of *Jersey Shore* is unimportant, just that it likely has no effect on the outcome, or that it has no effect on the outcome once other more easily observable variables are controlled for. The opinion is therefore unlikely to confound the effect of the study. SHADISH, COOK, & CAMPBELL, *supra* note 28, at 7.

64. For a history of the development of randomization in the social sciences, see generally Jamison, *supra* note 22.

groups.[65] In terms of probability theory, that unobservable variable is the same between our two groups "in expectation."[66] This means that, for example, if we added individuals to our pool randomly in an infinite amount, the probability of the two groups diverging in terms of innate propensity for hard work in any substantial manner goes to zero.

Randomized assignment to treatment and control groups implies probabilistic equality, which then allows us to realistically *assume* that the unobservable variable has not affected the outcome and our results are not biased away from the truth because of confounding variables. Even more, it allows us to realistically assume that *all* unobservable variables are equal in expectation and thus will not affect the

65. Crucially, one would expect this to be true of *any* potential confounding variable. Because of this, the researcher need not worry about identifying and controlling for every possible confounding variable.

66. Jamison, *supra* note 22, at 3 ("The value of randomized assignment is that it implies that the measured status of the condition which is randomized is uncorrelated (in expectation) with any of the other conditions, and hence that any variation in outcomes as a function of that status must be due to the influence of that randomized condition, at least within the set of potential conditions examined."). Expectation, put simply, is what one would expect from a random event if one could repeat that event infinitely and average the results. More formally, expectation of a random variable with a finite number of outcomes (such as the die example above) is the weighted average of the probability of each outcome. The expectation of a continuous random variable (such as a Gaussian or "normally distributed") variable is found through integrating the value of the variable, multiplied by its probability, over all possible values the variable can take. For example, if one rolls two dice and sums their number, the sum must be between two (a single pip plus a single pip) and twelve (six pips plus six pips), with the highest probability being the number seven, due to the possible combinations of the die faces. The number seven can be attained by six pips plus one pip, five pips plus two pips, and so on. The probability of rolling a total of seven on two dice is approximately 16.67%. In contrast, the probability of rolling a two or a twelve are both 2.78%. One can find these probabilities by listing the total number of combinations on the face of two dice (there are thirty-six) and then dividing the total number of ways to reach a number by thirty-six. Any individual roll is difficult to predict, yet if one rolled the dice hundreds and then thousands of times, taking the average of these rolls, the average would eventually converge to something very close to seven. Over an infinite number of rolls, probability theory would require this number to be precisely seven. To be slightly more technical, the law of large numbers would require the probability of observing a result different from seven to tend towards zero over time.

outcome.[67] To combine the terms from above, because of what happens in expectation in a randomized experiment, researchers are able to control for the effect of unobservable variables and make the reasonable assumption that the results have not been confounded or biased. The strength of that assumption depends on the details of the study, such as its size, the number of confounding causes, and so on.[68] Together, the language of expectation and assumptions is key to understanding causal reasoning in a statistical context.

67. Of course, whether this assumption is born out in practice depends on the nature of the research design. The larger the sample, the more one would expect potentially confounding causes to be evenly balanced between treatment and control groups, and the more potential confounding causes, the greater the assumption that each is balanced in practice. *See, e.g.*, Angus Deaton & Nancy Cartwright, *Understanding and Misunderstanding Randomized Controlled Trials*, 210 SOC. SCI. & MED. 2, 6 (2018). Deaton and Cartwright explain:

> Even with very large sample sizes, if there is a large number of causes, balance on *each* cause may be infeasible . . . . [T]here are three billion base pairs in the human genome, many or all of which could be relevant prognostic factors for the biological outcome that we are seeking to influence . . . . Out of all those billions, only one might be important, and if that one is unbalanced, the results of a single trial can be "randomly confounded" and far from the truth. Statements about large samples guaranteeing balance are not useful without guidelines about how large is large enough, and such statements cannot be made without knowledge of other causes and how they affect outcomes.

Id. These issues do not affect the *probabilistic* lack of bias, but rather provide nuance to the assumption that the lack of bias in expectation has been born out in practice. *See id.* ("Of course, lack of balance in the net effect of either observables or nonobservables . . . does not compromise the inference in [a] [randomized controlled trial] in the sense of obtaining a standard error for the unbiased [average treatment effect].").

68. One check on this assumption is to examine how balanced potential observable confounding variables are between treatment and control groups. *E.g.*, *id.* ("Having run a [randomized controlled trial], it makes good sense to examine any available covariates for balance between the treatments and controls; if we suspect that an observed variable x is a possible cause, and its means in the two groups are very different, we should treat our results with appropriate suspicion. In practice, researchers often carry out a statistical test for balance after randomization but before analysis, presumably with the aim of taking some appropriate action if balance fails.").

Even with the magic of randomization and its effect on the assumptions behind causal claims, experimental evidence still has limitations. One is particularly noteworthy: experimental results may not be generalizable.[69] The results of a generalizable experiment can be applied in other settings. They are "externally valid."[70] This means that, for instance, in a medical study the treatment would work on those in the experiment and the public at large. One can immediately see how this notion is tied to the idea of randomization—if the experiment was properly randomized among all potential population groups among whom we would like the results to apply, we expect the results to apply, at least on average, regardless of the personal attributes of the individual. Yet despite randomization, problems with generalizability may remain. For instance, if a weight loss study only had participants who were clinically obese, one worries about the generalizability of that study to those with other body compositions regardless of whether the participants in the study were randomized into treatment and control groups. Randomization would have a powerful effect on unobserved and possibly confounding variables *within* that population group, but that might be all.[71]

To summarize, the remarkable power of experiment evidence is that probabilistically or in expectation, unobservable variables are balanced between treatment and control groups. This allows us to believably make the statistical assumption that we identify causal effects after imposing a treatment in an experimental setting, without those effects being confounded by other variables. We would then expect those causal effects to be generalizable at least to the population groups included in our randomization.[72]

69. *See* Jamison, *supra* note 22, at 3 ("Randomization yields 'internal validity' . . . but not 'external validity'. [sic] Of course, the larger the relevant population or expanse of observed conditions, the more widespread and robust is the conclusion.").

70. *See, e.g.*, Allan Steckler & Kenneth R. McLeroy, *The Importance of External Validity*, 98 AM. J. PUB. HEALTH 9, 9–10 (2008).

71. *See* SHADISH, COOK, & CAMPBELL, *supra* note 28, at 18–19. Other issues may provide additional challenges. *See id.* at 9. This is particularly true of the burgeoning field of experimental evidence in psychology and economics, in which clever experiments conducted on, e.g., groups of undergraduate college students might not be generalizable to other groups with great confidence.

72. In the words of one of the founders of statistics, R.A. Fisher, randomization "relieves the experimenter from the anxiety of considering and estimating the

Now consider what this entails in settings in which experiments are impossible, as are often faced by issues in constitutional scrutiny challenges, which typically do not involve areas of social science in which randomized experimentation is possible. For example, the illustration above concerning education and earnings shows one area which cannot be studied easily through direct experimentation.[73] Because of these difficulties, researchers in the field of economics began approaching causality through a set of very different methods.[74] Experiments in the economic sphere were traditionally impossible to run—manipulating entire nation-state level economies as one would subjects in a laboratory experiment is difficult,[75] and isolating phenomena to affect single variables at a time in the "data of daily experience" is likewise fraught when studying human behavior.[76]

At the same time, policy makers relying on advice from economists would be most concerned with the causal conclusions from the economists' research, as they wish to know what effect their policies are likely to have. As such, economists began to look for ways to

magnitude of the innumerable causes by which [the] data may be disturbed." R.A. FISHER, THE DESIGN OF EXPERIMENTS 49 (1935).

73. *See, e.g.*, Clay, Lingwall, & Stephens, Jr., *supra* note 61.

74. *See* Jamison, *supra* note 22, at 12–14 (discussing the use of randomization in economics).

75. This is not to say that economists do not run experiments. There is a very active branch of modern economics which does employ experimental methods, though at a level typically less than national. *See* Abhijit V. Banerjee & Esther Duflo, *The Experimental Approach to Development Economics*, 1 ANN. REV. ECON. 151, 152 (2009).

76. Statisticians recognized this principle as early as 1897. *See* G. Udny Yule, *On the Theory of Correlation*, 60 J. ROYAL STAT. SOC. 812, 812 (1897). Quoting Yule, Stigler writes:

> The investigation of causal relations between economic phenomena presents many problems of peculiar difficulty, and offers many opportunities for fallacious conclusions. Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and discuss as best he can the relations of a whole group of changes; he cannot, like the physicist, narrow down the issue to the effect of one variation at a time.

STIGLER, *supra* note 42, at 348.

exploit quasi-experimental variation in clever ways. They did this by creating new sets of statistical skills that could ground causal conclusions in assumptions that did not depend on randomized division into treatment and control groups. By carefully studying the *qualitative* aspects of a social situation, and then combining that with rigorous *quantitative* work in statistics, conclusions could be drawn about causality based on assumptions that did not depend on randomization.

For example, as scholars studied the effect of education, they turned to what might be termed "natural experiments." A natural experiment is one in which quasi-experimental variation is created in a manner outside a laboratory. For instance, if one assumes that statewide passage of compulsory attendance laws is unrelated to the individual child's decision to attend school or drop out, then one could use variation in dates those laws were passed between states to study the causal effect of those laws on attendance and future earnings. In essence, the states that did not pass laws might serve as controls for the states that did, in the same way as one might allocate patients to treatment and control groups in an actual experiment.[77]

For a specific example of this type of non-experimental, causal evidence, suppose a researcher wished to study the effect a change in minimum wage law had on unemployment. Suppose also that New Jersey implemented a minimum wage policy that neighboring states did not. This is unlike randomly sorting states into "new minimum wage"

77. This type of analysis is part of a variety of other techniques known as "structural causation." In the discipline of econometrics—the combination of statistics and economics most concerned with development of causal theory outside direct experimentation—this means the science of making causal claims based on the structure of a mathematical system that captures (in model form) the quantitative notions of cause and effect. There are various ways in which this term is used, but this categorization suffices for our purposes. *See* Hamish Low & Costas Meghir, *The Use of Structural Models in Econometrics*, 31 J. ECON. PERSPS. 33, 35–42 (2017), https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.31.2.33. A second area of causation sometimes relied on in economics is "Granger" causation. This proceeds under a vastly different set of assumptions and in different scenarios. Granger causation is the idea that closely related temporal events in time series data provide evidence of causality. *See, e.g.*, Barbara Rossi, *Advances in Forecasting Under Instability*, *in* 2B HANDBOOK OF ECONOMIC FORECASTING 1203 (Graham Elliott & Allan Timmerman eds., 2013) (explaining how researchers establish whether there is Granger-causality when there are statistically significant instabilities in Granger-causality relationships).

and "old minimum wage" groups as one would in an experiment. However, if we believe that economic conditions in New Jersey and neighboring states are generally similar, we may feel comfortable making certain sets of statistical assumptions that can allow us to treat New Jersey's actions similar to an experiment.[78] For instance, so long as one assumes that the trend in unemployment in New Jersey and neighboring states would be the same in the absence of the New Jersey law, one can treat the unemployment trend outside New Jersey as a control or counterfactual for the unemployment trend within the state of New Jersey. This is a stronger assumption than what would be made in an experiment which randomized subjects into treatment and control groups. Here, if neighboring states are fundamentally different from New Jersey in confounding ways, the assumption of a valid counterfactual would be suspect.

Consider the details of how one might implement this technique. In a simple version of this methodology, we will have an early time (prior to the policy change) called t_1 and a time after the policy change, called t_2. If we observe unemployment in both New Jersey and neighboring states at both times, one could take the unemployment rate in New Jersey after the law at t_2, and subtract the unemployment rate in New Jersey before the law at t_1 to get the change in employment in the "treated" state. Then, one could take the average unemployment rate from the neighboring "control" states at t_2 and subtract the average unemployment rate at t_1 in those states to find the counterfactual trend. Then, one could simply subtract the second difference (the change in control states) from the first difference (the change in New Jersey, our treatment state) to obtain an estimate of the causal policy effect of the law.[79]

For example, if the unemployment rate in New Jersey were 10% before the law at t_1 and 15% after at t_2, that makes a difference of 5%.

78. For a treatment of this situation outside the example context, see David Card & Alan B. Krueger, *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply*, 90 AM. ECON. REV. 1397 (2000).

79. In mathematical form, we might estimate the causal effect as $\hat{y} = \left(\bar{u}_{NJ,2} - \bar{u}_{NJ,1}\right) - \left(\bar{u}_{C,2} - \bar{u}_{C,1}\right)$, where $\bar{u}_{NJ,1}$ is the average unemployment rate in New Jersey and *t1*, $\bar{u}_{C,1}$ is the average unemployment rate at *t1* in control states, and so on.

If the employment rate was, on average, 11% in other states at t_1 and 8% at t_2, the control difference would be 3%. We subtract 3% from 5% to obtain a causal effect of 2%—the minimum wage policy in New Jersey likely increased unemployment by 2% relative to the counterfactual trend in other states.[80] Because we are subtracting a difference from another difference, this is referred to as a "difference-in-difference" (DID) estimate.[81]

In this technique as in others, the key to a causal interpretation is the strength of the underlying assumptions. In the example above, we quickly described these assumptions as "the trend in unemployment between New Jersey and neighboring states would be the same in the absence of the New Jersey law." This sort of assumption can be surprisingly strong. For instance, states constantly pass various policies that affect the unemployment rate, from changes in taxes to incentives for businesses to move to the state. A variety of non-policy factors have tremendous implications on unemployment, such as the strength of local and regional economies, national and international trends, and so on. To treat our example as near-experimental requires us to assume that each of those factors affect New Jersey and its neighboring states equally.[82]

80. In terms of the prior footnote, $0.02 = (0.15 – 0.10) – (0.11 – 0.08)$.

81. This methodology can quickly become significantly more complicated. For instance, one might go a step further and employ *triple* differencing, such as looking at changes in levels between ages, states, and over time. *E.g.*, Carolyn M. Moehling, *State Child Labor Laws and the Decline of Child Labor*, 36 EXPLORATIONS ECON. HIST. 72 (1999) (using a triple difference strategy to examine the effect of state child labor laws). Such a technique has the advantage of relying on weaker causal assumptions than a standard two-level differencing technique, at the cost of creating results that are less generalizable.

