ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

BALANCING VERSUS STRUCTURED DECISION PROCEDURES: ANTITRUST, TITLE VII DISPARATE IMPACT, AND CONSTITUTIONAL LAW STRICT SCRUTINY

Louis Kaplow

Forthcoming in *University of Pennsylvania Law Review*, Vol. 167 (2019)

Discussion Paper No. 1000

04/2019

Harvard Law School
Cambridge, MA 02138

BALANCING VERSUS STRUCTURED DECISION PROCEDURES: ANTITRUST, TITLE VII DISPARATE IMPACT, AND CONSTITUTIONAL LAW STRICT SCRUTINY

LOUIS KAPLOW[*]

Abstract

Important doctrines in diverse areas of law employ structured decision procedures requiring, in rough terms, that the plaintiff first make some demonstration of harm; if but only if that is done, the defendant must make some showing of benefit; and if but only if that occurs, balancing is performed. In-depth analysis of such protocols reveals them to be inferior to unconstrained balancing with respect to the quality of final decisions and the guidance they provide for the collection of information and, accordingly, the conduct of adjudication. This article applies this analysis to the rule of reason and merger regulation under antitrust law, Title VII disparate impact law, and the practices of strict scrutiny and proportionality analysis in constitutional law. Longstanding controversies are addressed and unappreciated deficiencies are discovered. In all three domains, existing law is cast in a substantially different light, both descriptively and normatively.

Forthcoming, University of Pennsylvania Law Review, vol. 167 (2019)

JEL Codes: J78, K00, K21, K31, K41, L41

Keywords: legal rules, adjudication, information, balancing, rule of reason, mergers, disparate impact, Title VII, strict scrutiny, proportionality analysis

[*] Harvard Law School and National Bureau of Economic Research. I am grateful to Erwin Chemerinsky, Steve Churchill, John Donohue, Richard Fallon, James Greiner, Vicki Jackson, Benjamin Sachs, Holger Spamann, Matthew Stephenson, Adrian Vermeule, and workshop participants at Berkeley, Chicago, Harvard, Northwestern, and Stanford for helpful discussions and comments; Tracy Cui, Carys Johnson, Kevin Keller, Carsten Koenig, Iacopo Lash, Andrea Lowe, Asher Lowenstein, Clarissa Lu, Junesoo Seong, and Aluma Zernik for research assistance; and Harvard Law School's John M. Olin Center for Law, Economics, and Business for financial support. Disclaimer: I occasionally consult on antitrust cases, and my spouse is in the legal department of a financial services firm.

TABLE OF CONTENTS

INTRODUCTION

Balancing is a familiar mode of decisionmaking in the law and beyond. When one consideration favors a particular decision (say, liability) and another opposes it, it seems to be the essence of reason that the superior decision reflects the balance of the competing forces, taking into account the weight of the evidence and the importance of each factor. Many legal rules, such as the negligence test for tort liability, operate in this fashion.

Sometimes, however, structured decision procedures are used instead for these types of decisions. As a benchmark for comparison with balancing, this article takes the following stylized version as a point of departure:

(1) The plaintiff must show that the harm of the defendant's act exceeds some threshold. If not, there is no liability. If so:

(2) The defendant must show that the benefit of its act exceeds some other threshold. If not, there is liability. If so:

(3) The harm and benefit are balanced, and there is liability if and only if the harm is greater.

Similar or related schemes are thought to characterize some existing legal decision procedures or have been proposed in antitrust law (rule of reason, mergers), discrimination law (Title VII disparate impact), and constitutional law (strict scrutiny, proportionality analysis). Yet structured decision procedures of this sort are neither used nor advocated in most other areas of law.

It is natural to inquire when and why legal rules should employ structured decision procedures in lieu of balancing. And, when they are used, it is necessary to set the two decision thresholds, for their height is critical to the procedure's bite and, in particular, its difference from unconstrained balancing. This article aims to answer these questions and others, with a particular emphasis on the descriptive and normative implications in the aforementioned legal domains.

Part I presents a general, abstract comparison of structured decision procedures to unconstrained balancing in order to frame the analysis of the substantive doctrines examined in the next three parts.[1] Part I begins by comparing the two approaches as final decision rules when all of the information that will be considered is before the decisionmaker. In many cases, including all those that reach step 3 under the structured protocol, the decisions will be the same. In important settings, however, the outcomes will differ. Moreover, in all such cases the outcome under the structured rule is necessarily inferior in the sense that such cases involve either the assignment of liability when the benefit exceeds the harm or a failure to assign liability when the harm exceeds the benefit. In addition, the purported virtue of structured rules in avoiding difficult balancing turns out to be misleading. They avoid balancing in many easy cases, but neither effort nor error is reduced. And they avoid balancing in some hard cases, but that is precisely when they stop short of the balancing performed in step 3 even though the resulting outcome from step 1 or step 2 may well be incorrect. Moreover, they sometimes require close comparisons with the thresholds even though balancing would have been easy. On reflection, it is remarkable that structured decision procedures are believed to prevail in

[1] This abbreviated presentation draws on the extensive analysis in Louis Kaplow, *On the Design of Legal Rules: Balancing Versus Structured Decision Procedures*, 132 HARV. L. REV. 992 (forthcoming 2019).

important areas of law and are advanced as replacements for balancing in some others without even having asked the basic question of how outcomes under the two methods differ.

Part I then compares the two approaches as guides to information gathering, which is sometimes advanced as a virtue of structured decision procedures because stopping early (at step 1) economizes on effort. This advantage proves to be largely illusory. Optimal, unconstrained information gathering, which is what one ideally would do under a balancing approach, involves a number of principles that are sharply violated by the structured protocol, stemming from the latter's sequential separation of the investigation of the harm and the benefit of a challenged practice. First, much evidence is expressly comparative; indeed, characterization evidence is relevant precisely to the extent that it bears differentially on competing understandings of the defendant's alleged act. Attempting to separate the two is artificial and fraught, somewhat like using scissors, disjoined, one blade at a time. Second, evidence often naturally clumps by source rather than by subject: internal documents, witnesses, and experts; not harm and benefit. Third, even if all evidence bore only on harm or only on benefit and naturally clustered in single-issue bundles, it is a priori unlikely that the optimal order of gathering and assessing evidence would be to do first all of one type (harm), followed by all of the other (benefit). Instead, it is (roughly) sensible to collect first, second, and so forth whatever bundle has the highest diagnosticity to cost ratio; at any given point, the most promising bundle may just as plausibly involve benefit as harm. Indeed, since there tend to be diminishing returns with respect to the exploration of each issue, it is unlikely that all of the most promising avenues would concern only one and all of the least promising only the other. A final subsection explains that neither of the two approaches actually governs the conduct of U.S. civil litigation, which has a structure of its own—a fact that casts a curious light on many doctrinal and policy discussions of structured decision procedures.

Parts II–IV, the core of this article, apply this general analysis of balancing versus structured decision procedures to three areas of law. The reader should note that the analysis is qualified in a number of ways: Constraints of space result in partial analysis, mainly to illuminate core ideas. Each area of law is different from the others and from the stylized three-step protocol that is analyzed in Part I. Moreover, doctrine in each area is to a degree murky, and the foregoing analysis sometimes calls into question whether common descriptions of the rules reflect actual practice. Rather than repeating these caveats throughout, I will simply proceed, often presenting points in blunt terms for purposes of brevity and clarity, at some expense to fidelity. In spite of these qualifications, however, the general framework and analysis powerfully illuminate and in important ways reshape our understanding of the law in each of the three domains.

Part II examines antitrust law. The Supreme Court's canonical statements of the Sherman Act's rule of reason, which span a century, present it as a pure balancing test: liability turns on whether the challenged practice overall suppresses rather than promotes competition. Setting to the side some categorical carve-outs—such as the per se illegality of price fixing—commentators and courts increasingly restate this rule as a structured protocol that resemblances the three-step stylization examined here. To that extent, the foregoing criticisms are apt.

For mergers, the structure is similar, although not often explicitly stated as such. Even though most nontrivial horizontal mergers a priori generate some upward pricing pressure, few mergers are challenged and some challenges fail in court, and this is so (according to conventional lore) without reaching the question of whether there are offsetting efficiencies (akin to step 2). This suggests a distinct step 1 with a high threshold. On those occasions in which that hurdle is overcome, government guidelines and practice then move to step 2 but in turn find that it usually fails, suggesting a significant threshold there as well—particularly since it is commonly accepted that many mergers are motivated by greater efficiency (which is the rationale for challenging so few). Once placed in this article's framework, merger challenges can better be understood, criticized, and improved. One suspects, however, that regarding both information collection and decisionmaking, government agencies' internal analysis may reflect a less constrained balancing process than official pronouncements suggest.

Part III turns to disparate impact cases under Title VII. Under prevailing Supreme Court precedent and the 1991 Amendments to the Civil Rights Act, they are governed by a structured decision procedure that has some resemblance to that examined here.[2] The first step of disparate impact inquiries, which focuses on the plaintiff's prima facie case (typically proved with statistics) indeed suffers greatly from the problem of attempting to separate harm from justification, in a way that few have recognized. The second step (in some respects collapsed with the third, as will be discussed), regarding the defendant's business justification, is understood to raise a serious question regarding the threshold: specifically whether it is essentially zero ("job related" is taken to mean that any connection whatsoever between the discriminatory employment practice and productivity is sufficient), massive ("business necessity" is the operative phrase, and "necessity" means necessity!), or somewhere in between (which could be more akin to balancing). These steps, the role of alternative employment practices, queasiness about balancing in this context, and other matters are illuminated by this article's framework.

Part IV considers the doctrine of strict scrutiny—and proportionality analysis (employed in many countries and proposed by some for the United States)—which likewise has a structure that bears some similarity to the stylized inquiry examined here. (Strict scrutiny omits step 3, but may in part introduce balancing through its step 2 or its less restrictive alternatives analysis, referred to as narrow tailoring.) Central questions concern the two thresholds: their height and their nature. Specifically, the first threshold is largely taken to be qualitative and categorical rather than quantitative, essentially limiting the domain of review. Particularly with regard to constitutional provisions, it sometimes makes sense to employ constraints on balancing, here by the legislative and executive branches, and to focus judicial review on realms in which other governmental actors cannot be trusted. However, once the first step is triggered, the second step's requirement of a compelling state interest seems to entail a threshold that mixes qualitative and quantitative features in a manner that is difficult to interpret or justify. The combination of the two steps is accordingly obscure on the extent to which balancing by the court is avoided, and any such deviation seems difficult to square with the purposes of review.

Proportionality analysis, in some incarnations, involves a number of steps that are even more suggestive of the structured decision procedures considered here. Yet the stance taken toward the height of the thresholds seems inconsistent. Some of the elucidation and rationalization proceeds as though certain thresholds are tough, which allows cases to be disposed of before reaching proportionality review's final balancing step. If so, there are potentially significant costs and essentially no benefits (because, as noted earlier, the balances avoided are either easy or are ones that may well have favored an opposite conclusion). Other descriptions, seeming to sense the problem that early truncation may involve erroneous outcomes based on incomplete analysis, imply that the thresholds are instead negligible, in which event one essentially has unconstrained balancing that renders all but the final step moot. Matching these protocols against this article's stylized procedure and accompanying analysis makes more apparent these inconsistencies and illuminates the costs if structured proportionality review substantially deviates from unconstrained balancing.

These applications concretize and clarify the article's more general analysis and indicate its usefulness. This article neither advances definitive claims regarding the actual state of doctrine in any of these areas of law nor suggests what doctrinal formulations would be best. Instead, it seeks to provide a general framework for understanding the operation and implications

[2] This article does not address disparate treatment doctrine that employs a superficially similar regime, although some of the lessons developed here may be applicable.

of an important aspect of legal rule design. Along the way, it illuminates a number of broad questions regarding the legal system, including information collection as a central input to adjudication, various dimensions of balancing (quantification, commensurability, and constraints), less restrictive alternatives, and overlooked mismatches between stated legal doctrine and practice.

I. Balancing Versus Structured Decision Procedures[3]

A. *Final Decision Rule*

1. Analysis

Let us begin with a precise statement of the stylized setting that will be used as a benchmark for analysis in parts II–IV. As we will see, this bare-bones description of structured decision procedures does not fully depict any of the legal doctrines examined there; those doctrines are in various respects murky and contested, differ from each other, and (depending on the interpretation) diverge in various ways from this baseline. The stark formulation offered here is chosen for clarity, and it indeed will enhance our understanding of each application.

A legal decisionmaker—an agency, court, or other tribunal—is confronted by a case.[4] Its ultimate choice is whether or not to assign liability, which for ease of exposition will be taken to involve injunctive relief (the application to ex ante behavior is elaborated in the margin).[5] The imposition of liability, relative to a finding of no liability, results in the avoidance of a harm of H and forgoes a benefit of B. Either value might be zero, each may be highly uncertain, and part of the task (elaborated in section B) involves the gathering of information to sharpen these estimates.[6]

Under balancing,[7] liability is assigned if and only if $H > B$. (Ties are largely ignored, with conventional tie-breaking norms employed without further comment.) Because there is often uncertainty, which may remain significant even after information gathering is complete, H and B are best interpreted as expected values (with risk adjustments, as appropriate).

Under a structured decision procedure, it will be assumed that liability is determined by the following three-step protocol:

[3] This Part is a brief sketch of the ideas developed in much greater depth in Kaplow, *supra* note . Rather than restating numerous qualifications, elaborations, and subtleties, the reader is referred there for further discussion.

[4] Most of the discussion abstracts from the fact that the flow of cases is itself endogenous to the decision procedure. One important dimension concerns screening, including discouraging the filing of frivolous cases and avoiding the suppression of valid, valuable cases—a topic addressed briefly at the end of subsection B.1.

[5] In many legal settings, a significant, even primary function of liability is to deter harmful conduct while avoiding the chilling of beneficial behavior. The H and B employed here can be interpreted as stand-ins for deterrence benefits and chilling costs, although there are important (but subtle) differences between this function of liability and settings in which the decision concerns the prohibition or permission of an act, going forward (which is nominally true for merger review, zoning decisions, drug approval, and injunctions more broadly). For a formal analysis of the differences, see Louis Kaplow, *On the Optimal Burden of Proof*, 119 J. POL. ECON. 1104 (2011). Informal analysis and substantial elaboration appear in Louis Kaplow, *Burden of Proof*, 121 YALE L.J. 738 (2012) [hereinafter Kaplow, *Burden of Proof*], and Louis Kaplow, *Likelihood Ratio Tests and Legal Decision Rules*, 16 AM. L. & ECON. REV. 1, 5–10 (2014) [hereinafter Kaplow, *Likelihood Ratio Tests*]. The analysis is extended to multistage decisionmaking in Louis Kaplow, *Multistage Adjudication*, 126 HARV. L. REV. 1179 (2013) [hereinafter Kaplow, *Multistage Adjudication*], and Louis Kaplow, *Optimal Multistage Adjudication*, 33 J.L., ECON., & ORG. 613 (2017).

[6] Much of the discussion in this article abstracts from administrative costs, although section B's analysis of information gathering addresses a significant aspect of this subject.

[7] To avoid excessive verbiage, the term "balancing," standing alone, refers to pure or unconstrained balancing, in contrast to a structured decision procedure, unless the context clearly indicates otherwise (notably, when reference is being made to the balancing that occurs in step 3 of such a structured procedure).

(1) If $H > H^*$, proceed to step 2. Otherwise, assign no liability and stop.

(2) If $B > B^*$, proceed to step 3. Otherwise, assign liability and stop.

(3) If $H > B$, assign liability. Otherwise, assign no liability. And stop.

Let us now compare this stylized structured decision procedure to balancing as a final decision rule, taking as given that there is some set of information before the tribunal. The structured rule can, and sometimes will, err in two ways. First, it may call for no liability even though the harm exceeds the benefit. This possibility is a direct consequence of step 1's decision threshold. No liability is assigned whenever $H \leq H^*$. This outcome, moreover, is determined without regard to the magnitude of B, so it is possible that $H > B$. This happens whenever B falls in the range from 0 to H, that is, when we have $0 \leq B < H \leq H^*$. The only way to eliminate this possibility is to set H^* to zero, guaranteeing that the first step never matters.[8]

Second, the structured protocol can result in liability even though the benefit exceeds the harm. This arises in some cases in which step 2 is binding. There, liability is assigned when, having found that $H > H^*$ in step 1 (which is required to reach step 2), we also have $B \leq B^*$. This outcome does depend on the magnitude of both B and H, but it is not determined by a direct comparison of the two as it would be under balancing. A divergence in outcomes can arise when B^* is sufficiently large because then it is possible that $H < B$ even though step 2 fails. This happens whenever B falls in the range from H to B^*, that is, when we have $H^* < H < B \leq B^*$.

In reflecting on this result, it is useful to focus on the possible relationships between H^* and B^*. As is clear from the sequence of inequalities, this problem can occur when $H^* < B^*$. By contrast, when $H^* > B^*$, we can see that this situation cannot arise. Interestingly, the impossibility of mistakenly assigning liability in this case arises precisely because step 2 is rendered redundant: If, after step 1, we went straight to balancing, we would assign liability only when $H > B$. But we already know from step 1 that $H > H^*$, and we are assuming that our protocol sets $H^* > B^*$, which implies that, as we leave step 1, we know that $H > B^*$. It is pointless to ask first whether B is at least as high as B^* when, if it is, we will then immediately ask whether it is at least as high as H, a more demanding test.

Anticipating the discussion of applications in parts II–IV, it is sometimes suggested in particular legal contexts that structured decision procedures are appealing because they save effort by avoiding difficult balancing. But the balancing of H and B is hardly difficult in this case; instead, we may sometimes be making extra work for ourselves. When it is a close question whether $B > B^*$, we need to struggle with step 2 even though the final outcome would be immediately obvious if we peeked ahead to the balance required in step 3, which in this instance is not as close a call.[9]

It is also useful to examine step 2 from another perspective, asking what is the relationship between B^* and H rather than between B^* and H^*. If $B^* > H$, we know that we can get the wrong outcome. After all, step 2 asks whether $B > B^*$, knowing full well that $B^* > H$. On the other hand, if $B^* < H$, we cannot err. Here, we are asking a pointless question because the assessment of whether $B > B^*$ will immediately be followed (if it passes) by the more

[8] Subsection B.1 briefly explores whether setting H^* somewhat above zero may be useful to screen cases.

[9] This point about possible added effort also applies to step 1, if one now thinks ahead two steps from there: Even when step 1 passes, that is, $H > H^*$, we then have to determine B, which we would have to do under balancing as well. Also, if step 1 was close, we may have had to undertake extra effort even if it would ultimately turn out that the $H > B$ balancing in step 3 would have been easy because B was notably lower than H. Of course, step 1 sometimes does save work, namely, the need to examine B at all when step 1 fails—but this savings arises precisely in the cases in which we may be reaching the wrong outcome on account of step 1's decision threshold, as already explained.

stringent test of whether $B \geq H$. And, when the step 2 test, $B > B^*$, fails, it would have been clearer that the more stringent test of whether $B \geq H$ would have failed, so we are hardly easing the decision task.

The implication is that setting $B^* = H$ avoids these problems. But it does so by converting the second step into step 3's balancing inquiry. After all, asking whether $B > B^*$ and whether $B > H$ are the precisely same questions when $B^* = H$.

Summarizing this subsection, if in step 1 we always set $H^* = 0$, and then in step 2 we always set $B^* = H$, all the shortcomings of the structured decision procedure as a final decision rule are avoided—because we have converted it into an unconstrained balancing test. The most significant point, however, is that, whenever the two decision methods generate different results, the outcome under the structured decision procedure is always the one that is in error. Moreover, such errors can be made even when all the information required for balancing has already been processed (which occurs when errors are made at step 2). Perhaps most remarkable, structured protocols are thought to prevail in important areas of law, and are sometimes proposed as replacements for balancing in others, without having even asked how the outcomes under the two approaches differ.

2. Additional Considerations

This subsection examines less restrictive alternatives and offers some reflections on balancing as a decision rule. Beginning with the former, it is common under structured decision procedures to append in some fashion an inquiry into less restrictive alternatives, using here the terminology of antitrust law[10] for the sort of supplement that is referred to in Title VII disparate impact law as alternative employment practices and in constitutional law's strict scrutiny as narrow tailoring (or, in proportionality analysis, as minimal impairment). The central idea is that, when a defendant purports to justify an action by reference to its producing B, we should consider whether some or all of that B might be achieved through an alternative arrangement that causes less H.

Under unconstrained balancing, suppose that a proffered less restrictive alternative would generate its own levels of harm and benefit, which here will be denoted H' and B', respectively. If the alternative is indeed less restrictive, $H' < H$. For it to be more desirable, the net social harm must be less than that from the defendant's original action: $H' - B' < H - B$. Note that because this inquiry into less restrictive alternatives only matters when $H \leq B$, we can further state that $H' - B' < H - B \leq 0$. In considering this series of inequalities, one performs two balancing tests: the original one (which, if it had been $H > B$, would have resulted in an assignment of liability with no need to inquire into less restrictive alternatives) and a second one (comparing the alternative to the original practice).

It is also helpful to restate the less restrictive alternatives test as a "delta/delta" test. Starting with $H' - B' < H - B$ from just above, we can rearrange terms to express this equivalently as $H - H' > B - B'$. That is, a less restrictive alternative is superior to the original action when it reduces the harm by more than it reduces the benefit. Introducing the further notation $\Delta H = H - H'$ and $\Delta B = B - B'$, this rearranged version can also be written as the requirement that $\Delta H > \Delta B$. Instead of performing a second balancing test, we can directly compute the two deltas and see which is greater.

[10] This is the rubric under the rule of reason, as will be developed in section II.A. For mergers, examined in section II.B, the test is phrased as a requirement that purported efficiencies (B) be "merger specific."

Consider next how less restrictive alternatives analysis fits into structured decision procedures. The inquiry is usually placed in or (more often) after step 2 (and sometimes, as we will see in parts III and IV, this inquiry is in lieu of step 3's balancing inquiry). The core oddity is that, as we have seen, the proper way to examine less restrictive alternatives involves quantifying H, B, H', and B', in order to perform the second balancing test or the delta/delta test. Yet, as ordinarily imagined, this analysis is performed before undertaking any explicit comparison of H and B. So the decisionmaker must do all that is required for full balancing, and more, even though the balancing step has not yet been reached (or, under some protocols, is not even undertaken).[11] Another surprising feature is that the effort involved with less restrictive alternatives analysis seems to be mandated even if $H > B$ to begin with, in which case it seems pointless—and this may well become apparent in the course of analyzing a case. Worse, under protocols that consider less restrictive alternatives but have no step 3, a defendant may be found not liable when the less restrictive alternative fails even though the analysis thereof made clear that $B < H$ despite step 2 having passed. In summary, even though the consideration of less restrictive alternatives improves structured decision rules, the resulting mechanism is clumsy and fails to eliminate either important type of error.

Turn now to some perennial questions raised by the use of balancing as a decision rule. The reluctance to engage in explicit balancing in a number of legal settings seems to be part of the explanation for the embrace of structured decision procedures. Queasiness about balancing is often created by the difficulty of quantifying one or both of the desiderata that need to be balanced or from the felt inappropriateness of expressing the two in a common metric (often referred to as incommensurability). The core response is that, however great the challenges, something akin to balancing is the only plausible way to proceed when there are competing considerations, each sometimes powerful enough relative to the other to be decisive.