82. This analysis is presented at a level in which the arguments and assumptions behind DID estimation might be considered in litigation. More technically, a causal interpretation of these estimates requires that "the outcome variable in the absence of treatment is generated by a model that is (strictly) monotonic in the unobservables, and the distribution of such unobservables must be invariant over time." Low & Meghir, *supra* note 77, at 41. This means that the unobserved variables have a non-decreasing relationship with the variable of interest (so that as the unobservable variable increases, the outcome of interest does not decrease), and that the size of the unobserved variable does not change as time progresses. For instance, when looking at the effect of a compulsory attendance law, it means an unobservable trait related to

We can strengthen the causal interpretation in our example by lowering the number of required assumptions or weakening those assumptions. (Again, recall that *strong* assumptions are less likely to be satisfied than *weak* assumptions.) For example, if we believe that western Pennsylvania is an improper control group for employment trends in New Jersey, perhaps we narrow our analysis to only considering counties on the border between New Jersey and neighboring states. If we limit the analysis to these counties, we might expect that the same local trends affect unemployment in these areas, so our analysis does not require as strong an assumption to claim we are finding a causal effect of the law. This has a cost. By narrowing the analysis, to weaken the required assumptions, we have narrowed the potential scope of our study. This, like the experiment described above limited to the clinically obese, limits the generalizability of the study. We might be more certain we have found truth, but less certain that truth applies beyond the New Jersey border.

One way to strengthen our causal claim, by weakening the assumptions underlying it, is to adjust for various observable trends in our analysis. If the researcher feels, for example, that unemployment is closely related to other economic trends in the region, such as inflation, they might include local inflation rates in their model as a control. By including it in the model, the researcher no longer needs to assume that inflation has no effect on the outcome. Without that control, if inflation were related to unemployment, the necessary assumption would be that inflation's effect on unemployment in both New Jersey and neighboring states is identical. This might be a troubling assumption, and so including the control variable weakens the required assumptions and strengthens the analysis. Again, the key idea behind making causal claims based on statistical evidence is the qualitative analysis of assumptions underlying the analysis.

DID analysis has been relied on by many litigants. For example, litigation in the Northern District of Illinois considered whether a class of healthcare purchasers had been harmed by the acquisition of an alleged monopoly power created by a hospital merger.[83] To prove

school attendance such as work ethic must not lead to less school, and work ethic does not change over time.

83. *In re* Evanston Nw. Healthcare Corp. Antitrust Litig., 268 F.R.D. 56 (N.D. Ill. 2010).

classwide damages, plaintiffs turned to DID estimation on prices of healthcare services paid by every insurer who dealt with the hospitals at issue and a group of control hospitals.[84] By taking the difference in payments before and after the merger, for the "treated" hospitals that had merged and the group of unmerged "control" hospitals, plaintiffs argued they could ascertain the anticompetitive overcharge as the basis for damages.[85] Naturally, the hospital attacked this methodology, and its objections illustrate typical challenges facing DID estimation, which will be germane to whether they provide relevant causal estimates in this example, and so are worth examining in some detail.

First, defendants attacked the relevance of DID estimation to specific individual-level damages.[86] They argued that the DID estimates were insufficient, as "price increases should be expected to vary across payors, type of plan (HMO/PPO), patient cost-sharing arrangements, and type of service."[87] The statistics, they argued, assumed the price of services increased by the same amount for each patient.[88] This line of argument is common to any type of aggregate statistical analysis. The introductory examples above, and statistical theory in general, rely on using aggregate data to inform decision making. If the example unemployment situation above was followed in real life, average unemployment numbers would be used to calculate a causal effect rather than comparing any single individual in New Jersey to any single other individual in any other state. Because statistics rely on aggregation, their use is frequently attacked in litigation that requires individual-level damage calculations.

These attacks bely understanding of statistics: although averages are used to inform parameters of statistical models, those averages

84. *Id.* at 68–72.

85. *Id.* at 72.

86. The level of direct attack of statistical methodology common in litigation can be unsettling to practitioners. Attacks in academia tend to be more subtle and respectful, perhaps because often little money rides on the result. Yet, strong arguments can be made that this level of examination is what academia needs if it intends to influence public policy, as causal estimation in economics often aims to do. *See, e.g.*, David W. Schnare, *Academic Research Transparency and the Importance of Being Earnest*, 49 J.L. & EDUC. 1 (2020).

87. In re *Evanston*, 268 F.R.D. at 75.

88. *Id.*

can then be used to help make informed predictions at the individual level. In essence, the prediction for any single individual can often be a powerful aggregate of their own data plus aggregate statistical data.[89] The Supreme Court confirmed the propriety of the potential for such aggregate information in its little-heralded decision in *Tyson Foods, Inc. v. Bouaphakeo*, noting that statistical evidence, including aggregates, should be viewed procedurally as any other evidence:

> A representative or statistical sample, like all evidence, is a means to establish or defend against liability. Its permissibility turns not on the form a proceeding takes—be it a class or individual action—but on the degree to which the evidence is reliable in proving or disproving the elements of the relevant cause of action.[90]

Second, the defendant attacked the assumptions behind the DID estimation. As we have discussed, DID estimation relies on the assumption that the control group represents a valid counterfactual for the group experiencing a policy change. In our unemployment example, this assumption required that trends in unemployment between New Jersey and neighboring states would be the same in the absence of the New Jersey law. As with the attack on individual level damage calculations, this assumption is based on the use of aggregation. Here, the defendant mistakenly argued that "the control hospitals must be virtually identical" to the treated hospitals.[91] The court rightly rejected this argument. In the court's terminology:

89. For example, one's height can be predicted as a combination of the height of one's parents plus information on the average height of people one's age. The individual's personal information (parental height) plus aggregate information (average height) combine to yield an overall prediction. *See Height Calculator*, CALCULATOR.NET, https://www.calculator.net/height-calculator.html (last visited Mar. 13, 2021) ("Normally, a child's height is based on parental heights subject to regression toward the mean."). In simple linear regression, in which a line of best fit is used to describe the relationship between an explanatory and response variable, computing the slope and intercept of the line rely on using summary statistics from all the variables, and then the slope and intercept of the calculated line, together with individual level explanatory data, are used to form predictions for the individual.

90. Tyson Foods, Inc. v. Bouaphakeo, 577 U.S. 442, 454–55 (2016).

91. In re *Evanston*, 268 F.R.D. at 85.

> At the most basic level, DID analysis is a comparison of averages—the average prices at [the treated hospitals] are compared with the average prices [of] hospitals similar to [them]. If DID analysis required control hospitals to be identical to [the treated group], it would not make sense to have more than one hospital in the control group, as they would all provide the exact same data. Of course, the court acknowledges that the control group must be composed of hospitals sufficiently similar to ENH . . . . [T]he court is satisfied that here . . . [the] control group [is] valid and reliable. [92]

This statement is correct and understandable, but it is worth being precise in language, as such precision will show even more clearly why the plaintiffs were more correct in this instance. In just slightly more technical terms, we can interpret a DID estimate as causal if we can assume the trend in the control group matches what would happen to the treated group in the absence of the policy.[93] Recall from the discussion of experimental causality above that the key causal assumption in a randomized controlled experiment is that the treated and control groups are identical *in expectation*. These groups will rarely be identical in practice, but the randomization makes the groups equal in the theoretical, probabilistic averages of their unobserved variables. Since the DID assumptions rely on a probabilistic expectation for their causal validity, the expression of that expectation in the analysis is through an *average* trend, precisely as the court observed. In terms of the reasoning employed in this Article, defendants were arguing that the statistics should require individual-level identicality with the treated hospital, which would be an incredibly strong assumption to make. The court rightly rejected this and correctly reasoned that the weaker assumption was that the control hospitals were similar *in aggregate*.[94]

92. *Id.*

93. The same more technical caveats as in *supra* note 82 apply.

94. The court's level of examination here mirrors what might be found by many empirical economists who regularly estimate DID models, while falling short of the more technical assumptions for causality emphasized by econometric theorists. *See supra* note 82 (discussing approaches to the assumptions behind DID reasoning).

Yet another attack remains. In an experiment, the expectations match because of randomization. The randomization in essence breaks any correlation between unobservable variables and group selection. DID estimation is strongest when the treatment can be assumed to be quasi-random, or more specifically, the selection into treatment and control should be uncorrelated with unobservable characteristics that could bias the regression. Here, the underlying basis of the entire litigation is that one particular group of hospitals merged, while others did not. If the reasons those particular hospitals merged is correlated with an unobserved variable, such as financial health, scope of services, or so on, then we fail to create quasi-experimental variation and our estimates might be biased. In terms of the paragraph above, if we have this kind of bias then even in expectation our control group is no longer valid.[95]

In sum, the development of correlation and causal analysis has spanned centuries, beginning in ancient notions of cause and effect before treatment by Enlightenment scholars and modern theorists. In its present application, outside the context of randomized experiments, causality is inherently a quantitative analysis based on the assumptions one makes about the world. The example of DID analysis developed above illustrates the kind of common assumptions, arguments, and

95. For other examples, in antitrust litigation over car pricing, plaintiffs alleged that conspiracy among manufacturers created policies that limited the ability to import vehicles from Canada, which would have lowered prices in the United States. *In re* New Motor Vehicles Canadian Exp. Antitrust Litg., 632 F. Supp. 2d 42, 45–46 (D. Me. 2009). Plaintiffs attempted to show how change in background pricing passed through to the ultimate consumers of the vehicles. *Id.* at 57. Plaintiff used a study using DID estimation to show that change in effective price through a manufacturer's rebate lowered the transaction price substantially for consumers. *Id.* The court ultimately failed to rule on the validity of this statistical evidence, ruling that introducing the argument required expert testimony which had not been provided. *Id.* at 61.

Or in litigation over Keurig coffee pods labeled as "recyclable" when many California recycling plants would not accept them, DID methodology was proposed as a potential damages model to ascertain what price premium had been paid for the "recyclable" label. Smith v. Keurig Green Mountain, Inc., No. 18-cv-06690, 2020 WL 5630051, at *9 (N.D. Cal. Sept. 21, 2020).

Finally, defendants may rely on DID methodology, e.g., to attempt to show that differences in market choices between companies are unrelated to antitrust conspiracy. Kleen Prods. LLC v. Int'l Paper, No. 10 C 5711, 2017 WL 2362567 (N.D. Ill. May 31, 2017).

counterarguments that must be considered outside the context of randomized experiments.[96] The next subsection illustrates additional ways in which statistical evidence may be utilized.

C. A Spectrum of Causal Assumptions

This subsection begins to lay a framework for evaluating statistical evidence of causality in a legal context.[97] First, it differentiates between how courts conceive of evidence of correlation and causation. It then argues that causal evidence should be treated along a spectrum of weak to strong assumptions. The following Section will build off this spectrum when considering causal evidence in the constitutional scrutiny context.

Consider again the most basic piece of analytical literacy drilled into statistics students worldwide: the difference between correlation and causation. When someone refers to something as being "correlated," there are two distinct pieces of information they may be attempting to convey. They might be stating the simple idea that there is an observable relationship between things.[98] Or, they might be implying that a certain mathematical relationship has been calculated: correlation in statistical terms is the sample standard deviation of one variable, multiplied by that of another, and then divided by the covariance.[99] This results in a number bounded between negative one and positive

96. In the constitutional scrutiny context, DID analysis was a key element in the Eastern District of Arkansas's consideration of abortion restrictions in the state. There, an expert used data before and after an admitting-privileges law took effect in Texas to show the impact of decreased access. Planned Parenthood Ark. & E. Okla. v. Jegley, No. 4:15-CV-00784, 2018 WL 3816925, at *58 (E.D. Ark. July 2, 2018), *vacated*, No. 4:15-CV-00784, 2018 WL 9944527 (E.D. Ark. Nov. 9, 2018), *appeal dismissed*, No. 18-2463, 2018 WL 9944528 (8th Cir. Nov. 9, 2018). "[A]lternative causal explanations" were controlled for because "baseline" effects were captured by measuring abortion rates before and after enforcement of the Texas bill. *Id.*

97. We offer these tools not because we believe courts should necessarily hold statistical evidence to the same level as, for example, an econometrician under peer review, but because many of the principles that would be encountered under that level of review can be applied usefully and instructively by courts.

98. *Infra* note 148, and accompanying text.

99. Stephanie Glen, *Correlation Coefficient: Simple Definition, Formula, Easy Steps*, STATISTICS HOW TO, https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/ (last visited Mar. 13, 2022).

one, in which positive one indicates a perfect correlation (the two variables are identical, and to know one is to know another), zero indicates no correlation (knowing information about one tells you nothing about the other), and negative one indicates a perfect negative correlation (the two variables are opposites, such as if you multiplied each value of the first variable by the negative one). In these senses, the idea of correlation may express a dichotomous idea—concepts are related, or not—or, it may express the mathematically calculated level of a relationship, e.g., the correlation between X and Y is 0.75.