As a matter of logic, if there should be liability when harm vastly exceeds the benefit in some appropriate sense, but no liability when the harm is minuscule and the benefit immense, then consistent, coherent decisionmaking requires quantification (even if hunches and guesstimates may be required) and comparison. Regarding the latter, even if the decisionmaker does not consciously or explicitly state two quantified factors in a common denominator, the decisions can be viewed *as if* determined by balancing as long as greater harm, all else equal, always favors liability and greater benefit, all else equal, always opposes liability.

Note further that, as a practical matter, such balancing is routinely employed in many settings in which both quantification and comparison may be extremely difficult. Medical decisionmaking is an obvious example. Diagnosis is often uncertain, evidence is limited, each patient is unique, and the stakes often involve life, quality of life, and cost, all of which must be somehow weighed in reaching a treatment decision. Would any sensible individual follow a doctor's advise that eschewed quantification or harm, benefit, or both? And, whatever the challenge, can one really decide whether the benefit of treatment is worth the risk without considering the relative importance of each?

Finally, observe that structured decision procedures hardly avoid these problems. H must be quantified at step 1, and B as well if step 2 is reached. Moreover, not only must comparisons be made when at step 3, but any plausible attempt to set the two thresholds, H^* and B^*, inevitably involves judgments about the relative importance of harms and benefits.[12]

[11] Sometimes, however, it will be possible to apply the delta/delta test directly, without separately assessing the harm and benefit of each practice relative to the situation with neither, by comparing the two practices to each other.

[12] In operation, structured decision procedures may succeed in allowing decisionmakers (such as judges) to avoid having to *state* any of their quantitative conclusions because, at step 1 and step 2, they can merely announce whether the

A different type of objection to balancing concerns the virtue in some settings of designing rules to constrain balancing, particularly by agents with limited institutional competence or who may be untrustworthy. With respect to government actors, who are constrained variously by constitutions (see Part IV) and myriad other rules, regulations, and standard operating procedures, one may be particularly concerned in some settings about the abuse of power (notably, to entrench incumbents) and under- (or negative-) weighting of certain groups' interests.

The relevant question for present purposes is the extent to which structured decision rules may be helpful in this regard. On reflection, the fit is imperfect. Often it is the primary actor rather than the reviewer that needs to be controlled. And when the reviewer itself cannot be trusted—perhaps it would manipulate the reported *H* and *B* to distort the balancing decision—similar efforts would often circumvent structured decision rules. Indeed, the matter can be worse, for even if the manipulation of *H* and *B* might sometimes be detected, the decisionmaker has the further degree of freedom of manipulating *H** and *B** since both are usually stated in fuzzy rather than explicit, quantitative terms. Nevertheless, aspects of structured decision rules might sometimes be helpful,[13] and employing a categorical (rather than quantitative) inquiry at step 1 to limit the reviewer's jurisdiction might be regarded to be appropriate in some settings, as suggested in Part IV's discussion of constitutional law.

B. *Information Gathering*

1. Analysis

Section A examines final decisionmaking, taking as given the information (evidence) before the decisionmaker. This section focuses on information gathering, a critical but under-analyzed feature of legal decision processes.[14] Moreover, structured decision procedures are sometimes favored on the ground that they economize on information costs because they consider first only information on *H* and, when step 1 fails, proceedings are terminated without having to examine information on *B*.

Let us begin with how information should optimally be collected—the procedure that ideally would be followed under unconstrained balancing. Whether to collect some clump of information is determined by what is referred to as the value of information.[15] One assesses what decisions would be made in light of what one might learn, determines the net of expected harm and benefit under each, and weights these outcomes by their respective probabilities. That total expected value is then compared to the expected value that would be generated if one made the best possible decision, liability or no liability, as currently informed. The excess of the former over the latter is the value of information, and the information should be collected if this

(unquantified) thresholds are exceeded. Moreover, even if they believe that both steps pass so that balancing is required, they can obfuscate by adjusting one of their conclusions at an earlier step — for example, by stating that step 2 fails if they believe that $H > B$, thereby avoiding the need to articulate how they balanced *H* and *B*.

[13] The most promising situation is one in which *H* is externally observable to some extent and *H** is set explicitly, whereas *B* is externally inscrutable. Then step 1 could be helpful in this regard (but not step 2).

[14] For further discussion of a number of related issues concerning information and adjudication, see Louis Kaplow, *Information and the Aim of Adjudication: Truth or Consequences?*, 67 STAN. L. REV. 1303 (2015); Kaplow, *Multistage Adjudication*, *supra* note ; and Louis Kaplow, *The Value of Accuracy in Adjudication: An Economic Analysis*, 23 J. LEGAL STUD. 307 (1994)

[15] For a simple exposition and illustration aimed at a legal audience, see HOWELL E. JACKSON ET AL., ANALYTICAL METHODS FOR LAWYERS 13–19 (3d ed. 2017).

overall value exceeds the cost. Favoring the collection of additional information are low information costs and high diagnosticity of the information. This latter notion reflects how close is the initial decision, how much uncertainty is present, and how informative is the information one contemplates collecting.

When considering more complex settings with many possible clumps of information that might be collected, a number of key principles guide the way. First, one prioritizes information with the highest diagnosticity to cost ratio. Second, which clump to collect next (and so forth) depends on what is learned along the way. Third, whether to collect, say, two clumps together rather than sequentially depends on a tradeoff of synergies from simultaneous collection and option value (the likelihood that what is learned from the first clump will render collection of the second suboptimal, and the cost savings therefrom, net of any information loss from stopping).

Each of these principles is intuitive and can be understood by contemplating medical diagnosis. One begins with the cheapest, least invasive, and most informative tests. Whether to stop and decide (initiate treatment or send the patient home), and what if anything to learn next (a scan? a biopsy?) depends on the initial test results; hence, the full plan is not determined a priori. And one might, say, do multiple blood tests simultaneously because of savings in cost and time, whereas one might hold off on an expensive and painful biopsy that would be unnecessary if the blood test is negative.

Turn now to structured decision procedures. In principle, they would collect first all (but only) information pertaining to H and then proceed to B if but only if step 1 passes. Despite some superficial appeal, this preset plan grossly violates the foregoing principles and is subject to additional, significant infirmities.

Its first major deficiency arises from the implicit supposition that most evidence pertains only to H or only to B, whereas in fact much illuminates both. A notable case, which is quite important in many of the applications considered below, concerns characterization evidence. If the question is whether the defendant's act is of the harmful or the beneficial type—and suppose for ease of exposition that there are only two, mutually exclusive possibilities—then by definition any evidence that affects the probability that it is of the harmful type also affects the probability that it is instead beneficial (indeed, by precisely the same amount, although in the opposite direction). The predicate that inquiries into harm and benefit are distinct is thus incoherent. To illustrate this problem, ask how one could interpret an arguably ambiguous internal document if allowed to consider only one of the possible meanings. Or, anticipating section II.A on antitrust's rule of reason, how can one assess whether a practice is anticompetitive (of a type causing H) when that is typically *defined* as action *other than* "competition on the merits" (of a type causing B).[16]

Second, in practice information often clumps by source (sets of documents, particular witnesses) rather than by issue. There would be huge synergy loss from making two passes: for example, reviewing the same sets of documents twice, first for H and later, if step 1 passes, for B. In medical diagnosis, it would be akin to taking a blood sample and only collecting enough to run one test, waiting for the result, and then resampling to run the other test. (That may sometimes be appropriate, but often not.) Observe that an implication of these first two problems is that, in applying step 1, the decisionmaker may already have in hand much information pertaining to B, which may indicate that $B < H$, yet it is supposed to determine whether $H > H^*$ and, furthermore, to assign no liability (and stop) if it does not—without any regard for B.

[16] *See infra* note .

Third, and going to the heart of the foregoing analysis of the value of information, the structured procedure's sequencing is usually wrong, often dramatically so, even if one ignores the first two defects. Sometimes, the information cluster with the highest diagnosticity to cost ratio will pertain to *B*. Due to diminishing returns, it would be atypical for all of the information pertaining to *H* to rank higher than any pertaining to *B*. Also, as explained, the optimal sequence is contingent, depending on what is learned at each step, not preset in advance. And the tradeoff of synergy and option value across information pertaining to *H* and to *B* is entirely ignored. In every respect, structured decision procedures would often lead decisionmakers badly astray with regard to information gathering.

Before leaving this subject, consider the implications of these three points for setting *H** modestly above zero, as a first step, in order to screen cases.[17] On one hand, in light of administrative costs and the prospect that, otherwise, many low-merit cases might be filed, this approach has some appeal. On the other hand, in light of the foregoing, unconstrained balancing—perhaps requiring an early indication that *H* nontrivially exceeds *B*—seems to be a superior screening strategy. After all, if much information concerns both *H* and *B* in any event, and if any distinct information that is readily obtainable may often pertain to *B* rather than *H*, then screening based on whatever may be known about both *H* and *B* makes more sense than straitjacketing the decisionmaker to screen based only on *H*.

2. *Conduct of Legal Proceedings*

The actual conduct of U.S. civil litigation (focusing here on federal courts) follows neither these structured protocols (when applying doctrines under which they are applicable) nor optimal ones. By contrast, one suspects that specialized agencies often proceed roughly in accord with the latter, particularly regarding cases that they choose at some point to terminate without liability. This may be so to a significant extent even when structured decision rules govern, although agencies may need to "repackage" their analysis in issuing an opinion or seeking enforcement in court when they do seek to impose liability.

Once a government agency or a private plaintiff finds itself in court, the game changes substantially. If a complaint's adequacy is challenged at a motion to dismiss, the only question before the court is whether the challenger has stated a plausible claim.[18] Under a pure balancing test, the plaintiff must allege that $H > B$, whereas under the structured decision procedure the plaintiff must instead allege that $H > H^*$. It is unclear how much this difference matters in practice or, when it does, which hurdle would be easier to overcome (which, among other things, depends on the magnitude of *H**).

When a motion to dismiss is denied or none was filed, the case proceeds to discovery. Ordinarily, the scope of discovery covers all issues and all types of evidence, subject to limits regarding burdensomeness, what is now called "proportional to the needs of the case."[19] The key point is that, unless a judge chooses to engage in substantial case management, the ordinary conduct of discovery does not involve sequencing. It does not adhere to the principles of optimal information collection, which would require interim assessments and stopping decisions that are associated with a particular decision regarding liability. Nor does discovery follow the dictates of structured information protocols, which would call for discovery only on *H*, followed by a determinative resolution of whether $H > H^*$, which would have to be answered affirmatively

[17] *See* Kaplow, *supra* note , at 1043–47.

[18] *See* Bell Atlantic Corp. v. Twombly, 550 U.S. 544 (2007); Ashcroft v. Iqbal, 556 U.S. 662 (2009).

[19] FED. R. CIV. P. 26(b)(1).

(requiring complete factfinding) before proceeding to discovery pertaining to *B*. Thus, when some advance structured decision procedures because they sometimes save the costs of collecting information on *B* in the course of civil litigation, it is mysterious what they have in mind.

After discovery, a party may move for summary judgment. Under either balancing or a structured decision procedure, this would ordinarily involve a motion by the defendant, typically claiming in essence that the there is a negligible evidentiary basis (sufficient to create a "genuine dispute") for believing that *H* is nontrivial.[20] Note, as just explained, that at this point discovery would ordinarily have been completed on all issues, so even if the motion is granted, resulting in no liability, all information gathering pertaining to *B* will have occurred even under a structured decision procedure.