In contrast for "causality," the term is typically used in the former, dichotomous sense. A proposed relationship is either causal or not: does the rotation of the earth *cause* the sun to move through the sky? Yes. Does increased consumption of ice cream *cause* an increase in summer temperature? No. Causality is fundamentally not a continuous idea, unlike mathematical correlation, so that a statement such as "the causality between X and Y is 0.75" is meaningless.[100] While this is true, some nuance remains. No simple mathematical term can be calculated for causality, because unlike correlation, causality is a quantitative issue informed by reasoning, observation, and assumptions. Calculating a correlation coefficient expresses a strictly mathematical relationship, whereas the notion of causality attempts to summarize often complex real-life phenomena. While we cannot specify the level of causality on a continuous scale between two variables, we *can* express the confidence with which we can claim a causal effect. In particular, we can express the confidence we have in the assumptions behind causality, and these assumptions *do* range, not between negative one and positive one (as for correlation, as that would imply a quantitative calculation that does not exist), but between weak and strong.[101]

This range of assumptions can be considered along a roughly ordered spectrum. For shorthand purposes, we refer to this as a

100. The closest we come to this sentiment is when considering multiple causes, which we might apportion into some part of a whole. Note in addition, once we have established the assumptions behind causality, one *can* statistically say that, for example, increasing one variable will result in an expected change in another variable by a certain amount. But this is an application of the level of the relationship, not a statement of the level of causality.

101. *See supra* note 82 and accompanying text (introducing the idea of weak versus strong assumptions).

"causality spectrum." This spectrum or range of assumptions behind causal claims is not strictly ordered—that is, it is not the case that in every application one type of assumption and how those assumptions were treated in particular research will be stronger or weaker than another. Rather, it shows what in many applications can be considered weak versus strong assumptions behind causal evidence, and hence provide perspective on the strength of the evidence of causality. Section IV will place this spectrum in the context of useful non-mathematical narratives that could be applied to this type of evidence and use that narrative to examine evidence in constitutional scrutiny cases.

In the social-scientific context in which scrutiny cases often draw their evidence, the following presents the type of research one might see, roughly in order of the strength of the assumptions needed to plausibly claim evidence of a causal effect. The models presented are ones commonly used in policy analysis, and are sometimes called "treatment effect" models, to distinguish them from models based on more complex assumptions about human behavior.[102]

1. Assumptions in Large Randomized Experiments

First, large, randomized experiments or "trials" are the gold standard for establishing causal mechanisms. As discussed above, the ability to randomize cuts through a multitude of potential problems with analysis, such as the presence of confounding variables. Large

102. *See* Low & Meghir, *supra* note 77, at 41. Low & Meghir write:

> Not all treatment effect models are created equal: it is important to distinguish those estimated through randomized experiments from those estimated through quasi-experimental methods, such as difference-in-differences, regression discontinuity, matching, and others. The point of randomized experiments is that results do not depend on strong assumptions about individual behavior, if we are able to exclude the important issues discussed above. However, this clarity is lost with quasi-experimental approaches such as difference-in-differences, where the validity of the results typically depends on assumptions relating to the underlying economic behavior that are left unspecified.

Id.

studies, as well as studies repeated by multiple researchers, provide the strongest evidence of causal effects. If a randomized study is large and repeated with confirmation of the study results, the study should be generalizable to many circumstances and provides compelling support for causal claims. Even with their limitations, this type of evidence should powerfully show that government action is likely to have the policy effect being sought.[103] An early social scientific example is the famous RAND study on the effect of insurance plans.[104] Over 20,000 individuals in multiple cities were randomly allocated to insurance plans with various parameters and then studied over multiple years, showing that individuals were highly responsive to changes in out-of-pocket insurance costs.[105]

2. Assumptions in Limited Randomized Experiments

Second, randomized experiments in more limited circumstances provide similar assurances that randomization has helped cut through potential confounding variables. Yet, in the very specialized and limited context in which many social-scientific experiments are conducted, we may need to make somewhat stronger assumptions about the world to apply their results in other contexts.[106] As always, the randomization guarantees that the treatment and control groups are probabilistically equal, or equal in expectation across potential factors influencing the outcome. Yet the limited nature of these experiments calls for a stronger assumption that all potentially confounding causes have been appropriately apportioned among treatment and control

103. For examples of the limitations and issues involved in randomized experiments, see *supra* note 67 and accompanying text.

104. Willard G. Manning, Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, Arleen Leibowitz, & M. Susan Marquis, *Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment*, 77 AM. ECON. REV. 251 (1987).

105. *Id.* at 258.

106. *See, e.g.*, Deaton & Cartwright, *supra* note 67, at 5 (arguing that randomized controlled trials (RCTs) have received excessive attention relative to their merit, particularly as the relatively few assumptions which must be made to yield unbiased estimates mean prior knowledge may fail to be built upon) ("The gold standard or 'truth' view does harm when it undermines the obligation of science to reconcile RCTs results with other evidence in a process of cumulative understanding.").

groups, that the results are generalizable, and so on. For example, economic experiments conducted in limited settings have been critiqued as having problems with generalization outside their specific contexts.[107]

3. Assumptions in Quasi-Experiments

Third, some causal evidence may be derived from experiments in which the researcher has some ability to control selection into treatment and control groups but does so without full randomization. These are sometimes referred to as "quasi-experiments."[108] For example, students may self-select into a volunteer reading program created by a researcher. A policy evaluator would need to carefully assess the role confounding variables may have, compare balance between treatment and control groups in terms of observable characteristics and so on. To show a causal effect, the researcher would need to assume that important confounding variables are sufficiently controlled for or rendered unimportant by aspects of the study design. For instance, in the reading group example, potentially confounding family background variables could be controlled for by gathering information on parental education levels, prior reading levels of students, and so on.[109] The assumptions required to assume a causal effect are stronger than those needed to establish causality through randomization. This is not to say this type of research is ineffective, just that stronger assumptions are required to claim an unbiased, causal effect has been found.

4. Assumptions in Natural Experiments

Fourth, "natural experiments" occur when events beyond a researcher's control create situations which mimic aspects of experimental design. With such experiments, we leave behind the experimental setting in which the researcher has caused (or observed, in the case of quasi-experiments) direct manipulation of research parameters. We label these natural experiments "strong" when the source of

107. See Kaushik Basu, *New Empirical Development Economics: Remarks on Its Philosophical Foundations*, 40 ECON. & POL. WKLY. 4336, 4336–37 (2005).

108. SHADISH, COOK, & CAMPBELL, *supra* note 28, at 14.

109. *Id.* at 15.

variation is plausibly unrelated to the subject matter of the study and "control" groups may plausibly serve as a valid counterfactual. A potential example of these are the "difference-in-difference" studies considered above. There, the researchers did not directly manipulate unemployment rates or education levels, but rather relied on the assumption that the policies which did create such variation between states were independent of the individual level choices faced by those affected by the policy.[110]

The assumptions needed to establish causal estimates in this context begin to be strong, for at least three reasons. First, there is no randomization that plausibly divides individuals into treatment and control groups in ways uncorrelated with confounding variables.[111] Because of this, selection bias becomes a concern. Second, the

110. Courts may find evidence based on DID and other non-experimental evidence convincing, while economic theorists might be more skeptical. This reflects a foundational divide among researchers themselves. Empiricists may generally find assumptions behind these models reasonable, for example, in Clay, Lingwall, & Stephens, Jr., *supra* note 61, while econometric theorists might argue that precise causal estimates are unlikely to be found from these methods. *E.g.*, Low & Meghir, *supra* note 77, at 41. Low and Meghir write:

> [S]uppose we want to estimate the effects of an intervention to increase the years of education. The difference-in-differences approach assumes that the level of education (in the absence of intervention) will be a strictly monotonic function of just one unobservable. Education is typically driven by the comparison of the benefits of education and the costs of education . . . . The costs are also likely to be heterogeneous. So the education choice will generally depend on at least two unobserved components, which are unlikely to collapse into one element of heterogeneity. In this case, . . . a difference-in-differences analysis of an intervention will not have a causal interpretation.
>
> . . . These models look simple and straightforward, but the interpretation of their results as causal impacts rely on strong behavioral and functional-form assumptions.

Id.

111. Low & Meghir, *supra* note 77, at 41 ("Not all treatment effect models are created equal: it is important to distinguish those estimated through randomized experiments from those estimated through quasi-experimental methods, such as difference-in-differences, regression discontinuity, matching, and others. The point of randomized experiments is that results do not depend on strong assumptions about individual behavior . . . .").

researcher is not involved in the creation of the design, as they are relying on events outside their control which may be influenced by a host of potentially confounding real-world factors. Finally, these studies often involve large policy issues decided at national or state levels, which again suggest a wealth of potentially confounding variables related to policy, such as the composition of state legislatures, and so on.[112]

5. Assumptions in Weak Natural Experiments

Fifth, natural experiments with weak designs may provide causal estimates under only very strong assumptions. For example, while neighboring states may serve as plausible controls for a state-level policy change, relying on distant states to provide a reasonable counterfactual may be less realistic.[113] With these issues, the researcher must rely on very strong assumptions about causality, or carefully control for each factor that would plausibly impact causality. Because the assumptions required to establish causality in a weak natural experiment begin to be very strong, whether the estimates provided by the researcher show a causal effect must be examined with a great deal of scrutiny.

6. Assumptions in Observational Studies

Sixth, pure observational studies with no plausible source of exogenous variation require incredibly strong—or heroic—assumptions to establish causal estimates. These studies, which are essentially just reporting correlations, would need to have precise and well-observed mechanisms behind any kind of causal claim that could be made. This is not to say that observational studies may not provide evidence of causality, but rather that the quantitative analysis behind causal claims

112. Together with strong natural experiments, we would list Granger causality as providing reasonably credible pathways to causal estimates in a setting in which the researcher has not directly manipulated the subject matter.

113. *See, e.g.*, Melvin Stephens Jr. & Dou-Yan Yang, *Compulsory Education and the Benefits of Schooling*, 104 AM. ECON. REV. 1777, 1777 (2014) (finding that comparisons between states nationally rather than within regions produced inaccurate estimates of the effect of education laws).

would need to be rigorous and assumptions made very clear. For example, whether smoking causes lung cancer is a topic in which it is unethical to run experiments on human beings, yet a host of observational correlations—coupled with medical knowledge that provides convincing mechanisms relating tobacco use and cancer—make the claim that smoking causes cancer plausible.[114]

***

Table 1, below, shows the steps on the causality spectrum.[115] As discussed above, the *weaker* the assumptions needed to justify causality, the *stronger* the causal evidence.[116]

Table 1: Causal Assumptions in Common Social-Scientific Research Designs

114. *See* Lawrence A. Loeb, Virginia L. Emster, Kenneth E. Warner, John Abbotts & John Laszlo, *Smoking and Lung Cancer: An Overview*, 44 CANCER RSCH. 5940 (1984).

115. By providing the list of study designs in Table 1, we do not intend to say that these encompass the universe of causal thinking. In particular, causal thinking in algorithm design from computer scientists provides fascinating ways to model and describe causality—ways that may improve the generally social-scientific techniques described in the table. *See, e.g.*, JONAS PETERS, DOMINIK JANZING, & BERNHARD SCHÖLKOPF, ELEMENTS OF CAUSAL INFERENCE: FOUNDATION AND LEARNING ALGORITHMS (2017).

116. There may be social situations of such tremendous complexity, particularly in the way feedback is created between various aspects of a study design, that deciphering what truly causes something else may be impossible. The science of complexity theory suggests that in this type of scenario, unless a researcher can reconstruct the development of a system as a whole, they likely do not understand the causal mechanisms at play. An example might be recreating an entire biosphere or the evolutionary development of an organism. The number of causal factors, each interrelated with other factors, may make understanding causality impossible without being able to recreate such a system from scratch. Taken further, complexity theorists may reject causal conclusions from many social-scientific settings, believing that these problems make causal inference impossible. This might be the case in many scrutiny situations faced by courts considering statistical evidence. This more extreme conclusion would hinder courts from considering much potentially useful evidence, and so we leave mention of it to the end of our causal list and do not include it in the table below. *See* Andreas Wagner, *Causality in Complex Systems*, (Santa Fe Inst., Working Paper No. 1997-08-075, 1997), https://sfi-edu.s3.amazonaws.com/sfi-edu/production/uploads/sfi-com/dev/uploads/filer/e1/57/e157031b-1423-47b2-b3a1-efc67bacb59e/97-08-075.pdf.

Type of Research Design	**Observational studies with no causal design**	**Weak natural experiments**	**Strong natural experiments**	**Quasi-experiments**	**Weak experimental evidence**	**Large, randomized control trials**
Typical Assumptions Required to Establish Causality	**Heroic** Most researchers would not find assumptions satisfied	**Strong** Many researchers may question whether assumptions are violated	**Medium** Often considered "reasonable" by researchers	**Medium** Often considered "reasonable" by researchers	**Medium** Often considered "reasonable" by researchers	**Weak** Assumptions are typically considered satisfied, and hence "reasonable"
Strength of Causal Evidence	**Weak**	**Weak**	**Medium**	**Medium**	**Medium**	**Strong**

III. CORRELATION, CAUSATION, AND CONSTITUTIONAL SCRUTINY

This Section develops the legal counterpart to the history of causal analysis in Section II by outlining the development of constitutional scrutiny analysis. The general theme is a long history of balancing principles of judicial restraint with constitutional concerns. It then discusses how courts have viewed causation and statistical evidence in general. Finally, it combines those two concepts by examining how constitutional scrutiny claims have treated causal evidence. In the context of statistical evidence, the general theme from the development of scrutiny is reversed: courts have often used restraint with regards to statistical evidence to *strike down* legislation.

A. The Development of Constitutional Scrutiny Analysis

When a court overturns a law by finding it unconstitutional, the scrutiny given the law is based on centuries of constitutional precedent. This precedent reflects courts' struggles to cabin how they police the relationship between legislative actions and constitutional principles. Article VI of the Constitution provides that the Constitution is the supreme law of the land and tasks the judiciary with stating what the law

is.[117] Whether this included the judicial ability to overturn legislation that conflicted with the Constitution was debated early on. Some believed that this was a power left to the legislative branch, and that it was just as likely that judges would "declare laws unconstitutional which are not so, as it is that the legislature will exceed their constitutional authority."[118] In fact, judges early in the nineteenth century were *impeached* for striking down acts of the legislature.[119] If there was a scope for judicial review, it would necessarily be narrow: "there may be cases so monstrous,— e.g., an Act authorizing conviction for crime without evidence, or securing to the legislature their own seats for life,—'so manifestly unconstitutional that it would seem wrong to require the judges to regard it in their decisions.'"[120] For an example of this limited view of judicial review, in 1811 Virginia, the state supreme court acknowledged its power to declare an act of the state legislature void but declined to do so. The court rested its decision on a widespread rule that "an Act of the legislature is not to be declared void unless the violation of the constitution is so manifest as to leave no room for reasonable doubt."[121]

This quite limited view of judicial review was "more than . . . a mere expression of courtesy and deference," it narrowed the power of

117. "An act of congress repugnant to the constitution cannot become a law." Marbury v. Madison, 5 U.S. 137, 138 (1803). This is because the legislature is limited in its power to alter the constitution and if a court could not declare an unconstitutional act void, the constitution would not be worth the parchment on which it was drafted.

118. James B. Thayer, *The Origin and Scope of the American Doctrine of Constitutional Law*, 7 HARV. L. REV. 129, 134 (1893) (quoting 1 ZEPHANIAH SWIFT, SYSTEM OF THE LAWS OF CONNECTICUT 52 (1795)); *see also* Richard A. Posner, *The Rise and Fall of Judicial Self-Restraint*, 100 CALIF. L. REV. 519, 522 (2012).

119. Thayer, *supra* note 118, at 134.

120. *Id.*

121. *Id.* at 140. (quoting Commonwealth *ex rel.* O'Hara v. Smith, 4 Binn. 117, 123 (Pa. 1811)). Thayer quotes a long line of precedent along these lines, *id.* at 142–43, such as from Georgia in 1808, *Grimball v. Ross*, Charlton 175, 178 (Ga. 1808) ("But when it remains doubtful whether the legislature have or have not trespassed on the constitution, a conflict ought to be avoided, because there is a possibility in such a case of the constitution being with the legislature."), and South Carolina in 1812, *Adm'rs of Byrne v. Adm'rs of Stewart*, 3 Des. Eq. 466, 477 ("The validity of the law ought not then to be questioned unless it is so obviously repugnant to the constitution that when pointed out by the judges, all men of sense and reflection in the community may perceive the repugnancy.").

the judiciary to void a legislative act to only when "those who have the right to make laws have not merely made a mistake, but have made a very clear one—so clear that it is not open to rational question."[122] Application of this rule required acknowledging that what may seem unconstitutional to one can reasonably be considered constitutional by another. Since reasonable minds can differ as to which choice is the best option, the legislature has the power to choose any of the options, if they are rational and potentially constitutional.[123] Judicial power was secondary to the exercise of legislative power over policy decisions, and the legislature did not have to adapt their legislation to reflect what the judiciary believed was "prudent or reasonable legislation."[124] Since the power of courts is strictly a judicial one, the judiciary could not deprive the legislature of their power to make discretionary decisions.[125] Thus, in situations where the legislature must choose between policies, the judiciary did not have the power to interfere in that decision.[126] Rather, "the judicial function is merely that of fixing the outside border of reasonable legislative action, the boundary beyond which the taxing power, the power of eminent domain, police power and legislative power in general, cannot go without violating the prohibitions of the constitution or crossing the line of its grants."[127]

Tension between limited judicial line-drawing and judicial restraint vis-à-vis legislation in the modern world emerged in *Lochner v. New York*.[128] The Supreme Court struck down a statute prohibiting

122. Thayer, *supra* note 118, at 143–44.

123. *Id.*

124. *Id.* at 148.

125. *Id.* at 135.

126. *Id.*

127. *Id.* at 148. In this view, the post-hoc nature of judicial review further limited the scope of judicial power. The judiciary does not have a role in determining whether a legislative action is constitutional unless and until someone challenges the action in court. The result is that harmful or unconstitutional legislative acts can take effect and the judiciary will not have an opportunity to stop them. Since the judiciary was not given the power to revise laws before they become effective, and have limited opportunities to review legislative action, the scope of their power would be narrow. *Id.* at 137.

128. Lochner v. New York, 198 U.S. 45 (1905). *Lochner* was reaffirmed by a number of cases, including *Coppage v. Kansas*, 236 U.S. 1 (1915), in which a Kansas law that forbid employers from conditioning employment upon a promise not to join

bakeries from requiring or permitting their employees to work more than sixty hours per week or ten hours per day on the basis that it violated the right to contract under the Fourteenth Amendment. The Court distinguished contracts that violate a statute enacted as a legitimate exercise of police power (e.g., contracts for unlawful acts), as well as dangerous workplace environments where employee safety warranted the legitimate exercise of police power.[129] In a dissenting opinion, Justice Holmes—reflecting the traditional view—noted that the Constitution was not drafted to support one economic theory over another, and that the majority had allowed their disagreement with the economic theory supported by the legislature to influence their decision.[130] Justice Holmes asserted that the proper test is not whether the judiciary agreed or disagreed with the economic theory, but whether a reasonable or rational person would admit the statute infringed on fundamental principles.[131]

Thirty years later, in *Nebbia v. New York*, the Court agreed with Holmes and captured the idea of rational basis review of judicial actions in the economic arena.[132] In determining whether the state had exceeded its police powers in enacting and enforcing a statute that fixed the price of milk, the Court stated that due process, under the Fifth and Fourteenth Amendments, requires only that the legislation is not "unreasonable, arbitrary or capricious and that the means selected have a real and substantial relation to the object sought to be attained."[133] In contrast to *Lochner*, the Court stated:

> So far as the requirement of due process is concerned, and in the absence of other constitutional restriction, a state is free to adopt whatever economic policy may reasonably be deemed to promote public welfare, and to enforce that policy by legislation adapted to its purpose

or retain membership of a labor organization was struck down, and *Adkins v. Children's Hospital of the District of Columbia*, 261 U.S. 525 (1923), in which a federal statute prescribing minimum wage for women in Washington, D.C., was struck down as a violation of due process.

129. *Lochner*, 198 U.S. at 75 (Holmes, J., dissenting).

130. *Id.* at 76.

131. *Id.*

132. Nebbia v. New York, 291 U.S 502 (1934).

133. *Id.* at 525.

> . . . . If the laws passed are seen to have a reasonable relation to a proper legislative purpose, and are neither arbitrary nor discriminatory, the requirements of due process are satisfied . . . . With the wisdom of the policy adopted, with the adequacy or practicability of the law enacted to forward it, the courts are both incompetent and unauthorized to deal . . . . Times without number we have said that the Legislature is primarily the judge of the necessity of such an enactment, that every possible presumption is in favor of its validity, and that though the court may hold views inconsistent with the wisdom of the law, it may not be annulled unless palpably in excess of legislative power. [134]

Thus, the emergence of rational basis review hearkened back to a long tradition of judicial restraint against exercising power to strike down laws.

However, within a decade of *Nebbia* a more expansive scope of judicial review began to emerge. The famous footnote four of *Carolene Products* suggested cases in which there may be stricter scrutiny: "There may be a narrow scope for operation of the presumption of constitutionality when legislation appears on its face to be within a specific prohibition of the Constitution, such as those of the first ten Amendments."[135] The Court acted on this suggestion in 1942's *Skinner v. Oklahoma* with regard to a state statute which provided for the sterilization of "habitual criminals" who committed two or more felonies.[136] The court struck down the law under strict scrutiny because "[m]arriage and procreation are fundamental to the very existence and survival of the race[,]" and discrimination "as if it had selected a particular race or nationality for oppressive treatment" has taken place "[w]hen the law

134. *Id.* at 537–38 (internal references omitted).

135. United States v. Carolene Prods. Co., 304 U.S. 144, 152 n.4 (1938). *Carolene* also suggested requiring stricter scrutiny for restrictions on political processes, statutes directed at particular religious or racial minorities, and prejudice against "discrete and insular minorities." *Id.*

136. Skinner v. Oklahoma, 316 U.S. 535 (1942).

lays an unequal hand on those who have committed intrinsically the same quality of offense and sterilizes one and not the other."[137]

Over time, the existing classifications for intermediate and strict scrutiny gradually emerged as courts wrestled with balancing government policies against a variety of rights. For example, one important step along this path to modern scrutiny analysis was *Brown v. Board of Education* in 1954.[138] Although *Brown* did not use the language of scrutiny analysis, by holding that "separate educational facilities are inherently unequal" based in part on social-scientific evidence, the Court took a significant step towards creating heightened scrutiny levels supported by empirical evidence (though the particular evidence would perhaps not stand the test of time).[139] In 1976, the modern intermediate scrutiny test was established in *Craig v. Boren*, in which an Oklahoma statute prohibiting the sale of 3.2% beer to males under 21 and to females under 18 was challenged under equal protection.[140] The

137. *Id*. at 541.

138. Brown v. Board of Education, 347 U.S. 483 (1954). Through its notable footnote 11, which invoked a variety of social-scientific evidence, the Court also paved the way for increased multi-disciplinary evidence in the law. Michael Heise, Brown v. Board of Education*, Footnote 11, and Multidisciplinarity*, 90 CORNELL L. REV. 279, 280 (2005). Although "much-maligned" due to limitations in the research cited, *id.* at 293–94, the footnote helped draw empirical analysis into the field. *Id.* at 296. Finally, as far as *Brown* can be viewed through the lens of strict scrutiny, many scholars would suggest that the Court's analysis did not need to reach through to empirical evidence and that application of the Equal Protection clause without such evidence was sufficient. *Id*. at 295, 295 n.91 (citing JACK M. BALKIN, BRUCE ACKERMAN, DERRICK A. BELL, DREW S. DAYS III, JOHN HART ELY, CATHARINE A. MACKINNON, MICHAEL W. MCCONNELL, FRANK I. MICHELMAN, & CASS R. SUNSTEIN, INTRODUCTION TO WHAT *BROWN V. BOARD OF EDUCATION* SHOULD HAVE SAID 44–53 (Jack M. Balkin ed., 2001) (noting that modern constitutional law scholars would not generally rely on the statistical evidence in Footnote 11)). In terms of our Table 2 analysis, this suggests that the reasoning underlying *Brown*'s finding of the harms of segregation required few inferential steps, feedback loops, or confounding causes, and hence statistical evidence used, if any, merely needed to support a plausible inference of causality. *See infra* Table 2.

139. *Brown*, 347 U.S. at 495; *see* David Zimmerman, *Five Supreme Court Constitutions: Race-Based Scrutiny Past, Present, and Future*, 10 BYU J. PUB. L. 161, 165–66 (1996).

140. Craig v. Boren, 492 U.S. 190 (1976) (citing Reed v. Reed, 404 U.S. 71 (1971)). The development of intermediate scrutiny in *Craig* will be addressed in *infra* Section III.C.

Court acknowledged that enhancing traffic safety is an important governmental objective, but rejected that the gender-based distinctions underpinning the statute's differential in age were substantially related to achieve that objective.[141]

In sum, the long history of scrutiny analysis is one of traditional restraint followed by an expansive version of judicial review in certain circumstances. As existing frameworks were found insufficiently tailored, new criteria were developed which drew in social-scientific evidence in selective ways. The next subsection examines how courts have typically treated causal issues before examining causal issues in scrutiny cases themselves.

B. Correlation and Causation in Court

Statistical relationships outside, and often inside, laboratories are complex. This is because attempts to codify real-life phenomena in statistical terms are inherently fraught: real life scientific relationships, particularly in the social sciences, often resist simple analysis. As discussed above, to deal with this real-life messiness, the sciences have adopted terminology meant to summarize what one knows about relationships through the artful use of two contrasting terms: correlation and causation. This subsection examines how courts have used these terms before addressing causal issues in the constitutional context.[142]

Law is steeped in the business of proving and disproving causality.[143] Because causation is embedded at the heart of legal matters large and small, litigators and courts have deep experience approaching causation from a legal perspective.[144] Indeed, outside of philosophers

141. *Craig*, 492 U.S. at 197–99 (concluding that the state's statistics, concerning the relationship between gender and driving under the influence, were too weak to support the gender-based distinctions in the statute).

142. Our goal is not to provide a positive taxonomy of causal or statistical errors courts may make. Rather, the examples of scrutiny cases suggest the need for a normative framework through which to process statistical evidence.

143. *See supra* note 3 and accompanying text.

144. Although beyond the scope of this Article which focuses on improving the framework under which causal evidence is evaluated, studying the process by which probabilistic and causal evidence empirically translate into judicial decision-making remains a crucial area for future study.

and other specialized academics, there are likely no others who deal with causation so routinely in their professions. Because of the practical need for courts to resolve the business of everyday life expeditiously, causation in law has evolved very differently than in the hard and social sciences. This has produced disconnects in vocabulary and approaches to evidence between disciplines.[145] Unlike in many scientific studies, causality may be proven in court through several non-statistical and non-experimental techniques, and if statistical evidence is brought to bear, assumptions behind causal claims are treated differently than they would be in scientific journals.[146]

As a starting point for attempting to reconcile these views, consider how courts treat the idea of correlation versus causation. First, courts are correct to distinguish these ideas. For instance, in litigation over flooded property in Michigan, one party argued that modifications to a property preceded a flood and hence caused it. The plaintiff's expert "opined that the insufficiently-sized culverts that [the defendant]

145. One potential reason for these disconnects may be that the practice of law is fundamentally about persuasion—persuading courts to rule a certain way in litigation, persuading parties to settle in mediation, and so on. Attempts to persuade may rely on any number of non-statistical theories of causality, and as data or causal evidence enters a field focused on persuasion, it takes on the persuasive role to which it is put by attorneys.