If a case goes to trial, any decision under either balancing or a structured decision procedure will typically not be made until the end. (The exception involves a motion for judgment as a matter of law at the end of the plaintiff's case, which relates to the core situation in which a defendant should prevail on a motion for summary judgment.[21]) In that event, even under a structured decision procedure, in a case in which there is no liability because step 1 fails (that is, the factfinder ultimately concludes that $H \leq H^*$), there will not even be a savings in trial costs (and, as noted, certainly not in discovery costs) as long as there had been a genuine dispute about whether this was so. The only savings would be in the factfinder's final deliberation efforts, for a judge or a jury may enter a finding of no liability if it concludes that step 1 fails, without deciding steps 2 and 3. As a practical matter, verdicts and (with bench trials) opinions often complete all of the steps, in part because of the possibility of an appeal. The belief that structured decision rules economize substantially on litigation costs seems to be a mirage.

II. ANTITRUST

Parts II–IV, the core of this article, apply Part I's analysis to a number of legal settings. The characterizations of various legal doctrines are simplified for present purposes, often stating them bluntly at the expense of various subtleties. Moreover, there is significant ambiguity surrounding all of them, and each differs from the others and from the stylized structured decision procedure just examined. Indeed, some of the benefit of applying this article's framework is to bring these ambiguities and variations into sharper focus, raising new questions in addition to suggesting answers to some familiar ones.

A. *Rule of Reason*

This section focuses on what is sometimes referred to as a "structured" rule of reason. Challenges to restraints of trade under Section 1 of the Sherman Act[22] have been governed for

[20] FED. R. CIV. P. 56(a); *see* Celotex Corp. v. Catrett, 477 U.S. 317 (1986); Anderson v. Liberty Lobby, Inc., 477 U.S. 242 (1986); Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574 (1986).

[21] Under *Anderson*, 477 U.S. at 249–50, the standard for summary judgment under FED. R. CIV. P. 56 is the same as that for judgment as a matter of law under FED. R. CIV. P. 50.

[22] 15 U.S.C. § 1. In a broad sense, the rule of reason also governs Section 2, 15 U.S.C. § 2, and some suggestions have been made regarding the deployment of a structured rule of reason inquiry in this monopolization context as well. The rule of reason, as announced in *Standard Oil Co. v. United States*, 221 U.S. 1 (1911), was explicitly directed at interpreting Sherman Act § 1, but the Court indicated that the inquiry is the same under § 2. *See id*. at 61–62; *see also, e.g.*, United States v. Microsoft Corp., 253 F.3d 34, 59 (D.C. Cir. 2001) (en banc) (per curiam) (suggesting a "similar

over a century by the rule of reason, which is understood to impose liability when anticompetitive effects (H) exceed procompetitive effects (B), essentially a balancing test.[23] In recent decades, however, courts and a number of commentators,[24] as well as model jury

balancing approach" under the two sections and citing *Standard Oil*). Nevertheless, much of the development of the rule of reason, under that phraseology, as a formal test, has occurred under Section 1, growing out of *Chicago Board*, as discussed later in this section. The development of the law of monopolization, interpreting Section 2, also post-dates *Standard Oil*. Courts examining particular practices under Section 2 tend not to mention the rule of reason as such. *See, e.g.*, Brooke Grp. Ltd. v. Brown & Williamson Tobacco Corp., 509 U.S. 209 (1993) (not mentioning the rule of reason in its decision on predatory pricing); United States v. Dentsply Intl., Inc., 399 F.3d 181 (3d Cir. 2005) (same, for exclusive dealing, in an opinion examining the practice under Section 2). But some do. *See, e.g.*, McWane, Inc. v. FTC, 783 F.3d 814, 833 (11th Cir. 2015) (describing the burden-shifting technique as "a structured, 'rule of reason'-*style* approach" (emphasis added)); *see also* Mark S. Popofsky, *Defining Exclusionary Conduct: Section 2, The Rule of Reason, and the Unifying Principle Underlying Antitrust Rules*, 73 ANTITRUST L.J. 435, 437 (2006) ("[T]he few clear guideposts in Section 2 case law demonstrate that courts properly apply different Section 2 legal tests to different conduct. The unifying principle is that each Section 2 legal test reflects a specific expression of the same underlying 'rule of reason.' Although courts usually describe the rule of reason as a particular step-wise test for assessing the legality of concerted action, the rule of reason more generally provides a *principle* for generating antitrust liability tests in a common-law fashion.").

Further suggestions on the use of a structured rule of reason for Section 2 cases appear in U.S. DEP'T OF JUSTICE, COMPETITION AND MONOPOLY: SINGLE-FIRM CONDUCT UNDER SECTION 2 OF THE SHERMAN ACT, at viii (2008) (The executive summary of the section on "General Conduct Standards" in the monopolization context begins as follows: "The plaintiff should have the initial burden of establishing that challenged conduct harms the competitive process and therefore has a potentially anticompetitive effect. If plaintiff carries that burden, defendant should have the opportunity to proffer and substantiate a procompetitive justification for the challenged conduct. If defendant does so, plaintiff then should have the burden of establishing that the challenged conduct is anticompetitive under the applicable standard.") (this report was not joined by the Federal Trade Commission, which had participated jointly in the hearings and other work leading up to the report (*see* Press Release, Fed. Trade Comm'n, FTC Commissioners React to Department of Justice Report, Competition and Monopoly: Single-Firm Conduct Under Section 2 of the Sherman Act (Sept. 8, 2008), http://www.ftc.gov/news-events/press-releases/2008/09/ftc-commissioners-react-department-justice-report-competition-and), and the report was withdrawn the next year when the administration changed (*see* Press Release, U.S. Dep't of Justice, Justice Department Withdraws Report on Antitrust Monopoly Law (May 11, 2009), https://www.justice.gov/opa/pr/justice-department-withdraws-report-antitrust-monopoly-law); it appears that most of the disagreement concerned the report's statement of substantive rules governing single-firm conduct in a manner that objectors regarded to be too lenient, such as being too generous in safe-harboring behavior or requiring that anticompetitive effects significantly outweigh procompetitive ones, with no suggestion that the overall framework was problematic). *See also* Jonathan B. Baker, *Exclusion as a Core Competition Concern*, 78 ANTITRUST L.J. 527, 543–51 (2013) (arguing that the sequenced, multi-element, burden-shifting framework under Section 1's rule of reason is increasingly being applied to exclusion claims, including under Section 2). In addition, similar decisionmaking rubrics have been proposed by academics to address particular exclusionary practices. *See, e.g.*, Patrick Bolton, Joseph F. Brodley & Michael H. Riordan, *Predatory Pricing: Strategic Theory and Legal Policy*, 88 GEO. L.J. 2239, 2262–85 (2000) (proposing a sequential, burden-shifting rule to govern predatory pricing cases).

For analogous statements and advocacy in the European Union, see, for example, *Communication from the Commission — Guidance on the Commission's Enforcement Priorities in Applying Article 82 of the EC Treaty to Abusive Exclusionary Conduct by Dominant Undertakings*, 2009 O.J. (C 45) 7, ¶ 31 [hereinafter *Guidance on Article 82*] (concluding its statement regarding procompetitive justifications by stating: "It is incumbent upon the dominant undertaking to provide all the evidence necessary to demonstrate that the conduct concerned is objectively justified. It then falls to the Commission to make the ultimate assessment of whether the conduct concerned is not objectively necessary and, based on a weighing-up of any apparent anti-competitive effects against any advanced and substantiated efficiencies, is likely to result in consumer harm."), and Miguel de la Mano & Benoît Durand, *A Three-Step Structured Rule of Reason to Assess Predation Under Article 102* (DG Competition, European Commission, Office of the Chief Economist Discussion Paper, 2010) (proposing a structured rule for predatory pricing). For criticism, see Hans W. Friederiszick & Linda Gratz, *Hidden Efficiencies: The Relevance of Business Justifications in Abuse of Dominance Cases*, 11 J. COMPETITION L. & ECON. 671 (2015).

[23] The per se rule and variations such as the "quick look" are examined at the end of this section.

[24] As an illustration, consider the formulation presented in 1 ABA SECTION OF ANTITRUST LAW, ANTITRUST LAW DEVELOPMENTS 61–62 (Darren S. Tucker et al. eds., 8th ed. 2017) (emphasis added, footnotes omitted) [hereinafter ABA ANTITRUST LAW DEVELOPMENTS]:

> Since the early 1980s, lower courts have imposed greater structure on rule of reason analysis by casting it in terms of shifting burdens of proof. Although the precise formulation varies somewhat from circuit to circuit, the

> approaches are generally similar. Under the more structured rule of reason, the plaintiff bears the initial burden of *proving* that an agreement has had or is likely to have a *substantially* adverse effect on competition. If the plaintiff meets its initial burden, the burden *shifts* to the defendant *to produce evidence* of the procompetitive virtues of the conduct. *If* the defendant does *produce evidence* of procompetitive virtues, then the plaintiff *must show* that the challenged conduct is not reasonably necessary to achieve the stated objective *or* that the anticompetitive effects nonetheless outweigh the procompetitive virtues. *The ultimate issue*, then, is whether the restraint's anticompetitive effect *substantially outweighs* the procompetitive effect for which the restraint is reasonably necessary.

The emphasized language highlights features (not shared by all formulations) that will be discussed below: the plaintiff's initial burden is one of *proving* (suggesting a persuasion burden) anticompetitive effects that are *substantial* (suggesting $H^* > 0$); if that is done, the burden *shifts* to the defendant *to produce evidence* (suggesting a production burden) of procompetitive effects; *if* (but only if) that is done, the plaintiff *must show* (suggesting a persuasion burden) either that the restraint is unnecessary (phraseology often associated with less restrictive alternatives analysis) *or* that the anticompetitive effects outweigh the procompetitive ones (note the lack of a modifier to "outweigh" and the lack of explicit mention of direct rebuttal even though only a production burden regarding B has been met); and that *the ultimate issue* involves a balancing test, and it requires that the anticompetitive effect *substantially outweighs* the procompetitive effect. *But see id.* at 79 (stating an equipoised balance).

Consider another, seemingly similar version (which cites an earlier edition of the preceding source in support and which, in turn, is quoted as the exemplar of the structured rule of reason in Andrew I. Gavil, *Burden of Proof in U.S. Antitrust Law, in* 1 ISSUES IN COMPETITION LAW AND POLICY 125, 146 (Wayne Dale Collins ed., 2008), a survey chapter on the operation of burdens of proof in U.S. antitrust law):

> Courts have imposed a consistent structure on rule of reason analysis by casting it in terms of shifting burdens of proof. . . . . Under this approach, the plaintiff bears the initial burden of *showing* that an agreement had a *substantially* adverse effect on competition. . . . If the plaintiff meets this burden, the burden *shifts* to the defendant *to come forward with evidence* of the procompetitive virtues of the alleged wrongful conduct. . . . *If* the defendant is able *to demonstrate* procompetitive effects, the plaintiff then *must prove* that the challenged conduct is *not reasonably necessary* to achieve the legitimate objectives *or* that *those objectives can be achieved in a substantially less restrictive manner*. . . . *Ultimately, if these steps are met*, the harms and benefits *must be weighed* against each other in order to judge whether the challenged behavior is, *on balance*, reasonable.