146. For example, causation in courts may be based on such non-experimental evidence as differential etiology analysis. *See, e.g.*, Brown v. Burlington N. Santa Fe Ry. Co., 765 F.3d 765, 772 (7th Cir. 2014) (holding that differential etiology, is an accepted methodology for an expert to render an opinion about the cause of an ailment if the expert uses reliable methods); Hendrix *ex rel.* G.P. v. Evenflo Co., 609 F.3d 1183, 1195 (11th Cir. 2010). Some courts, moreover, are beginning to require more than mere time and sequence reasoning to determine causality. *See, e.g.*, Michael Hooker, Guy P. McConnell, & Jason A. Pill, *Have We Reached the Tipping Point? Emerging Causation Issues in Data-Breach Litigation*, 94 FLA. BAR J. 8 (2020) (explaining how courts are beginning to require more than mere time and sequence reasoning to determine the sufficiency of pleadings in data breach litigation).

Much of the litigation over expert testimony under *Daubert* concerns what is sufficient proof of causal mechanisms along these lines. *See, e.g.*, Kumho Tire Co. v. Carmichael, 526 U.S. 137, 141 (1999) (holding that *Daubert*'s "gatekeeping" obligation applies to all expert testimony beyond "scientific" knowledge, including "technical," and "other specialized" knowledge); C.W. *ex rel.* Wood v. Textron, Inc., 807 F.3d 827, 835 (7th Cir. 2015) (holding that the district court properly adhered to the *Daubert* guidelines in conducting a review of, and subsequently rejecting, the studies that experts in the case relied upon in generating their differential etiology).

had placed 'definitely' were 'the cause of flooding.'"[147] The court rejected this contextualization, noting that the expert's argument "is simply a logical fallacy of concluding that correlation is causation: that because [defendant] made modifications to her property prior to the flood, those modifications must have caused the flood."[148]

The mantra that "correlation is not causation" has been repeated by multiple courts at multiple times. The District of New Mexico provides a good example. In a case discussing the effect of teen sexting on other behavior, the court cited research that "underscored the correlation between sexting and other sexual activity . . . although it could not determine causality" and that "while consensual sexting is related to risk behaviors, causality is unknown."[149] Numerous other courts make this distinction, noting that, e.g., "[m]ere correlation does not demonstrate causation,"[150] that "'[c]orrelation does not prove causation [.] . . . Just because two things move together, doesn't mean the one is pushing the other; the sun doesn't rise every day because the rooster crowed,"[151] and that "[i]n law, as in science, '[c]orrelation is not causation.'"[152]

147. Bechtol v. Allen, No. 307716, 2013 WL 5857642, at *2 (Mich. Ct. App. Oct. 31, 2013).

148. *Id.* at *4. The court continued, noting "[w]hile correlation may be suggested, the evidence also shows that the old drainage tile system . . . had gone unmaintained for decades, [plaintiffs] were already beginning to have water problems before [defendant] started her modifications, [and] the rainfall on the date of the flood was exceptional . . . ." *Id.*

149. United States v. Streett, 434 F. Supp. 3d 1125, 1187 (D.N.M. 2020).

150. Sergeants Benevolent Ass'n Health & Welfare Fund v. Sanofi-Aventis U.S. LLP, 806 F.3d 71, 92 (2d Cir. 2015) (holding that correlation between issuance of a public health advisory and decline in a drug's sales does not necessarily mean that the decline in sales was *caused* by the advisory).

151. *In re* Gonzalez, No. 8:12-bk-19213, 2019 WL 1087093, at *8 n.114 (Bankr. M.D. Fla. Mar. 5, 2019) (quoting M. Elizabeth Karns, *Statistical Misperceptions*, 47-JUN FED. LAW. 19, 20 (2000)).

152. Manuel v. Pepsi-Cola Co., No. 17 CIV. 7955, 2018 WL 2269247, at *11 (S.D.N.Y. May 17, 2018), *aff'd*, 763 F. App'x 108 (2d Cir. 2019) (citing Norfolk & W. Ry. Co. v. Ayers, 538 U.S. 135, 173 (2003) (Kennedy, J., concurring in part and dissenting in part)); *see also, e.g.*, Norris v. Baxter Healthcare Corp., 397 F.3d 878, 885 (10th Cir. 2005) (noting that "[a] correlation does not equal causation" when discrediting plaintiff's expert testimony from two doctors who claimed that breast implants caused plaintiff's autoimmune disease based on their own observations,

At the same time, courts often repeat that correlation may provide *evidence* of causation. "Although correlation alone may be insufficient to establish causation . . . , it is nonetheless relevant to identifying causal relationships. Indeed, it may be 'a necessary but not sufficient condition for causation.'"[153] As such, "scientific induction *means* inferring causality from correlation."[154] Indeed, one court showed its philosophical chops by hearkening to Hume in its causal analysis, noting:

> But-for causation is an inference drawn from regularly observed correlation. But it must be stressed that causation is an inference, not an observation, as philosophers since at least Hume have reminded us. The only empirical facts that we can discover about the world are facts about correlation. We cannot observe causal relationships . . . . What we observe is correlation, and when we see it regularly enough, we hypothesize causation.[155]

contrary to defendant's epidemiological evidence which found there is no general causational relationship between breast implants and immune system diseases); Tagatz v. Marquette Univ., 861 F.2d 1040, 1044 (7th Cir. 1988) (noting that "[c]orrelation is not causation" when holding that Plaintiff's data did not show that the difference in employees' salaries was due to those employees having a particular attribute, because members of the group with that attribute had many other things in common which could have caused the difference in salaries); Brumbaugh v. Sandoz Pharm. Corp., 77 F. Supp. 2d 1153, 1157 (D. Mont. 1999) ("Correlation of two events in time does not necessarily establish causation.").

153. Etherton v. Owners Ins. Co., 829 F.3d 1209, 1220–21 (10th Cir. 2016) (quoting JOSEPH F. HEALEY, THE ESSENTIALS OF STATISTICS 350 (4th ed. 2015)) (determining that a doctor's testimony did not mistake correlation for causation when he used a temporal relationship between an injury and a purported cause because his analysis did not rely solely on correlation).

154. *Tagatz*, 861 F.2d at 1044 (holding that plaintiff's evidence of difference in salary for employees with a certain attribute in common shows that it is possible that the employer was discriminating against employees lacking that attribute).

155. United States v. Mosley, 454 F.3d 249, 266 (3d Cir. 2006) (determining whether the court should hypothesize a causal relationship between the seizure of a passenger in a vehicle and discovery of evidence); *see also* Albuquerque Cab Co., v. Lyft, Inc., 460 F. Supp. 3d 1215, 1225 (D.N.M. 2020) (noting that "correlation is evidence of causation" when holding that plaintiff's factual allegations of correlation between defendant entering the taxi market and plaintiff's revenue decrease made

As seen in these examples, courts traffic regularly with the concept of correlation versus causation, and—given the practical limitation involved in most cases—appropriately situate correlation as being able to potentially provide evidence *of* causation, particularly when paired with further evidence to strengthen causal claims. Perhaps because courts traffic regularly with these concepts, when confronted with evidence that approaches causality from a more technical aspect, they may fail to scrutinize it with the care it deserves. The next subsection explores how courts have used social scientific evidence beyond correlation in the constitutional context.

C. Social-Scientific Evidence and Causal Reasoning in Constitutional Claims

In this subsection, we consider how courts have applied causal reasoning and statistical evidence in constitutional scrutiny cases. After considering several examples, we return to the motivating example of university firearm restrictions from the introduction.[156] In the following Section, once a framework is developed for analyzing scrutiny cases relying on statistical evidence, we will return to these cases.

The most prominent case considering statistical evidence and causality in the constitutional context is *Brown v. Entertainment*

plaintiff's causation claim plausible to satisfy pleading requirements); Ohio Valley Env'l Coal., Inc. v. U.S. Army Corps of Eng'rs, No. 2:12-6689, 2014 WL 4102478, at *18 n.14 (S.D.W. Va. Aug. 18, 2014) (finding that "[i]t is true that correlation does not necessarily imply causation. Yet, science proceeds through finding correlations, and causation may in a given instance be inferred therefrom by ruling out certain hypotheses[,]" and noting that even though defendant was justified in disregarding articles showing correlation between coal mining and adverse health effects because they reasonably determined the articles were not within their scope of review, the court's finding should not be construed as an endorsement that the studies were incorrect). In at least one area of the law, courts have rejected even the requirement of a correlation, let alone causality. *E.g.*, Prison Legal News v. Sec'y, Fla. Dep't of Corr., 890 F.3d 954, 968 (11th Cir. 2018) ("We have rejected the misconception that prison officials are required to adduce specific evidence of a causal link between a prison policy and actual incidents of violence . . . . Requiring proof of such a correlation constitutes insufficient deference to the judgment of the prison authorities with respect to security needs.") (internal punctuation and citations removed).

156. *See supra* note 12 and accompanying text.

Merchants Ass'n.[157] California had imposed a statutory restriction on violent video game sales to minors.[158] This included "killing, maiming, dismembering, or sexually assaulting an image of a human being" with some restrictions.[159] The act was challenged on First Amendment grounds and found its way to the Supreme Court. As California's law consisted of a content-specific restriction on speech, the Court proceeded with a strict scrutiny analysis. To survive, the state would need to show "a compelling government interest" which the act was "narrowly drawn to serve."[160] This required showing an "actual problem" in need of solving, with government action being "actually necessary to the solution."[161]

California failed to make this showing, based in part on the type of causal evidence brought to court. "At the outset, [California] acknowledges that it cannot show a direct causal link between violent video games and harm to minors."[162] The state argued that it need not produce such evidence—claiming an inference by the legislature that such a link existed would suffice.[163] The Court rejected this view and tied the level of statistical evidence to the level of scrutiny. Such a posited link by the legislature could have satisfied an intermediate scrutiny case in which a content-neutral restriction was at issue, but in a strict scrutiny case "ambiguous proof" was insufficient.[164] The state had relied on psychological research that purported "to show a connection between exposure to violent video games and harmful effects on children."[165] This research failed to establish causal relationships and so failed to satisfy strict scrutiny:

157. 564 U.S. 786 (2011).

158. *Id.* at 789.

159. *Id.*

160. *Id.* at 799.

161. *Id.* (quoting in part United States v. Playboy Ent. Grp., Inc., 529 U.S. 803, 822–23 (2000)).

162. *Id.*

163. *Id.*

164. *Id.* at 800.

165. *Id.*; *see also* United States v. Alvarez, 132 S. Ct. 2537, 2540 (2012) ("[T]he First Amendment requires that there be a direct causal link between the restriction imposed and the injury to be prevented.").

> They do not prove that violent video games *cause* minors to *act* aggressively (which would at least be a beginning). Instead, "[n]early all of the research is based on correlation, not evidence of causation, and most of the studies suffer from significant, admitted flaws in methodology." They show at best some correlation between exposure to violent entertainment and minuscule real-world effects, such as children's feeling more aggressive or making louder noises in the few minutes after playing a violent game than after playing a nonviolent game.[166]

This holding—that the level of causal evidence brought through statistical analysis should be correlated with the level of scrutiny applied—is striking. The holding creates a searching evidentiary standard for statistics akin to the demanding legal reasoning in strict scrutiny analysis—essentially using causal statistical standards to continue to move away from early views of deference to legislative acts.

Despite this, *Entertainment Merchants Ass'n* has received little analytical follow-up from courts. Of the variety of cases citing this section of *Entertainment Merchants Ass'n* that include the words "correlation," "cause," or its variants, *Entertainment Merchants Ass'n*'s application is widely varied, and numerous citing cases simply focus on the question of First Amendment protection for videogames.[167] Some courts have applied the causality standard to statistical evidence in like manner as *Entertainment Merchants Ass'n*, such as the Eastern District of California in *Welch v. Brown*. There, the court examined a restriction on sexual orientation change efforts for minors.[168] "[E]vidence that [such efforts] 'may' cause harm to minors based on questionable and scientifically incomplete studies that may not have included minors is unlikely to satisfy the demands of strict scrutiny."[169] Or when the District of Maryland considered the constitutional validity of a requirement that a Limited Service Pregnancy Resource Center

166. *Ent. Merchs. Ass'n*, 564 U.S. at 800 (citation omitted).

167. Westlaw returns 51 cases matching these criteria, as of January 2021.

168. Welch v. Brown, 907 F. Supp. 2d. 1102, 1120 (E.D. Cal. 2012), *rev'd on other grounds sub nom.* Pickup v. Brown, 740 F.3d 1208, 1222 (9th Cir. 2014), *abrogated by* Nat'l Inst. of Fam. & Life Advocs. v. Becerra, 138 S. Ct. 2361 (2018).

169. *Id.*

(LSPRC) post information disclaiming its lack of licensed medical professionals, the court noted:

> But as with California's violent video game law, when core First Amendment interests are implicated, mere intuition is not sufficient. Yet that is all the County has brought forth: intuition and suppositions. "This is not to suggest that a 10,000–page record must be compiled in every case or that the [g]overnment must delay in acting to address a real problem; *but the [g]overnment must present more than anecdote and supposition . . . ."* The County has not demonstrated how the practices of LSPRCs are *causing* the harm it has a compelling interest in addressing.[170]

On the other hand, some courts have cited *Entertainment Merchants Ass'n*'s causality standard and then proceeded without statistical evidence at all, arguing essentially that such evidence is not needed when dealing with "billiard-ball" causality questions in which cause and effect are physically proximate or temporally immediate. For example, applying *Entertainment Merchants Ass'n* and its causal standard to whether an injunction against displaying graphic anti-abortion imagery to children satisfied the First Amendment, the Colorado Court of Appeals in *St. John's Church in the Wilderness v. Scott* found anecdotal evidence sufficient to establish causality.[171] The court failed to

170. Tepeyac v. Montgomery Cnty., 5 F. Supp. 3d 745, 769 (D. Md. 2014) (emphasis on "causing" added) (citation omitted); *see also* Driscoll v. Stapleton, 473 P.3d 386, 398–400 (Mont. 2020) (Sandefur, J., concurring in part and dissenting in part) (applying *Entertainment Merchants Ass'n* to voting suppression facts and finding that "the evidentiary record in this case is devoid of any substantial non-speculative evidence that [such policy] will likely be a cause of any decrease in absentee voter turnout"); Susan B. Anthony List v. Ohio Elections Comm'n, 45 F. Supp. 3d 765, 776, 776 n.6 (2014) (noting *Entertainment Merchants Ass'n* "requir[es] empirical proof of a 'direct causal link' . . . not just 'some correlation'" and finding that the issue before it was "inherently difficult to quantify").

171. Saint John's Church in the Wilderness v. Scott, 296 P.3d 273, 283–84 (Colo. App. 2012) (citing past precedent on protecting children from psychological harm and then noting that (1) parents were concerned, (2) images were highly disturbing to children, and (3) the priest's daughter buried her face in her hymnal and remained upset over the course of several days).

cite statistical evidence, perhaps because the link between displaying the imagery and the effect on children was so immediate and pronounced.[172] Or, when the Minnesota Supreme Court in *State v. Melchert-Dinkel* considered whether a speech assisting another to commit suicide was constitutionally protected, the court noted *Entertainment Merchants Ass'n* and then proceeded without statistical evidence because "[p]rohibiting only speech that assists suicide, combined with the statutory limitation that such enablement must be targeted at a specific individual, narrows the reach to only the most direct, causal links between speech and the suicide."[173]

In our introductory example of strict scrutiny analysis of university firearm restrictions in *State* ex rel. *Schmitt v. Mun Choi*, on appeal the State argued that *Entertainment Merchants Ass'n* applied and strict scrutiny analysis required evidence of causation: "weak" statistical evidence was insufficient.[174] But rather than hold to *Entertainment Merchants Ass'n*'s standard of causal evidence from the United States Supreme Court, the Missouri courts looked to the their own Supreme Court's rather opposite standard, that "the ever-changing body of science and statistics is ill-suited to constitutional analysis."[175] Rather than statistics, "simple common sense," history, and consensus were sufficient to satisfy strict scrutiny.[176] In particular, statistical evidence was insufficient to overcome *anecdotal* evidence.[177] *Entertainment Merchants Ass'n*'s standard did not apply.

In cases pre-*Entertainment Merchants Ass'n* or outside the strict scrutiny context, statistical evidence with bearing on causality is common. As with courts' application of *Entertainment Merchants Ass'n*, courts have also treated this evidence in very different ways. For example, a strong counterpart to *Entertainment Merchants Ass'n*'s call for causal evidence was the Supreme Court's dicta on statistical

172. *See id*.

173. State v. Melchert-Dinkel, 844 N.W.2d 13, 23 (Minn. 2014).

174. Brief of Appellant at 60, State *ex rel.* Schmitt v. Mun Choi, 627 S.W.3d 1 (Mo. Ct. App. 2021).

175. State v. Merritt, 467 S.W.3d 808, 814 n.6 (Mo. 2015) (en banc).

176. *Mun Choi*, 627 S.W.3d at 17 n.13.

177. State *ex rel.* Schmitt v. Mun Choi, Nos. 16BA-CV03144, 16BA-CV02758, 2020 WL 5093608, at *8 (Mo. Cir. Ct. 2020).

evidence when establishing intermediate scrutiny in *Craig v. Boren*.[178] In *Craig*, the Court considered statistical evidence of gender differences in drinking arrests in Oklahoma. The state introduced evidence that showed a higher rate of male arrests (2%) than female arrests (0.18%) for driving while under the influence of alcohol.[179] The Court noted "[w]hile such a disparity is not trivial in a statistical sense, it hardly can form the basis for employment of a gender line as a classifying device. Certainly if maleness is to serve as a proxy for drinking and driving, a correlation of 2% must be considered an unduly tenuous 'fit.'"[180] The Court also found fault with the statistics themselves, noting "methodological problems" and that the studies "make no effort to relate their findings to age-sex differentials as involved here."[181]

After rejecting the specific statistical evidence brought, the Court issued a broader statement distancing itself from substantiating similar policies with statistical evidence:

> It is unrealistic to expect either members of the judiciary or state officials to be well versed in the rigors of experimental or statistical technique. But this merely illustrates that proving broad sociological propositions by statistics is a dubious business, and one that inevitably is in tension with the normative philosophy that underlies the Equal Protection Clause.[182]

In a footnote, the Court noted that "if statistics were to govern the permissibility of state alcohol regulation without regard to the Equal protection Clause as a limiting principle, it might follow that States could freely favor" any number of groups which studies had shown drank at different rates, such as between religious denominations.[183]

178. 429 U.S. 190 (1976).

179. *Id.* at 201.

180. *Id.* at 201–02. The Court here uses the term correlation in its colloquial sense of "association," rather than specifying that a correlation coefficient between two variables is 0.02.

181. *Id.* at 202–03.

182. *Id.* at 204.

183. *Id.* at 208 n.22.

Courts have found *Craig* persuasive. Consider two examples with similar skepticism toward statistical evidence. First, in *People v. Chairez* the Illinois Supreme Court applied a heightened form of intermediate scrutiny to consider whether a state statute prohibiting possession of a firearm within 1,000 feet of a park violated the Second Amendment.[184] Under this analysis, the government needed to show a "close fit" between the means (the firearm restriction) and the ends (protecting children from gun violence).[185] In practice, the court looked in vain for statistical evidence of a correlation between means and end, finding "no evidentiary support" or "direct correlation" between the restriction and its purported purposes.[186] Finding the state's evidence "devoid of any useful statistics or empirically supported conclusions,"[187] the lack of statistical evidence led to the same outcome as in *Craig*, and the government's policy was found unconstitutional.

Second, in *Lamprecht v. FCC*, the D.C. Circuit considered whether the FCC could constitutionally award "extra credit" towards female applicants for radio station permits.[188] The purpose behind the policy was to promote underrepresented programming, with the government introducing evidence that "women who own radio or television stations are likelier than white men to broadcast these distinct types of programming."[189] The court examined this evidence and found the data "fail[ed] to establish any statistically meaningful link between ownership by women and programming of any particular kind."[190] Then, as in *Craig*, the court used this to make a broader statement regarding statistical evidence in equal protection analysis:

> The study, in short, highlights the hazards associated with government endeavors like this one . . . . When the government treats people differently because of their sex, equal-protection principles at the very least require that

184. People v. Chairez, 104 N.E.3d 1158, 1163 (Ill. 2018).
185. *Id.* at 1174.
186. *Id.* at 1176.
187. *Id.* at 1175.
188. 958 F.2d 382, 383 (D.C. Cir. 1992).
189. *Id.* at 395.
190. *Id.* at 398.

> there be a meaningful factual predicate supporting a link between the government's means and its ends.[191]

In sum, courts have a long and varied history with statistical evidence, particularly as it bears on causation and how the evidence ties into established frameworks for resolving constitutional questions. When one court might call for strong causal statistical evidence (e.g., *Entertainment Merchants Ass'n* and *Craig*), another might apply that precedent to anecdotal evidence alone (e.g., *St. John's Church* and *Melchert-Dinkel*), still another might call for statistics without specificity (e.g., *Chairez* and *Lamprecht*), and yet another might dismiss the need for statistics altogether (e.g., *Mun Choi*). From the disparate ways in which courts have treated correlation and causation, to the way courts have found or dismissed causal evidence, what is lacking from each of the above is a unified, consistent framework for analyzing or incorporating evidence of causality into constitutional reasoning. The next Section offers a possibility and then re-examines the cases above.

IV. PAIRING CAUSAL ANALYSIS WITH CONSTITUTIONAL SCRUTINY: A FRAMEWORK

In this Section we apply the principle of a causality spectrum from Section II to the use of causal evidence in constitutional scrutiny analysis from Section III. We argue that, in general, courts should match the strength of the causal evidence to the strictness of their scrutiny analysis. That is, when a government policy requires a high level of scrutiny to be accepted as constitutional, and when statistical evidence is used to tie that policy to the government interest, that causal evidence ought to be strong: towards the right on the causality spectrum. If the government policy must only satisfy a low level of scrutiny, statistical evidence may only need to show a correlation or lie towards the left on the causality spectrum. In this sense, the framework here builds off the holding in *Entertainment Merchants Ass'n* that specifically paired strict scrutiny to a high level of causal evidence, while extending it to multiple levels of scrutiny and multiple "levels" of the

191. *Id.*

complexity of the relationship between government means and ends. We then offer a qualitative, Socratic approach to analyzing the strength of causal statistical evidence and apply it across the scrutiny cases discussed in Section III.

A. Pairing Scrutiny with Statistics, in General

Over the decades in which constitutional scrutiny analysis developed, courts developed rough tiers of scrutiny which correlated with the importance of the right or the fundamental nature of the trait under consideration. The more important the right, or the more fundamental the trait, the higher the level of scrutiny to be applied. Then, at least for Second Amendment concerns, even within that analysis the closer the restriction is to the "core" of the right, again the higher the level of scrutiny that is applied.[192] In like manner, the higher the level of scrutiny, the higher the general standard should be for the causal evidence supporting the policy. Table 2, below, summarizes this framework.

In this framework, a natural starting point for analysis is to examine the relationship between government means and ends. Drawing on a simplified version of complexity theory, we characterize that relationship as either simple or complicated.[193] A simple setting lies close to the "billiard-ball" causality discussed previously. In these situations, there is little temporal or physical distance between government means and ends, which means little worry about the presence of confounding variables or feedback loops that render the relationship between cause and effect difficult to analyze, and few inferential steps are needed to suggest that the government means will affect the ends in the stated ways.[194] On the other hand, in a complicated setting, which is more

192. *See, e.g.*, Ezell v. City of Chi., 651 F.3d 684, 708 (7th Cir. 2011) ("[L]aws restricting activity lying closer to the margins of the Second Amendment right, laws that merely regulate rather than restrict, and modest burdens on the right may be more easily justified. How much more easily depends on the relative severity of the burden and its proximity to the core of the right.").

193. In complexity theory, common levels of complexity might be "simple," "complicated," and "complex." In that categorization, "complex" systems are ones which exhibit "nontrivial emergent and self-organizing behaviors." MELANIE MITCHELL, COMPLEXITY: A GUIDED TOUR 13 (2009).

194. Examples of these might be *State v. Melchert-Dinkel*, 844 N.W.2d 13, 23 (Minn. 2014), and *Saint John's Church in the Wilderness v. Scott*, 296 P.3d 273, 283–

typical in social policy, many inferential steps are needed, as many potentially confounding causes exist, and those causes might be embedded in feedback loops as policy interventions cause behavioral changes, which then affect policy application, and so on.[195] Generally speaking, in simple situations little causal evidence might be needed, as, like when striking a billiard ball, effect may be easily attributed to cause. For complicated situations, more evidence should be sought.

Consider now how the interactions between levels of scrutiny and the complexity of the policy environment alter the need for causal evidence. First, when the case involves rational basis review or minimal scrutiny, the causal evidence linking the government action to the government interest should likewise be given minimal scrutiny.[196] Although from a policy perspective one might question government policy created with no basis to link cause and effect, legally, causal evidence might not be needed at all.[197] To require otherwise would fundamentally change the scope of rational basis review, which only requires a "rational" connection between government ends and means. In this way, the framework continues the long trend of judicial deference to legislative actions, at least in certain areas of review. By its nature, this holds true in situations in which there is a simple or complex relationship between those ends and means.

For a somewhat higher level of scrutiny, such as the various forms of intermediate scrutiny, courts should increase their scrutiny of causality to an intermediate level. In a simple policy environment, evidence should be brought that at least supports a plausible inference of causality, even if that evidence is not quantitative. When a more complex relationship between government means and ends exists, a court should seriously question government action based on no statistical

84 (Colo. App. 2012), discussed in *supra* notes 171 through 173. This might explain the courts' lack of concern for statistical evidence to establish causal relationships between government means and ends.

195. These situations might be characterized as having "complex and multiple channels of causality." Banerjee & Duflo, *supra* note 75, at 152.

196. We have not discussed rational basis cases in detailed examples in the prior sections, as the need for so little evidence in general for rational basis review means there is little need to press for causal connections to justify government policy.

197. If stronger evidence was needed for garden-variety economic policy, the effect on government decision making would be paralyzing.

evidence at all. In such cases, the government should bring potent justifications for how the ends will affect the means without scientific evidence causally connecting the two. If statistical evidence *is* brought, a court would look for evidence not at the far left on the causality spectrum where the assumptions behind establishing causality are very strong and one would justifiably wonder if there *is* a cause-and-effect relationship between means and end. The evidence should be at least *suggestive* of a causal relationship, even though the assumptions on which causality is based may be strong.

For cases involving strict scrutiny analysis, in which a fundamental right or attribute is at issue, the Supreme Court's standard in *Entertainment Merchants Ass'n* is instructive.[198] In our framework, this means first that courts considering such cases *without* statistical evidence should specifically recognize and address this fact. If the causal mechanisms involved in the case are so clear that such evidence is not needed, courts should make this apparent. This might be the case for simple policy environments. Next, if statistical evidence *is* brought in a complicated environment, causal connections should be based on relatively weak assumptions, likely towards the right of the causality spectrum. The evidence should be *conclusive* or *strongly suggestive* of a causal relationship, based on reasonable assumptions. In line with this, if a case involves such complexity that causal mechanisms are unclear and causal evidence is *impossible* to bring, courts should take a particularly hard look at the potential regulation, as it is unlikely that the government regulation can be shown to be related to the policy in appropriate causal ways.