Law v. NCAA, 134 F.3d 1010, 1019 (10th Cir. 1998) (emphasis added). Some notable differences are: although step 2 refers to the defendant *coming forward with evidence*, this is then referenced as the defendant having been able *to demonstrate* procompetitive effects; if that is done (and with no reference to an opportunity to rebut or whether the demonstrated B exceeds H), the plaintiff *must prove* either that the conduct is not necessary or that there exists a less restrictive manner of achieving them, which is to say that the plaintiff apparently loses if it cannot (even if it can directly rebut B or show that $H > B$, because, as stated next, the balancing step is reached only *if these steps are met*); the ultimate balancing itself is stated as a final step and not as the ultimate and thus perhaps superseding question; the final balance is unweighted, even though at step 1 the plaintiff had to show a *substantially* adverse effect on competition (meaning that it is possible for the plaintiff to lose at step 1 even though it would win if it had reached the final balance); and it is not clear why balancing would ever be required if the only way one can reach this step is when plaintiff has demonstrated a substantial H *and* it has shown that the challenged restraint is *not* necessary to achieve the defendant's B. It is worth reflecting on how many and substantial are the differences, particularly since this is a decision that repeatedly (in the ellipsed portions) cites the previous version in support. Moreover, as noted, this version is quoted in a survey essay as the exemplar.

Having already taken substantial space, I will at this point merely assert that, if one reviews such statements by every federal circuit court and in other sources, including the treatise by Phillip Areeda and Herbert Hovenkamp, one will see rough similarity from ten thousand feet but many such critical differences on close examination. Moreover, these differences appear within authorities: a single circuit may state the formulation differently across cases, and this treatise has multiple, conflicting formulations. *Compare* 7 PHILLIP E. AREEDA & HERBERT HOVENKAMP, ANTITRUST LAW ¶ 1507a (4th ed. 2017), *with id.* ¶ 1507c. *But see id.* at ¶ 1507f (concluding, in spite of the prior discussion advancing structured, multi-step inquiries, with a more flexible, case-specific approach). In addition, as best I can tell, no one seems to have noticed that apparently canonical formulations regarded to be largely the same are critically different in multiple ways and that many single statements have sharp internal conflicts. (In addition to those noted above, see, for example, *id.* at 443–44 (stating that if the plaintiff establishes, in their first step, that "the challenged practice *arguably* threaten[s] either to reduce output or raise price", and in their third step, that market power is "*plausible*"—the requirements ordinary associated with surviving a defendant's motion to dismiss or perhaps for summary judgment—the defendant will then lose the case, in their step 4, unless "there [is] *strong evidence* that the challenged practice creates *substantial* efficiencies" (emphasis added)).) Hence, even putting aside other issues raised in this section—notably the apparent conflict with Supreme Court precedent and the fact that nearly all such language is dicta given the procedural posture of the cases instructions,[25] increasingly state or advocate formulations involving a three-step structured decision procedure[26] of the sort sketched in subsection I.A.1 and examined throughout this article.[27] (The structured rule of reason under discussion here should be contrasted with the use of checklists of sorts that serve merely to remind the decisionmaker of pertinent considerations.[28]) The main rationales offered for this structured rule of reason are to economize

involved—it seems truly difficult to state what the law on this matter actually is. (In *Law v. NCAA*, quoted and discussed earlier in this footnote, the court rejected all of the defendant's procompetitive justifications—in a "quick look" rule of reason decision—so none of the subsequent steps were material to its decision. More often, the cases involve defendants' motions to dismiss or for summary judgment, so all of the action is at step 1.) Likewise, it is difficult to interpret commentators' references to this so-called structured rule of reason.

[25] *See* ABA SECTION OF ANTITRUST LAW, MODEL JURY INSTRUCTIONS IN CIVIL ANTITRUST CASES, 2005 EDITION, at A-4 (2005) [hereinafter MODEL JURY INSTRUCTIONS] ("Instruction 3A Rule of Reason – Overview. Under Section 1 of the Sherman Act, a restraint of trade is illegal only if it is found to be unreasonable. . . . [Y]ou must first determine whether the plaintiff has proven that the challenged restraint has resulted in [or is likely to result in] a *substantial harm to competition* in a relevant product and geographic market. If you [do], then you must consider whether the restraint produces *countervailing competitive benefits*. If you find that it does, then you must *balance* the competitive harm against the competitive benefit." (emphasis added)); *id.* at A-10 ("3C Rule of Reason – Evidence of Competitive Benefits. . . . The defendant has the burden of *producing evidence* regarding the existence of competitive benefits, and if the defendant *produces* such evidence, the burden *shifts* to the plaintiff to *prove* that the *restraint was not reasonably necessary* to achieve the benefits." (emphasis added)); *id.* at A-12 ("3D Rule of Reason – Balancing the Competitive Effects. *If* you find that the challenged restraint *was reasonably necessary* to achieve competitive benefits, *then you must balance* those competitive effects against the competitive harm resulting from the same restraint. If the competitive harm *substantially outweighs* the competitive benefits, then the challenged restraint is unreasonable." (emphasis added)). Comparison of the italicized phrases with those from the exemplars in the preceding footnotes reveals many similarities but also key differences, including that these instructions have both internally inconsistent and nonsensical aspects. Interestingly, this source further remarks in a footnote that is not part of the instruction itself: "In an effort to make the rule of reason instruction less confusing, it has been separated into four separate, but interrelated, instructions." *Id.* at A-4 n.1. This statement suggests that the drafters do not in fact view the sequenced structure to be part of the formal legal rule but rather as merely an aid in communicating the essence of the rule to juries, although the drafters also chose to withhold this explanation, including about the interrelationships, from the jury. (The more recent edition of these Model Jury Instructions, ABA SECTION OF ANTITRUST LAW, MODEL JURY INSTRUCTIONS IN CIVIL ANTITRUST CASES, 2016 EDITION (2016), is nearly identical. The only differences are the omission of the aforementioned note and that instruction 3C on competitive benefits omits mention of any burden shift to the defendant, although the notes that follow, seemingly not part of the instruction, do state that the defendant has a production burden in this regard. *See id.* at 8.)

[26] For previous advocacy of a structured rule of reason, see William F. Baxter, *The Viability of Vertical Restraints Doctrine*, 75 CAL. L. REV. 933 (1987), and Howard H. Chang, David S. Evans & Richard Schmalensee, *Some Economic Principles for Guiding Antitrust Policy Towards Joint Ventures*, 1998 COLUM. BUS. L. REV. 223. In the European Union's analysis of horizontal agreements that restrain competition under TFEU Article 101 (formerly Article 81), there is a two-step approach (wherein the latter two steps outlined in this article appear to be combined). *See Guidelines on the Applicability of Article 101 of the Treaty on the Functioning of the European Union to Horizontal Co-operation Agreements* (2011/C 11/01) ¶ 20 ("The assessment under Article 101 consists of two steps. The first step, under Article 101(1), is to assess whether an agreement between undertakings, which is capable of affecting trade between Member States, has an anti-competitive object or actual or potential restrictive effects on competition. The second step, under Article 101(3), which only becomes relevant when an agreement is found to be restrictive of competition within the meaning of Article 101(1), is to determine the pro-competitive benefits produced by that agreement and to assess whether those pro-competitive effects outweigh the restrictive effects on competition." (footnotes omitted)); *Communication from the Commission—Guidelines on the Application of Article 81(3) of the Treaty*, 2004 O.J. (C 101/08), 8, ¶ 11 (same, nearly verbatim).

[27] Often inquiries into market power, including determination of the relevant market, are included in step 1 or inserted, up front, as an additional step. This important feature will be abstracted from here, although it is important to emphasize that much of this article's criticism of structured decision procedures applies to this separation as well . *See* Louis Kaplow, *On the Relevance of Market Power*, 130 HARV. L. REV. 1303 (2017) [hereinafter Kaplow, *Market Power*]. In addition, the particular role of market definition is highly problematic regardless of how the inquiry is structured. *See, e.g.*, Louis Kaplow, *Why (Ever) Define Markets?*, 124 HARV. L. REV. 437 (2010) [hereinafter Kaplow, *Market Definition*].

[28] *Cf.* Cal. Dental Ass'n v. FTC, 526 U.S. 756, 782 (1999) (Breyer, J., dissenting) (". . . I would not simply ask whether the restraints at issue are anticompetitive overall. Rather, like the Court of Appeals (and the Commission), I would break that question down into four classical, subsidiary antitrust questions: (1) What is the specific restraint at

on the conduct of litigation[29] and to avoid the need for difficult balancing,[30] considerations taken to be particularly important given the complexity of many antitrust cases.

Analysis.—Step 1 requires, as typically stated, that the plaintiff show, demonstrate, or establish the existence of anticompetitive effects. Such language ordinarily indicates a burden of persuasion, which would be applicable only at the end of a trial,[31] yet these articulations are usually offered in the context of deciding motions to dismiss and for summary judgment. As

issue? (2) What are its likely anticompetitive effects? (3) Are there offsetting procompetitive justifications? (4) Do the parties have sufficient market power to make a difference?”). Interestingly, *United States v. Microsoft Corp.*, 253 F.3d 34, 58-59 (D.C. Cir. 2001) (en banc) (per curiam)—one of the cases often cited to illustrate circuit courts’ adoption of a structured rule of reason (in this instance, in the monopolization context)—introduces the listed items by describing them as “several principles,” *id*. at 58, even though their phrasing suggests a structured protocol.

[29] *See, e.g.*, U.S. Dep’t of Justice, *supra* note , at viii (“This allocation can enable courts to resolve cases more quickly and efficiently.”); *id.* at 36 ("Requiring plaintiffs to make a showing of harm to the competitive process at the outset facilitates the disposition of non-meritorious claims. . . . Likewise, requiring a defendant, upon a prima facie showing of harm to the competitive process, to come forward with a nonpretextual justification for its conduct enables courts and juries to condemn patently anticompetitive conduct without any weighing of offsetting effects. These steps can spare courts and juries difficult questions.”).

[30] *See, e.g.*, Areeda & Hovenkamp, *supra* note , at 442 (“Because both theory and data are usually insufficient, and because quantification in terms of a common denominator is usually impossible, balancing will inevitably be crude and should be avoided unless absolutely necessary.”); *see also* Rothery Storage & Van Co. v. Atlas Van Lines, Inc., 792 F.2d 210, 229 n.11 (DC Cir. 1986) (Bork, J.) (“[T]hough it is sometimes said that, in the case of restraints like these, it is necessary to weigh procompetitive effects against anticompetitive effects, we do not think that a useable formula if it implies an ability to quantify the two effects and compare the values found. . . . Weighing effects in any direct sense will usually be beyond judicial capabilities . . . .”); *id*. (presenting, as an alternative to quantifying and balancing the supposedly noncomparable effects, that courts should instead “draw[] inferences from market share and structure,” noting that “[a]ntitrust adjudication has always proceeded through inferences about market power drawn from market shares,” but failing to state what is inferred (presumably anti- and procompetitive effects), how the relevant inferences are to be quantified, and how the thus-inferred and quantified effects are then to be weighed when it is said to be impossible to do so when they are known explicitly); Herbert Hovenkamp, *Antitrust Balancing*, 12 N.Y.U. J.L. & Bus. 369, 370 (2016) (“‘Balancing’ requires values that can be cardinally measured and weighed against each other. The factors that are supposedly balanced in Sherman Act cases almost never fit this description. Even if the things requiring balancing did come in cardinal units, most times the courts would not have the tools necessary to make and apply the measurements.”). Paradoxically, Areeda and Hovenkamp seek to avoid balancing because of the insufficiency of “both theory and data,” yet seek to replace it with “tentative presumptions drawn from theory, experience, and the evidence at hand.” Areeda & Hovenkamp, *supra* note , at 442. As another means of avoidance, they later suggest: “If possible, we would quantify the magnitude, discounted by the probability, of the negative harm and positive benefit in terms of statutory values and hold conduct reasonable when the net is positive or unreasonable when negative. Because we almost never know enough to do this, we must find another way. The possibilities are to rule generally that harm always condemns or that benefit always saves, to delegate balancing to the jury, or to ask the judge to make a qualitative judgment guided by theory and experience. In most cases, the last course is the most sensible, but even it has significant limitations.” *Id*. at 448. Taken literally, it seems that, in cases in which a jury has found both a positive *H* and a positive *B*, they would have the judge then step in and perform the balance (*see also id*. (“Although juries have sometimes been left to do the balancing, this is clearly wrong.”)), which, moreover, would be done *qualitatively*, whatever that means, and, again, guided by the “theory and experience” that they previously stated does not exist. *See also* 3 Phillip E. Areeda & Herbert Hovenkamp, Antitrust Law ¶ 651e3 (4th ed. 2015) (subparagraph entitled “ Balancing generally to be avoided; burden-shifting.”); *id*. at 124 (stating that “[a] burden-shifting analysis should enable courts to avoid ‘close’ balancing in most situations,” but, as explained earlier in section I.A, when the measurements required for the earlier steps are sufficiently clear that the proper outcome is obvious, the balancing required by step 3 would not in fact be difficult, and when the balancing would be difficult, it can only be avoided at the prior steps if the pertinent decision thresholds are set in ways that may often generate suboptimal liability determinations); 11 Herbert Hovenkamp, Antitrust Law ¶ 1912i (4th ed. 2018) (subparagraph entitled “‘Balancing’ generally to be avoided”); Rebecca Haw Allensworth, *The Commensurability Myth in Antitrust*, 69 Vand. L. Rev. 1 (2015) (arguing that antitrust law suppresses value judgments involved in comparing different types of competitive effects); Hillary Greene, *Muzzling Antitrust: Information Products, Innovation and Free Speech*, 95 B.U. L. Rev. 35 (2015) (advocating balancing despite challenges relating to commensurability).