In sum, while each case and the evidence upon which it is based is inherently an individual matter, existing caselaw and prudential reasoning suggest a general match between the level of scrutiny and the level of its underlying causal analysis. Rational-basis cases have generally not required evidence at all, and so making causal requirements would dramatically change their review. Cases with a higher, intermediate level of scrutiny should bring evidence at least suggestive of causality, even though assumptions may be strong, while strict scrutiny cases should rely on evidence that is strongly suggestive of causality under reasonable assumptions.

198. *See supra* note 157 and accompanying text.

Table 2: Guidelines for Strength of Causal Evidence by Level of Scrutiny

Policy Environment/ Relationship Between Government Means and Ends	*Rational Basis Review*	*Intermediate Scrutiny*	*Strict Scrutiny*
Complicated Many inferential steps needed, feedback loops and confounding causes require assumptions	**Minimal**	**Medium** Evidence is suggestive of a causal relationship although assumptions may be strong	**High** Under reasonable assumptions, evidence is conclusive or strongly suggestive of a causal relationship
Simple Few inferential steps needed, no feedback loops or confounding causes	**Minimal**	**Low** Evidence, even if qualitative, supports a plausible inference of causality	**Low** Evidence, even if qualitative, supports a plausible inference of causality

Level of Scrutiny

To assist courts in examining statistical-based evidence of causality in this framework, the following subsection outlines a series of Socratic-style questions that can be used to address fundamental statistical issues in causality in a non-technical manner. These suggested dialogues will help show when the evidence supports, suggests, or provides conclusive evidence of causal relationships as in Table 2 above. Recall, causality itself is inherently *qualitative*, which means that probing causal claims can proceed in a relatively thorough manner without legal practitioners first obtaining PhDs.[199] That is not to say that generating the evidence of these claims is trivial, but that *understanding the reasoning* behind causal mechanisms and avoiding common pitfalls in causality is inherently quantitative and within the scope of non-specialized courts.[200]

B. A Socratic Narrative Framework for Causal Claims

In this subsection, we propose a Socratic-style narrative for courts and litigators examining causal claims. These questions are

199. SHADISH, COOK, & CAMPBELL, *supra* note 28, at 6.

200. If causal mechanisms and evidence cannot be explained in terms courts and their learned assistants can interpret, then those causal mechanisms should be met with skepticism. While the statistical methodologies themselves may be quite complex, the causal assumptions underpinning those relationships should be explainable.

meant to serve as an approach, or overlay, to the more technical aspects involved in establishing causality through statistics. This narrative could be used in actual oral arguments but could also serve as a framing device for briefs, deposition outlines, conversations with experts, and so on. It probes many of the common problems with establishing causal evidence. In the subsequent subsection, we then apply elements of this narrative to the constitutional scrutiny cases discussed earlier.[201]

1. Do we need causal evidence in this context?

Perhaps the first question to ask when confronting evidence in the scrutiny context is whether causal claims are needed, such as whether the case belongs in the first column of Table 2 under rational basis review. If the level of legal scrutiny applied is so minimal that almost any connection between government means and ends would be sufficient, perhaps no causal evidence is needed at all. If that is the case, the court evaluating such a claim should consider being upfront with the lack of need for such evidence. Highlighting the lack of causal connections between government means and ends may not change the outcome of a rational basis analysis, but it can help make clear to observers the broad basis on which rational basis review allows policy to proceed.

2. Should the causal evidence be based on statistical studies?

If the answer to the first question is yes—that some evidence of causality is needed—then the next question addresses whether causal mechanisms are so clear that no scientific or statistical evidence need be relied on to establish causality. In simple policy frameworks (the bottom row in Table 2), such as when dealing with "billiard-ball"-style cause and effect, this may be true.[202] In more complicated cases, as is often the case with government policies, courts should appropriately ask if studies exist or could be performed which might establish causal links. If the answer is no, courts should ask *why* no such studies exist or could be performed. The response to that question may reveal assumptions behind the policy relevant to the court's analysis regardless

201. *See supra* Section III.C.

202. *See supra* Section II.B (discussing strength of causal associations).

of the ultimate answer to the question of whether statistics are needed.[203]

If the answers to questions (1) and (2) are both yes, it means the court is likely in a situation covered in the "medium" or "high" cells of Table 2. Further questioning—such as given below—should probe the statistical evidence brought to bear under those standards. These questions do not jump immediately to tests of statistical significance before probing the nature of the causal relationships embedded in the data. As discussed above, statistically significant results may merely convey the veneer of scientific proof while masking strong assumptions or biases in the data. [204]

3. Does the statistical evidence claim to establish causality?

As an initial probe of the qualitative evidence of causality in statistical evidence, it is generally worthwhile to examine what the research itself claims about causality. Often, research will be up-front on whether causal mechanisms are being claimed, and on what assumptions.[205] If the research is silent on these points, then it is unlikely to have a convincing causal interpretation, and courts may be hesitant to rely on that research without first making clear that the evidence does not claim to establish causal effects. Should the court wish to proceed further, question (4) begins to probe causality directly.

203. If causal evidence is needed yet not forthcoming, then courts should be clear that government action has no statistical evidentiary basis.

204. The "significance" of significant p-values is discussed in *infra* Section IV.B.6.

205. *E.g.*, Ian Ayres, Jeff Lingwall, & Sonia Steinway, *Skeletons in the Database: An Early Analysis of the CFPB's Consumer Complaints*, 19 FORDHAM J. CORP. & FIN. L. 343, 364 (2014) ("Note that our reduced-form model cannot prove causality: for example, if ZIP codes with higher proportions of senior citizens have more complaints per mortgage, we cannot determine if mortgage companies are treating older borrowers worse than younger ones, or if senior citizens are just more likely to complain to the CFPB. However, even in the absence of causal evidence, the results at least suggest that mortgage lenders might need to show increased care toward certain populations.").

4. In the study at hand, how can causality be established?

If the research does not aim to establish causality, or if it is unclear on what assumptions causal claims based on the data may be made, it is then worthwhile to probe what assumptions would be needed to make causal claims. Here, the spectrum of causality assumptions detailed in Section II may be useful. This spectrum details types of studies and roughly identifies whether the assumptions on which causal claims can be made in that type of research design are weak, strong, or heroic.[206] If strong or heroic assumptions must be made, then a causal interpretation of the data should be treated with caution.

Once the causality assumptions behind a particular type of study are situated relative to general social-scientific standards, it is fundamental to question the details on how causality is understood in the particular evidence at hand. A series of relevant questions would first ask for an explanation, in simple terms, of which causal relationships are being explored.[207] Removed of technical jargon, this could be a statement such as "the study aims to find the causal effect that restricting handguns near public schools will have on school violence." In the constitutional scrutiny context, this statement should connect government means and ends.

Next, unless the study involves randomized experimentation, a critical question to ask is what theoretical experiment *would* be able to assess the causal relationship of interest.[208] For instance, even if not feasible, one could imagine taking a sample of schools across the United States and randomly allocating them to be included in gun-free zones. Once these zones were created, school violence would be measured over a time period sufficient to capture likely effects of the law.[209] Following this thought experiment, one would then ask how the research at hand proxies the theoretical experiment. In what ways is it different? In what ways is it similar? Crucially, what kind of

206. *See supra* note 116 and accompanying text.

207. *See* ANGRIST & PISCHKE, *supra* note 31.

208. *Id.*

209. In many, if not most, situations involving scrutiny analysis this type of experiment will be impractical because of cost or ethical concerns, but it still serves as a correct starting point for analysis.

assumptions must be met to derive the same kind of conclusions one would obtain from the theoretical experiment?

Key parts of a response will identify what in the present research serves in place of randomized allocation to a treatment group, and what in the present research serves in place of allocation to a control, or counterfactual group. For example, when considering how a DID study might proxy a randomized experiment for the school violence example, research might focus on two groups of similar schools, across county lines from each other, in which one county adopted a gun-free zone policy and the other did not. We would question how the policy was adopted, to see if we can claim it was independent of violence levels at the schools, in the same manner as randomization.[210] We would question whether the schools across county lines properly serve as counterfactuals, or if they are fundamentally different in ways we would expect to confound the comparison of school violence rates over time. We would then be clear on the assumption about common trends in school violence in the absence of the policy.[211]

5. What are examples of how the assumptions behind causality may be violated?

As an additional check on the prior answer, it is worth questioning specifically how the assumptions established in the prior question may be violated, and how likely that is to be the case. Framing the questioning of assumptions in this manner may elicit different nuance that sheds light on potential problems with the research. In our school example, perhaps the research notes that control-group schools

210. Often this concept is expressed through the term "exogenous." An "exogenous" variable is independent of other variables in the model. *See* Will Kenton, *Endogenous Variable*, INVESTOPEDIA https://www.investopedia.com/terms/e/endogenous-variable.asp (last updated Oct. 30, 2020).

211. Part of this discussion often includes which observable variables were controlled for in the statistical analysis. If a potentially confounding variable is observable, then controlling for that variable allows the researcher to avoid assuming that the variable has no impact on the analysis. *See supra* note 61 and accompanying text. Properly controlling for potentially confounding variables thus weakens the assumptions required to make causal claims and strengthens the analysis. One examining research with these variables included should ask which variables were controlled for, and what effect they had on the causal relationship of interest.

established various other anti-violence policies, so that a proper counterfactual trend becomes hard to establish. If the ways that assumptions may be violated appear common and difficult to control for, then the research requires strong assumptions to make causal claims.

6. After the above, do statistically significant results exist?

Only after probing the research as described above should the question of statistical significance be addressed.[212] Statistical significance provides evidence on whether the observable relationship is due to random chance. It is typically denoted using "p-values," which denote the probability of observing a result as extreme, or more extreme, than an observed value if no real relationship existed in the data.[213] Often this part of the analysis is examined without the benefit of questioning the validity of the study design and assumptions behind the significance itself. Without this backdrop, claims of statistical significance may be misleading.[214] For example, with a confounding variable not accounted for in the study design, a result may be fundamentally biased yet still reflect a significant p-value. This is because a p-value only reflects statistical results, not the validity of the study design itself.

For instance, one might find a statistically significant p-value when comparing the two groups of schools in our example, while ignoring that treatment group schools were shut down due to a pandemic,

212. On the more technical side of these questions, a proper query would be to ask what mode of statistical inference will be used. *See* ANGRIST & PISCHKE, *supra* note 31.

213. For example, in regression problems the typical null hypothesis is that there is no linear relationship between variables. A low p-value indicates that it is unlikely the observed relationship would be found if the null hypothesis were true. The typical threshold for calling a relationship statistically significant is a p-value of less than 0.05, which would require a probability of less than one in twenty that the observed results would be found, given that no relationship actually exists. P-values between 0.05 and 0.10 are often called "marginally" significant, depending on the circumstances. P-values substantially higher than those values are typically deemed not-significant, meaning the results could quite likely have occurred by random chance rather than because a null hypothesis is false.

214. Often a significant p-value conveys the veneer of scientific certainty, which would then mask underlying issues in the analysis. In a mis-specified model, a significant p-value is thus worse than a finding of no statistical significance.

and thus violence rates declined to zero. Calculations of statistical significance are independent of the assumptions underlying causal models—therefore, the prior line of qualitative questions is essential to explore before examining quantitative significance levels. As with the early development of correlation versus causal analysis, in which correlation offered clear mathematical precision compared to the complex qualitative study of causality, courts should not rely on quantitative levels of statistical significance or correlation as a crutch to ignore questioning the qualitative assumptions behind those numbers.[215] The essence of causal reasoning is careful thinking and persuasive storytelling about hypothetical situations, the precise craft which courts practice every day. Manipulating data and calculating statistical significance may remain the domain of specialized experts, but understanding and communicating the *story* behind why those statistics are valid is a skill we believe courts can, and should, develop.

7. If a causal relationship is established and significant, is it generalizable?

This final question applies even to the results from randomized experiments. In any study which purports to influence policy, the fundamental question of generalizability remains. Even if a causal effect is likely established, if it is not generalizable outside the context of the study to the government policy at issue, its effect should be treated with caution.[216] If the study was based on, for example, widespread sampling across relevant populations, the study will be more generalizable than without. If applying the study to the situation at hand requires extrapolation or interpolation, again caution should be exercised.[217] Some general principles to consider are whether the study has surface similarity to the target population, whether important factors in the

215. If the study design has been properly engaged to establish the assumptions behind causal claims, then additional assumptions beyond the scope of this paper could be examined to validate the claim of statistical significance itself. For example, whether standard errors have been calculated properly in weighted data or data with non-constant variance may be addressed.

216. SHADISH, COOK, & CAMPBELL, *supra* note 28, at 21–22.

217. *Id.*

research limit generalizability, and whether the study includes any factors essential to cause and effect in the target population.[218]

C. Applying the Framework to Scrutiny Cases

In this subsection, we illustrate how an expanded causal narrative may be applied to existing cases. In these examples, the analysis is limited to the courts' narrative and discussion of causal evidence, but even with those limitations, it is possible to see how an expanded causal narrative as we suggest in this Article may lead to more nuanced outcomes.

First, consider again *Brown v. Entertainment Merchants Ass'n*.[219] There, the Supreme Court expressly called for causal evidence to apply strict scrutiny analysis. This matches the suggestions in our framework for situations in which there is a complicated relationship between government means and ends, as there likely is when trying to match videogame sales to real-life violence. Specifically, our framework suggests that the evidence should be conclusive or strongly suggestive of a causal relationship under reasonable assumptions. The State argued that it need not produce such evidence, as a legislative finding would arguably be sufficient.[220]

Rather than punting on the idea of causality, the State might have situated what evidence it *could* bring along a causal spectrum, making it clear under what assumptions, if any, that evidence could be construed as causal. Then, rather than treating causality lightly, a narrative about the strength and assumptions of that evidence could be engaged. Such discussion could lend precision to the strength of the causal link required by the Court in this context. This holds even if ultimately the last point on our causal narrative—generalizability—would appear to still be an issue because the studies cited applied to "feeling more aggressive" rather than committing violent actions.[221] As it requires a strong assumption to tie a causal effect between aggressive "feeling" and performing the kind of violent acts at issue in the California law, a causal interpretation was likely unwarranted.

218. *Id.* at 24–25.

219. 564 U.S. 785 (2011).

220. *Id.* at 790.

221. *Id.* at 800.

In *Craig v. Boren*, the intermediate-scrutiny case regarding male versus female drinking legislation, analysis along our lines would again have proceeded differently than in the opinion.[222] First, we would suggest that it *is* realistic to expect members of the judiciary to be well-versed in at least the quantitative aspects of causal analysis, as such analysis is well within the grasp of both the judiciary and government officials considering policy. While "proving broad sociological propositions by statistics"[223] *alone* is inherently troubling, using statistical evidence to support or attack legislative actions should certainly be within the capacity of members of the judiciary.

Leaving *Craig*'s broad dicta about the scientific capacity of attorneys aside, the narrative in the case could also proceed differently. Our framework would suggest the case was presented in a complicated situation, with many inferential steps needed between age limits and DUI laws. In a complicated situation under intermediate scrutiny, the evidence should be at least suggestive of a causal relationship under even potentially strong assumptions. The Court rejected the State's statistical evidence on gender-based DUI differentials as a poor "proxy" for drinking and driving.[224] In our framework, this type of evidence would fall on the far left of the causality spectrum, and so would be subject to attack not only because of its "fit" as a proxy for the evidence at hand (a problem of generalizability), but because the observational study considered would need *heroic* assumptions to show a causal effect. For instance, claiming causality would require assuming that DUI rates were reflective of actual rates of drinking and driving, that a policy changing age limits would then affect those underlying rates of drinking and driving, and so on. Because of the heroic nature of these assumptions, there was little guarantee the government's policy would actually affect the means. Again, precision in matching causal language to scrutiny analysis makes clear *why* the evidence failed: government means could not match government ends on the causal evidence presented. The evidence did not reasonably suggest a causal relationship without making heroic assumptions.

222. 429 U.S. 190 (1976)

223. *Id.* at 204.

224. *Id.* at 201–02.

In both *Entertainment Merchants Ass'n* and *Craig*, a particularly relevant series of questions would engage what kind of experimental evidence *would* provide causal evidence in those contexts. In *Entertainment Merchants Ass'n*, this would require something such as randomizing children into two groups, one of which would be required to play violent video games, and the other would not. These two groups would then be observed for violent behavior. In *Craig*, it would imply a study that separated teenagers into two groups, applying different drinking ages to each group, and observing the outcomes. In both cases, those type of experiments are unlikely to occur. Yet again in both cases, the thought experiment sheds light on why the statistical evidence was insufficient. In *Entertainment Merchants Ass'n*, what was observed was not generalizable to the issues considered in the legislation or the hypothetical experiment, as it requires heroic assumptions to believe an effect showing "louder noises" has bearing on, for example, killing another human being.[225] In *Craig*, the available observational study (at the far left of the causality spectrum) about DUI arrest ratios had nothing at all to do with altered drinking age standards, and so provided poor evidence of what result altered standards would bring.[226]

The other scrutiny cases considered above fall along similar lines. In *Chairez*, the Illinois Supreme Court applied intermediate scrutiny and found no "direct correlation" between restricting firearms near parks and protecting children from gun violence.[227] The court's main complaint was the lack of *any* empirical evidence.[228] While the Court's focus was on the lack of empirical evidence, a narrative focused on causality would first examine the need for causal evidence. In the case of intermediate scrutiny, our framework recommends evidence that is suggestive of a causal relationship, even though assumptions underlying that evidence may be strong. If billiard-ball style causality existed between the government means and end, then no statistical evidence should be required, as the qualitative reasoning might be convincing. Here, relating gun restrictions to lower gun violence does not

225. *Ent. Merchs. Ass'n*, 64 U.S. at 799.

226. *Craig*, 429 U.S. at 201–02.

227. People v. Chairez, 104 N.E.3d 1158, 1176 (Ill. 2018).

228. The court complained the state's arguments were "devoid of any useful statistics or empirically supported conclusions." *Id.* at 1175.

necessarily involve the simple issues involved in the physics problem of striking billiard balls. As such, the court was then correct to look for at least *some* suggestive empirical evidence. Finding none, it correctly found intermediate scrutiny unsatisfied.

In *Lamprecht*, the D.C. Circuit applied intermediate scrutiny analysis to a government policy favoring female applicants for radio station permits.[229] In our framework an intermediate level of scrutiny pairs with an intermediate examination of the causal evidence brought to bear, meaning the evidence suggests a causal relationship, even if based on strong assumptions. The court noted the lack of a "statistically meaningful" link, which would imply that the evidence was *not* suggestive of a causal relationship or that establishing one would require making heroic assumptions. To be clearer in how it evaluated the evidence, or in how the statistical-based relationship between government means and ends was considered, the court could have specified whether causal evidence was warranted, and if not, why. Then if causal evidence were warranted, positioning the court's analysis in terms of a causal narrative would make clear how the existing evidence failed and how it might be improved.

A causal narrative in this context might proceed as follows. After addressing whether causal evidence was required, the court would ask what the study at hand claimed regarding causality. The court's evidence was observational, denoting radio station ownership by sex and noting what kind of programming each station engaged in.[230] It is unlikely that this kind of evidence made causal claims of any kind, as the assumptions required to do so would be very strong.[231] Next, the

229. *See supra* note 188 and accompanying text.

230. Lamprecht v. FCC, 958 F.2d 382, 397–98 (D.C. Cir. 1992).

231. Observational studies may provide evidence of causality but require strong assumptions to do so. As one textbook puts it:

> If great care is taken to control for the most likely lurking variables (and to avoid other pitfalls which we will discuss presently), and if common sense indicates that there is good reason for one variable to cause changes in the other, then researchers may assert that an observational study provides good evidence of causation.

Causation and Observational Studies, UNIV. OF FLA. BIOSTATISTICS: OPEN LEARNING TEXTBOOK, https://bolt.mph.ufl.edu/6050-6052/unit-2/causation-and-observational-

court might ask what kind of experiment could theoretically be constructed to establish this relationship. Taking radio station applicants, randomly dividing them into two groups, one of which would receive extra credit based on sex, and the other of which would not, allocating permits, and then observing what kind of programming was used, would be an impractical and unethical experiment to run. The study at hand differed from the ideal in fundamental ways. As an observational study, there was no attempt at randomization, which means making either strong or heroic assumptions about the allocation of the many variables related to minority programming. In our framework, litigators and courts would then engage in discussion about the strength of those assumptions, such as whether potentially confounding variables were a concern.

Finally, consider the motivating example of strict scrutiny analysis from Missouri in *State* ex rel. *Schmitt v. Mun Choi*. Under Missouri law, restrictions on firearms on college campuses are analyzed under a strict scrutiny framework. The relationship of gun restrictions and campus violence certainly qualifies as complicated: feedback loops may exist between policies and reactions to those policies, studies on campus violence may suffer from numerous unobservable confounding variables, and causal reasoning from those studies would necessitate making many assumptions. In our framework, we would thus look for a high level of causal evidence, meaning the evidence is conclusive or strongly suggestive of a causal relationship under reasonable assumptions.

The trial court weighed both qualitative testimony on the potential effects of the regulation and a handful of observational studies showing a somewhat weak relationship between such restrictions and campus violence. Expert testimony noted that "if a number of different tests all point in the same direction . . . this can imply causation, even if the tests individually do not rise to the level of statistical

studies/ (last updated Nov. 1, 2021). For example, one could assume that no confounding variables related to minority programming exist. This could be the case if radio station owners who differed on the basis of sex differed in no other way related to their choice of programming. This assumption would be strong. For example, other characteristics such as education, geographical background, or the radio programming one experienced as a child might presumably be related to programming choices as an adult.

significance."[232] The trial court looked at that general trend, and noted that the results were "inconclusive," but also that "[e]very test and model . . . points in the same direction: Violent crime . . . always increased . . . after colleges started allowing firearms on campus."[233] At any rate, the statistics did not "provide . . . a basis for second-guessing law enforcement"[234] and, in general, "a 'statistically significant' relationship between the regulation and the asserted interest is wholly unnecessary."[235] On appeal, the court went further, accepting the trial court's reasoning and noting that "the statisticians provided credible, competent evidence as to the compelling interest of promoting safety and reducing crimes."[236]

At both trial and on appeal, causal reasoning played an extremely limited role, essentially captured by the statement that studies may *imply* causality if generally agreeing with each other. This statement leaves unknown what assumptions the "causality in the aggregate" implication would rest on and how strong those assumptions might be, which are likely quite strong. Combining multiple studies for enhanced precision is a common statistical technique but combining studies can never fix methodological issues related to causality in each individual piece of research. In other words, if what can be said about causality in any individual study is limited, a combination of those studies might remain limited as well. This is because as studies are combined, greater precision, such as more precise correlations, may be obtained, but the assumptions and causal reasoning underlying those studies do not change in aggregation.

The widest-ranging study considered was based on DID methodology comparing violent crime rates in states that passed right to carry laws between 1977 and 2014. A causal interpretation of this research requires the assumption that some states serve as valid controls for other states and that national-level results are generalizable to the

232. State *ex rel.* Schmitt v. Mun Choi, Nos. 16BA-CV03144, 16BA-CV0758, 2020 WL 5093608, at *8 (Mo. Cir. Ct. 2020).

233. *Id.* at *6–7.

234. *Id.* at *5.

235. *Id.* at *14.

236. State *ex rel.* Schmitt v. Mun Choi, 627 S.W.3d 1, 18 (Mo. Ct. App. 2021).

specific circumstances present on university campuses.[237] Those are not necessarily weak assumptions to make. This, combined with the even stronger assumptions required to establish causality in the other studies considered, suggest that in our framework it is questionable whether the causal evidence is either conclusive or strongly suggestive of a causal relationship as should be expected in a strict scrutiny situation. At the least, examining the combination of qualitative and quantitative evidence in a more specific causal narrative would make clear what role causality played in the courts' reasoning, how the evidence matched the courts' expectations of causality, and how the courts viewed the relationship between government ends and means.

V. CONCLUSION

It may be argued, with some truth, that asking courts and litigants to address causal questions in statistical evidence is simply too much. Overburdened courts and harried litigators have little time to consider assumptions behind statistical evidence or proceed carefully down a series of arguments that probe causal assumptions. Presenting

237. John J. Donohue, Abhay Aneja, & Kyle D. Weber, *Right-to-Carry Laws and Violent Crime: A Comprehensive Assessment Using Panel Data and a State-Level Synthetic Control Analysis*, 16 J. EMPIRICAL LEGAL STUD. 198 (2019). The DID model employed showed that states that never adopted right to carry laws experienced a 42% decline in violent crime rates, while states that adopted those laws between 1977 and 2014 experienced a 4% reduction in violent crime. Finally, states that adopted right to carry laws before 1977 showed a 10% reduction in violent crime. Based on these statistics, Donohue and his collogues concluded that the average post-passage increase in violent crime was 20%. *Id.* at 213–14. Donohue's team noted that:

> Of course, it does not prove that RTC laws increase crime simply because RTC states experience a worse postpassage crime pattern . . . . If police and prisons were more effective in stopping crime, the 'no-controls' model might show that the crime experience in RTC states was worse than in other states even if this were not a true causal result of the adoption of RTC laws. As it turns out, though, RTC states not only experienced higher rates of violent crime but they also had larger increases in incarceration and police than other states.

Id. at 214.

statistical evidence and expert testimony *at all* shows a level of rigor lacking in many legal situations, and as one begins to take causal evidence seriously, one confronts literature skeptical of causal endeavors outside the laboratory at all.

While these are strong arguments in favor of the status quo, we respectfully disagree. Causal thinking lies at the heart of substantive law, as causation underlies liability of any sort. Because of this, legal practitioners are better prepared than many professionals to think in causal terms and are familiar with nuances in causal reasoning that might escape others, such as the difference between legal and proximate causation. "[C]ausality is not *mystical* or *metaphysical*. It can be understood . . . ."[238] As courts evaluate evidence that underlies potential causal relationships, strong causal thinking need not be abandoned. When it is, bias in evidence may flow through to bias in decision making. This is particularly worrisome in the context of constitutional scrutiny cases which seek to balance strong government interests against intrusion upon constitutional rights. Poor causal analysis leads to imprecision and vagueness in relating evidence to elements of scrutiny analysis and hides potential mismatches between that evidence and the legal reasoning flowing from it.

This Article proposes a narrative framework for evaluating evidence in the constitutional scrutiny context, based on the qualitative reasoning underlying all causal analysis. The framework matches levels of causal reasoning to levels of scrutiny, proposes lines of questioning that examine the assumptions upon which potentially causal evidence is based, and clarifies the evidentiary relationship and assumptions linking government ends and means. When considering foundational constitutional scrutiny cases considering this framework, lines of both legal and statistical reasoning that would match evidence to legal analysis become clearer. As legal practitioners increasingly apply principles of causal reasoning to constitutional scrutiny, the result will be increased precision for courts, clearer expectations for litigators, and better policy for the public.

238. PEARL, *supra* note 2, at 427.