[31] A burden of persuasion can also be applied at step 1 to find no liability at the close of a plaintiff’s case in a bench trial.

such, they would translate into a plaintiff's need, respectively, to plausibly allege and to have sufficient evidence to create a genuine dispute on the issue. Thus understood, this requirement does not depart from a plaintiff's general need to allege or produce evidence of a prima facie case under a balancing test, so it is unclear how the structured rule of reason differentially economizes on litigation or avoids difficult balancing.[32]

Step 1 has differential oomph only if, in our earlier terminology, H^* is set nontrivially above zero. Some formulations of the structured rule of reason suggest that this is so, notably, in demanding that demonstrated anticompetitive effects be substantial.[33] Then, as explored in section I.A, it is important to know how significant they must be, that is, how high H^* is set. Like in most areas of law, such questions are not given quantitative answers, making the force of step 1 unclear. A demand of significance might merely convey that effects need to be nontrivial, or it could be taken to require much more. Even if the latter, it would only increase early terminations if the translation into plausible allegations or evidence requisite to create a genuine dispute retained significant bite. Considering another dimension, if the requisite significance is contextual, and the context includes some looking ahead to procompetitive effects—the magnitude of B in the case at hand—then we really have relaxed the structured rule, morphing it toward a balancing test. In any case, as explained in subsection I.B.2, it is unclear whether requiring that $H > H^*$ at a motion to dismiss or for summary judgment is more or less stringent than requiring $H > B$.

Step 2 demands that the defendant advance a procompetitive justification (B) for its action. This component of the structured rule of reason is notably more mysterious, beginning with its trigger—that is, the predicate for shifting the burden to the defendant, which appears in virtually all formulations. As suggested by the analysis in subsection I.B.2, such shifting is moot at motions to dismiss and largely so at summary judgment and even at motions for judgment as a matter of law at the close of a plaintiff's case at trial.[34] And neither judge nor jury announces during a trial whether and when such a shift transpires. Therefore, step 2 can really only be operative during a factfinder's deliberations, after it resolves step 1 in the plaintiff's favor. Accordingly, as previously noted, any economization from the structured rule is at most limited to not having to address procompetitive justifications in a final decision. But if none are offered in any substantial manner, such would not be difficult, and if it is a close question it must be addressed in any event. In addition, no difficult balancing is avoided because, under unconstrained balancing, a factfinder who finds that anticompetitive effects have been demonstrated and procompetitive effects are absent will hardly find it difficult to balance the two.

[32] This point would hold even at the very end of a trial, because, even under balancing, a judge may choose not to address a defendant's procompetitive justifications if the plaintiff's prima facie case fails.

[33] *See, e.g.,* ABA ANTITRUST LAW DEVELOPMENTS, *supra* note , at 62 (stating that "[t]he cases reflect the additional consensus that the restraint's anticompetitive effect must be significant to support liability under the rule of reason," but leaving it unclear the extent to which this requirement merely rules out de minimus effects or demands something more); sources quoted *supra* note . Note that a balancing test would impose a similar demand if it was required that a plaintiff show that anticompetitive effects significantly outweighed (rather than merely outweighed) procompetitive effects, which is sometimes stated to be the case.

[34] *See* Kaplow, *supra* note , at 1034–43. Keep in mind the simple point that, even if a judge rules for the plaintiff on a defendant's motion for judgment as a matter of law at the close of the plaintiff's case, this does not mean that the plaintiff has met its persuasion burden, but only that there is a genuine dispute about whether this is so. Furthermore, even if the persuasion burden was announced to be satisfied at that point, the defense is permitted to and ordinarily does offer direct rebuttal; hence, only when the factfinder actually concludes, at the end of trial, that the plaintiff prevails at step 1 can any burden of justification be deemed to have shifted to the defendant.

Suppose now that step 2 is triggered and the defendant does advance procompetitive justifications. Two further puzzles arise. First, in most formulations of the structured rule of reason, the defendant has a mere production burden. As will be noted momentarily, this renders obscure suggestions that a plaintiff must counter with less restrictive alternatives or a demonstration that anticompetitive effects outweigh procompetitive effects, for procompetitive effects have not actually been established. Also, usually omitted from the list is the more straightforward response of directly rebutting the proffered procompetitive justifications and arguing that, in any event, the defendant's proffer is not sufficiently convincing.

Second, the requisite magnitude of the proffered procompetitive justifications is usually not mentioned. Echoing subsection I.A.1's analysis, we can ask whether they must merely be above zero, larger than some B^* (and what is that?), or greater than the H demonstrated by the plaintiff. If the former—and even if they were proved (under a persuasion burden), we would have no idea whether they exceeded H and hence it would not be clear why a plaintiff would then need to respond at all. One might think that, to make sense, a defendant could only purport to *justify* its action by reference to procompetitive effects if the B it advances exceeds H. But what then is the balancing that is required in step 3? And why does everyone say that, when step 2 fails, we have avoided the need to make a difficult balance if step 2 is the very balance we sought to avoid?[35] Note the tension (as well as the obscurity): if the B need not exceed H, then it is unclear how any burden shifts to the plaintiff, but if we require that B exceed H, then we have necessarily balanced. As best one can tell, these questions have been overlooked by all courts and commentators. The step 2 demand on the defendant regarding magnitude—what is here called the level of B^*—is unknown.

Referencing the stylized structured decision procedure presented here makes it obvious that one has not fully stated the rule without stating the thresholds, H^* and B^*. Moreover, as we also saw in subsection I.A.1, as a final decision rule these thresholds are everything. If both are essentially zero, we have unconstrained balancing (even if we do not seem to notice this). And if they are not, various burdens are indeed higher, and, importantly, outcomes will differ—undesirably, on account of the deviations from pure balancing.[36]

Step 3's balancing test is clear enough. The main questions, already noted, concern what it means to have reached step 3. How large of an H was established? (Was a particular conclusion reached, or only that $H > H^*$? If the latter, what was H^*?) Was B merely advanced or proved? Was the plaintiff's rebuttal already considered or not? (I.e., did the factfinder, at the end of the trial, first ignore all the plaintiff's evidence on B, construe all the defendant's evidence

[35] If instead the B required in step 2 is much below H (and keeping in mind as well that the defendant is only said to have a production burden at step 2), how can it be that a difficult balance has been avoided? When the defendant cannot even muster modest evidence of a low B, just how difficult would the balance have been?

[36] To this observer, it appears that the failure to specify the rule with precision has enabled many to simultaneously hold multiple views regarding the structured rule of reason that rest on inconsistent assumptions. Notably, in imagining that steps 1 and 2 avoid many difficult balances, it really must be that H^* and B^* are high and often decisive. But the rules are not described or advocated in this fashion and, if they were, it would be clear (as subsection I.A.1 elaborates) that one would often be recommending incorrect outcomes, which would be particularly stark at step 2 since one would realize, for example, that one was assigning liability when it may well be (or even is known) that $H < B$. In addition, as discussed, it seems common to state step 2 as placing only a production burden on the defendant whereas much discussion that follows in step 3, regarding less restrictive alternatives and balancing, seems to assume that the defendant has proven B. *Compare* ABA ANTITRUST LAW DEVELOPMENTS, *supra* note , at 62 (requiring only a production burden), *and id*. at 77 ("the defendant must *produce* evidence" (emphasis added), *with id*. ("If the defendant *demonstrates* that the restraint produces procompetitive effects, then the plaintiff . . . ." (emphasis added)). And some discussions, notably on less restrictive alternatives, make sense only if B has been proved to exceed H, whereas the remaining need to balance presumes that no such comparison has taken place.

most favorably to it, and conclude that there was a genuine dispute on *B*—that is, that the production burden was met, as the rule asks—and only then take a separate pass at deciding whether it believes *B*? If at a bench trial, would the judge's opinion have separate parts, following the structured rule, or would this aspect of the rule be ignored—even though this is the only point in the proceedings at which it can matter?) Was *B* also quantified? (Above zero? *B**? Or *H*?) Consider as well the corresponding questions on less restrictive alternatives, to which we will turn in a moment.

Taking all of this together, although the meaning of balancing at step 3 is clear, what analysis is actually conducted in step 3 itself or earlier on is not. For the reasons developed in subsection I.B.1 and elaborated below, attempts to sort and separate information collection and assessment in any such prespecified manner are largely counterproductive in any event. It is apparent that, despite decades of statements—by every circuit court and some leading commentators—we have little idea what the structured rule of reason means along a number of basic dimensions.

Consideration of less restrictive alternatives[37] under the rule of reason throws another monkey wrench[38] into the operation.[39] Whether inserted as part of step 2's consideration of *B*, as an additional, intermediate step after step 2 but before step 3's balancing, or as a part of step 3's balancing (but, regardless, generally understood to precede the act of balancing itself), we have the conundrum explained in subsection I.A.2 that the proper methodology for assessing less restrictive alternatives itself requires balancing: one must perform the second balance or, equivalently, the delta/delta test, both of which require knowing *H*, *B*, *H′*, and *B′*—and thus all that is needed to perform the final balance. Therefore, the common understanding that less restrictive alternatives analysis precedes balancing and is a way to avoid the need to balance reflects significant confusion[40] (although sometimes less restrictive alternatives analysis is easy, particularly when the anticompetitive effect results from an essentially naked restraint on competition attached to something that is procompetitive but unrelated).[41] The failure to appreciate the analytical connection between less restrictive alternatives analysis and balancing[42]

[37] Although the language of less restrictive alternatives is commonly used, it is often said instead (or in addition) that the plaintiff can challenge whether the defendant's allegedly anticompetitive conduct is "reasonably necessary" to achieve the procompetitive benefit, suggesting a similar sort of inquiry. *See, e.g.*, sources quoted *supra* note .

[38] This phrasing reflects the focus here on the structured rule of reason. Under the pure balancing version that seems to conform to longstanding Supreme Court precedent, less restrictive alternatives analysis, as difficult as it may be to conduct in practice, does not raise the problems noted here. *See supra* subsection I.A.2.

[39] For more extensive analyses of less restrictive alternatives under antitrust's rule of reason, see Gabriel A. Feldman, *The Misuse of the Less Restrictive Alternative Inquiry in Rule of Reason Analysis*, 58 AM. U. L. REV. 561 (2009), and C. Scott Hemphill, *Less Restrictive Alternatives in Antitrust Law*, 116 COLUM. L. REV. 927 (2016). Prior work has not, however, taken the perspective advanced here that emphasizes how the analysis of less restrictive alternatives fits into and is cast in a different light by other features of the structured rule of reason.

[40] For example, Areeda and Hovenkamp's reluctance to have courts engage in balancing, *see supra* note , leads them to urge, in one of their formulations of a structured rule of reason, that if the balancing step is reached, the court should revisit the previous step on less restrictive alternatives to see if it can get out of this predicament, apparently not appreciating that proper analysis of such alternatives itself requires balancing. *See* AREEDA & HOVENKAMP, *supra* note , at 445 ("Nevertheless, any court faced with the prospect of balancing [in step 6] must go back to step 5 and look hard for workable less restrictive alternatives."); *see also* 13 HERBERT HOVENKAMP, ANTITRUST LAW 55 (3d ed. 2012) ("[T]here is no way that a court can 'balance' the competitive benefits of apparently valuable information exchanges with the magnitude of the competitive threat. First and foremost, the antitrust decision maker must look for less restrictive alternatives.").

[41] This example can be viewed as an application of the ancillary restraint test of *United States v. Addyston Pipe & Steel Co.*, 85 F. 271 (6th Cir. 1898), *aff'd*, 175 U.S. 211 (1899).

[42] It is often noted and sometimes emphasized that less restrictive alternatives analysis does involve balancing, but I have not seen attempts to combine a precise statement of the requisite balancing, how it relates to the basic balancing of

may be attributable to never writing down the simple algebra of what less restrictive alternatives analysis means, a task that was central in framing the exploration in subsection I.A.2.

Additional questions were already noted in connection with step 2: In what sense should a plaintiff be seen as required to advance less restrictive alternatives when the defendant's procompetitive effect (*B*) was not taken to be proved, might be directly rebutted, and in any event was not shown to be greater than the anticompetitive effect (*H*)? It is also sometimes suggested that the proffered alternative must be equally effective in achieving *B*, although many formulations of antitrust's structured rule of reason do not sharply address the matter.[43] It may often be as hard or harder to determine whether the alternative is equally effective (the conceptually wrong question) than whether it warrants liability under the second balancing or delta/delta test (the right question).

Most of the foregoing discussion of the structured rule of reason considers it as a final decision rule because, beyond the plaintiff's step 1 hurdle (which may or may not differ much from that under balancing), the requirements matter only after trial, in reaching a conclusion about whether to assign liability. As noted, this suggests that few economization benefits can be realized. Moreover, subsection I.B.1's analysis of information gathering explains how, if litigation was shaped more in conformity with the structured rule of reason's separate steps rather than guided by unconstrained balancing, most principles of optimal information collection would be sharply violated.

Among them was the point that, often, evidence will bear simultaneously on *H* and on *B* and, indeed, its relevance may be explicitly comparative, in which case even thinking in separate buckets may be counterproductive. This point proves to be powerful in many antitrust settings because characterization is often a central challenge. Moreover, the most common definition of what counts as an exclusionary, anticompetitive practice is one that excludes other than by "competition on the merits."[44] When anticompetitive effects (*H*) are understood as ones that operate other than via procompetitive channels (*B*), the notion that one would either gather or process evidence in a sequentially siloed fashion, reaching a conclusion about *H* before even considering what *B* might be about, is counterproductive if not incoherent.[45] Another major

anti- and procompetitive effects required in step 3, and how that interaction renders bizarre any sense that less restrictive alternatives analysis can be separate from and prior to the core balancing test of the rule of reason.

[43] *See, e.g.*, sources quoted *supra* note .

[44] *See, e.g.*, U.S. DEP'T OF JUSTICE, *supra* note , at 166 n.6 ("All jurisdictions agree that unilateral conduct laws address specific conduct and its anticompetitive effects, rather than the mere possession of dominance/substantial market power or its creation through competition on the merits."); MODEL JURY INSTRUCTIONS, *supra* note , at C-26 to C-27; 3B PHILLIP E. AREEDA & HERBERT HOVENKAMP, ANTITRUST LAW 423 (4th ed. 2015); *Guidance on Article 82*, *supra* note , ¶ 1 ("Article 82 . . . prohibits abuses of a dominant position. In accordance with the case-law, it is not in itself illegal for an undertaking to be in a dominant position and such a dominant undertaking is entitled to compete on the merits. However, the undertaking concerned has a special responsibility not to allow its conduct to impair genuine undistorted competition on the common market."). Similarly, the test for monopolization under Section 2 defines the second element as "the willful acquisition or maintenance of [monopoly] power *as distinguished from* growth or development as a consequence of a superior product, business acumen, or historic accident." United States v. Grinnell Corp., 384 U.S. 563, 570–71 (1966) (emphasis added).

[45] As but one example that often arises in antitrust cases, consider the situation in which there is a dispute about the meaning of some of the defendant's internal documents, which the plaintiff claims indicate a motive to disadvantage competitors and thus injure competition and the defendant argues show a plan to outperform rivals, which promotes competition. Such disputes about interpretation, which depend on context, can only be understood by reference to the alternative hypotheses that each side advances. Similar analysis is applicable to inquiries into intent for the purposes of illuminating effects. *See, e.g.*, Chi. Bd. of Trade v. United States, 246 U.S. 231, 238 (1918) ("The history of the restraint, the evil believed to exist, the reason for adopting the particular remedy, the purpose or end sought to be attained, are all relevant facts. This is not because a good intention will save an otherwise objectionable regulation, or the reverse, but because knowledge of intent may help the court to interpret facts and to predict consequences.").

consideration is market power, central in many antitrust inquiries but often relevant precisely because of how it illuminates the relative plausibility of anti- and procompetitive explanations.[46] Also important are some of the other principles considered in subsection I.B.1, such as that some of the high diagnosticity/cost evidence may bear particularly on *B*[47] and that having some sense of the likely magnitude of *B* importantly feeds back on how demanding we should be with regard to *H*.

Law.—Consider three further angles that pertain to the state of the law. The first concerns the sense in which the structured rule of reason under discussion correctly states the doctrine. Despite some form of the structured rule being articulated by every circuit court (and leading commentators),[48] these statements are usually dicta that is rather far removed from the decision at hand. Most of these decisions regard motions to dismiss and for summary judgment, where the question is whether the plaintiff has alleged or offered evidence to create a genuine dispute about some level of anticompetitive effects—step 1, period. (This may help to explain both the variation in formulations across cases and internal inconsistencies.[49]) Recall as well that, under a pure balancing test, the plaintiff must likewise allege plausible anticompetitive harm to survive a motion to dismiss or present sufficient evidence thereof to create a genuine dispute at summary judgment.

Moreover, some of these statements of structured decision protocols are in any event followed by possibly superseding language that presents an unconstrained balancing test, often quoting or at least citing the Supreme Court. Indeed, clear Supreme Court precedent spanning a century rather clearly presents the rule of reason using *Chicago Board*'s famous statement: "The true test of legality is whether the restraint imposed is such as merely regulates and perhaps thereby promotes competition or whether it is such as may suppress or even destroy competition."[50] As subsection I.A.1 explains, a structured rule of reason with real bite (one with

[46] Indeed, another central shortcoming in the antitrust setting is that, in addition to the sequential siloing of anti- and procompetitive explanations, market power analysis is in turn separated from (and precedes) the analysis of either *H* or *B*. *See* Louis Kaplow, *Market Power*, *supra* note .

[47] For example, in analyzing a joint venture that includes arguably ancillary restrictions on competition between the venturers, it may often make sense early on to consider the nature of the *B* that the venture may generate. Determining whether the competitive restrictions bear a strong nexus with the procompetitive benefits typically requires an appreciation of what those benefits are.

[48] Areeda and Hovenkamp argue that, under one of their proffered formulations for a structured rule of reason, "the staged inquiry is particularly conducive to summary judgment motions or motions on the pleadings." AREEDA & HOVENKAMP, *supra* note , at 445. How this is true beyond their preliminary steps that address the plaintiff's allegations, however, is mysterious. Their next step declares the defendant's action illegal absent strong evidence of efficiencies, *see id*. at 444, but this, as usual, ignores that a factfinder would have to have concluded that the plaintiff's allegations of harm were true, not just that its allegations were plausible or that there was sufficient evidence to create a genuine dispute.

[49] Regarding the latter, it is commonly stated that step 2 places only a production burden on the defendant and, moreover, it is not even stated that the defendant must put forth a *B* as large as *H*, yet subsequent language on less restrictive alternatives or regarding step 3 seems to suggest that the plaintiff (having actually proved *H*) will necessarily lose if it does not meet these further obligations. *See supra* note .

[50] Chi. Bd. of Trade v. United States, 246 U.S. 231, 238 (1918). Although this statement is a century old, it is routinely cited and quoted by the Supreme Court as authoritative, and modern formulations are much more consistent with it than with a structured rule. *See, e.g.*, Am. Needle, Inc. v. Nat'l Football League, 560 U.S. 183, 203 n.10 (1993) (referring to Brandeis's "classic formulation" of the rule of reason in *Chicago Board*); Cal. Dental Ass'n v. FTC, 526 U.S. 756, 771 (1999) ("it seems to us that the CDA's advertising restrictions might plausibly be thought to have a net procompetitive effect, or possibly no effect at all on competition"); *id*. at 774 (" it does not obviously follow that such a ban would have a net anticompetitive effect here"); Eastman Kodak Co. v. Image Tech. Servs., Inc., 504 U.S. 451, 479 (1992) ("We need not decide whether Kodak's behavior has any procompetitive effects and, if so, whether they outweigh the anticompetitive effects. We note only that Kodak's service and parts policy is simply not one that appears always or almost always to enhance competition, and therefore to warrant a legal presumption without any evidence of its actual economic impact. In this case, when we weigh the risk of deterring procompetitive behavior by proceeding to trial against

an *H** or a *B** nontrivially above zero) would change outcomes from those under balancing, which has been consistently commanded by the Supreme Court. Accordingly, some skepticism seems applicable in attributing weight to seemingly contrary dicta, which itself is often accompanied by citations and quotations of the Supreme Court's articulation of the balancing test.[51] Perhaps not surprisingly, therefore, the U.S. antitrust enforcement agencies seem to follow the Supreme Court's open-ended formulation rather than the structured rule of reason.[52] On the other hand, once similar dicta appears consistently in circuit court decisions and is asserted by some leading commentators to state the law, it can readily take on a life of its own, even if it is contrary to a century of Supreme Court precedent and contravenes basic principles of decisionmaking and information collection. Indeed, the Court's recent decision in *American Express* may reflect and reinforce this transformation.[53]

the risk that illegal behavior will go unpunished, the balance tips against summary judgment."); *id*. at 486 ("It may be that its parts, service, and equipment are components of one unified market, or that the equipment market does discipline the aftermarkets so that all three are priced competitively overall, or that any anticompetitive effects of Kodak's behavior are outweighed by its competitive effects."); NCAA v. Bd. of Regents of Univ. of Okla., 468 U.S. 85, 104 (1984) ("But whether the ultimate finding is the product of a presumption or actual market analysis, the essential inquiry remains the same—whether or not the challenged restraint enhances competition. Under the Sherman Act the criterion to be used in judging the validity of a restraint on trade is its impact on competition." (footnote omitted)); Nat'l Soc'y of Prof'l Eng's v. United States, 435 U.S. 679, 691 (1978) ("From Mr. Justice Brandeis' opinion for the Court in *Chicago Board of Trade*, to the Court opinion written by Mr. Justice Powell in *Continental T. V., Inc*., the Court has adhered to the position that the inquiry mandated by the Rule of Reason is whether the challenged agreement is one that promotes competition or one that suppresses competition."). Note further that none the leading cases associated with truncated rule of reason analysis, cited in note —one after a full trial and two after proceedings at the FTC—indicate that anything like the structured rule of reason has refined or supplanted *Chicago Board*'s balancing test.

[51] *See, e.g.*, Capital Imaging Assocs., P.C. v. Mohawk Valley Med. Assocs., Inc., 996 F.2d 537, 543 (2d. Cir.), *cert. denied*, 510 U.S. 947 (1993) (concluding its presentation of what appears to be a structured rule of reason with the statement: "Ultimately, it remains for the factfinder to weigh the harms and benefits of the challenged behavior. The classic articulation of how the rule of reason analysis should be undertaken is found in [*Chicago Board*], where Justice Brandeis speaking for the Supreme Court said . . . . It at least seems clear that the factfinder must decide the overarching question of whether the challenged action purports to promote or to destroy competition.").

[52] *See* U.S. DEP'T OF JUSTICE & FED. TRADE COMM'N, ANTITRUST GUIDELINES FOR COLLABORATIONS AMONG COMPETITORS 10 (2000) [hereinafter U.S. COLLABORATIONS GUIDELINES] ("Rule of reason analysis entails a flexible inquiry and varies in focus and detail depending on the nature of the agreement and market circumstances."); *id*. ("Under the rule of reason, the Agencies' analysis begins with an examination of the nature of the relevant agreement, since the nature of the agreement determines the types of anticompetitive harms that may be of concern. *As part of this examination, the Agencies ask about the business purpose of the agreement* and examine whether the agreement, if already in operation, has caused anticompetitive harm." (citing *Chicago Board*) (emphasis added)). Other parts of these guidelines, however, do indicate some inclination to sequence the analysis. *See, e.g.*, *id*. at 12 ("The agencies do not undertake a full analysis of procompetitive benefits pursuant to Section 3.36 below, however, unless an anticompetitive harm appears likely.").

[53] *See* Ohio v. Am. Express Co., 138 S. Ct. 2274 (2018). A number of observations are in order: (1) Both the majority and dissent explicitly stated that both parties had agreed that the structured rule governed the case. *See id*. at 2284; *id*. at 2290 (Breyer, J., dissenting). (The majority, before noting the parties' agreement on this framework, offers its own characterization of the rule of reason: "The goal is to 'distinguish[] between restraints with anticompetitive effect that are harmful to the consumer and restraints stimulating competition that are in the consumer's best interest.'" *Id*. at 2284 (quoting *Leegin Creative Leather Products, Inc. v. PSKS, Inc.*, 551 U.S. 877, 886 (2007)).) (2) The formulations offered in the two opinions were, variously, subject to the various infirmities noted previously in this section (and in detail in various footnotes), inconsistent with each other in multiple ways, inconsistent with subsequent use in the opinions, and inconsistent with the authorities cited (by the majority) for the structured rule. (3) Much of the disagreement between the majority and dissent—giving a "two ships passing in the night" flavor to the opinions—relates to the majority invoking justifications for the restraint as part of step 1, which the dissent argued were properly assessed at later steps. *See id*. at 2303 (Breyer, J., dissenting) ("But the Court of Appeals would properly consider procompetitive justifications not at step 1, but at steps 2 and 3 of the 'rule of reason' inquiry. . . . The majority charts a different path. Notwithstanding its purported acceptance of the three-step, burden-shifting framework I have described . . . , the majority addresses American Express' procompetitive justifications now, at step 1 of the analysis . . . ."). Both views are puzzling since there had been

A second point concerns another aspect of the law in action that also relates to the procedural posture of decisions that purport to take place under the auspices of a structured rule of reason. Let's call this the "balancing myth" myth: more precisely, the confusion behind the commonly advanced claim that balancing under the rule of reason is a myth. The core rationale for believing that balancing is rare is the dearth of reported cases in which balancing occurs.[54] As mentioned, most antitrust opinions (just as in many areas of law) are on motions to dismiss and at summary judgment. Basic teachings of civil procedure make it clear that balancing cannot occur at these stages. Under unconstrained balancing, one would have had to resolve factual disputes regarding both *H* and *B* in order to compare them, and this can only be done at the end of a trial. Likewise under a structured rule of reason at step 3. By contrast, these procedural decisions are by nature confined to step 1.

Now, it may or may not be true that many cases would ultimately involve balancing (under either approach) if they reached the end of trial and produced complete opinions at that point. For those in which defendants indeed win on dispositive motions, this would not have occurred. For the rest—in which such motions are not filed or such motions are denied—we know that most cases settle. And for those that reach the end of trial, rule of reason cases are often tried to a jury and hence do not produce written opinions. A judicial opinion properly balancing anti- and procompetitive effects can only arise in the tiny—and probably quite unrepresentative—sample of completed bench trials that generate opinions.[55] Yet commentators' beliefs—and the leading studies of the question—concentrate on *dispositive* court opinions, which are almost entirely granted motions and hence, by nature, cannot have involved balancing; even worse, any motions that were denied are excluded from the sample,[56] including

a full trial covering all of the issues, rendering it unclear why it mattered what factors were part of which steps. (Relatedly, some of the dispute on whether the proper market definition encompassed both sides of the market or only the merchants' side concerned whether various of the defendant's arguments, referring to the cardholders' side, had to be considered as part of step 1 rather than deferred to later. If *American Express*'s statements under the rubric of market definition are tantamount to requiring procompetitive features of a restraint to be assessed at step 1, the central holding might be restated as rejecting the essence of the structured rule of reason that the majority purported to follow.) To this reader, attempts to shoehorn arguments and facts into the structured protocol confused rather than clarified the Court's analysis.

[54] *See, e.g.*, Hemphill, *supra* note , at 951 ("Although it is commonplace to understand the rule of reason as a fact-intensive search for net effects, cases are seldom decided on that explicit basis. An extensive survey of rule of reason final judgments [referring to the articles by Carrier discussed in note] concluded that very few are decided on net-effect balancing grounds. . . . Careful observers have gone so far as to declare that explicit balancing is a 'myth.'" (quoting Gavil, *supra* note , at 147) (footnotes omitted)); Allensworth, *supra* note , at 47–48 (similarly discussing Gavil and Carrier); *see also* U.S. DEP'T OF JUSTICE, *supra* note , at 38 ("[S]everal panelists and commentators have pointed out that, in practice, courts do not engage in the precise balancing called for by the effects-balancing test.").

[55] Even then, a judge may well reach a conclusion through explicit or implicit balancing but craft an opinion that suggests that the outweighed factor (say, the anticompetitive effect) was not established. *See, e.g.*, Benjamin Klein, *in* HEARINGS, U.S. DEP'T OF JUSTICE & FED. TRADE COMM'N, SINGLE-FIRM CONDUCT AND ANTITRUST LAW 201 (Nov. 15, 2006) ("I mean, I think [the courts] go backwards, and they figure out—you know, they do some kind of implicit balancing, and then they say—they make it easy and they say it was not an anticompetitive effect or there is no procompetitive efficiency rationale . . . ."); William J. Kolasky, *in* HEARINGS, *supra*, at 60 (May 1, 2007) ("But, in fact, when you look at the decisions, the courts never reach that final balancing stage, because they obviate the need for that by adjusting the degree of scrutiny that they engage in with respect to steps two and three [regarding procompetitive justifications and less restrictive alternatives], depending on how strong a showing the plaintiff makes in step one [regarding anticompetitive effects], an inquiry meet for the case . . . ."); Allensworth, *supra* note , at 48–50; Hemphill, *supra* note , at 951. That is, the extent of actual balancing can exceed what is explicit in written opinions.

[56] Notably, Michael Carrier's two studies, Michael A. Carrier, *The Real Rule of Reason: Bridging the Disconnect*, 1999 BYU L. REV. 1265 [hereinafter, Carrier, *First Study*], and Michael Carrier, *The Rule of Reason: An Empirical Update for the 21st Century*, 16 GEO. MASON L. REV. 827 (2009) [hereinafter, Carrier, *Second Study*], explicitly omit from the sample all motions that are denied. *See, e.g.*, Carrier, *First Study*, *supra*, at 1270 n.13. He also fails to report separately on the (presumably quite small) portion of his sample of dispositive opinions that are on the merits after a bench

motions that are denied because balancing seems to be required.[57] (It is like concluding that rain in the United States is rare based on an examination of measurements taken only at weather stations located in desserts.)

Third—and on a different tack—under the auspices of Section 1's rule of reason the Supreme Court has, over time, made some important refinements. Some types of acts—most importantly, horizontal price fixing—are per se illegal,[58] and some others may be subject to a more truncated or "quick look" analysis[59] under which anticompetitive effects may be presumed unless defendants are able to come forward with plausible procompetitive effects, which may then trigger a more complete (and more standard) inquiry.[60] The core rationale is that, with regard to categories of activity that are nearly always anticompetitive overall, it saves resources simply to deem the act illegal or to presume illegality but allow rebuttal. Such rules and

trial. (Carrier also drops difficult to categorize cases from his sample, which is worrisome because murkier cases may be precisely those in which competing effects are alive and well, thus requiring balancing.)

Suppose, for example, that 400 cases enter the legal system. From there, 200 are subjected to dispositive motions of which 100 are granted. 275 (including many of the 100 denied motions) leave the system via settlements. Of course, many that survived motions (particularly for summary judgment) may have been cases that were thought likely to require balancing, and many cases may not have had motions (or a judge may have deferred deciding them) because there was a serious contest, suggesting the possibility that balancing would have been required. Now, of the remaining 25, suppose that 20 are tried to a jury and 5 to a judge, the latter producing 5 opinions, 3 with balancing. By Carrier's method, only 3 of our 105 opinions that are dispositive—just under 3%—would involve balancing. The conclusion reached (and accepted by others) is that balancing is rare, a myth.

Compare this hypothetical to the introductory summary that Carrier offers in his 2009 update of his original study:

> In the first stage, the plaintiff must show a significant anticompetitive effect. The plaintiff's failure to make such a showing led to the courts' dismissal [apparently counting granted summary judgments as "dismissals" and possibly some opinions at the end of trial] of 84% of the cases. In the second stage, the defendant must demonstrate a legitimate procompetitive justification; its failure to do so led to invalidation of the restraint in 3% of the cases. [A reader wonders if these are all at the end of trial, or some were on plaintiffs' motions for summary judgment, with the defendant failing to successfully oppose regarding step 1 and also failing in its proffer on step 2.] If the defendant satisfies this burden, the plaintiff can show that the restraint is not reasonably necessary or that the defendant's objectives could be achieved by less restrictive alternatives. At most, 1% of the cases were dismissed because the plaintiff made this showing [presumably "dismissed" here refers both to the end of a trial and to a finding of liability]. Only after the completion of these three stages does the court balance anticompetitive and procompetitive effects. Balancing occurred in 4% of the cases [with no indication of what percentage they were of the presumably few cases yielding opinions at the end of trials].

Carrier, *Second Study*, *supra*, at 827.

[57] The latter would be atypical because, as explained earlier in this section, even motions for summary judgment and for judgment as a matter of law most often will focus on step 1. Hence, *B* will not be on the table, making balancing even further removed. Moreover, under unconstrained balancing, it is likewise true that balancing can only be done at the end of a trial.

[58] *See, e.g.*, United States v. Socony-Vacuum Oil Co., 310 U.S. 150, 223 (1940).

[59] The initial mode of analysis is often determined by categorical rules, a type considered briefly in the application to constitutional law in Part IV. On the optimal design of rules of this type in general, see Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557 (1992), and for different views on appropriate categorization in this antitrust application, see C. Frederick Beckner III & Steven C. Salop, *Decision Theory and Antitrust Rules*, 67 ANTITRUST L.J. 41 (1999); Arndt Christiansen & Wolfgang Kerber, *Competition Policy with Optimally Differentiated Rules Instead of "Per Se Rules vs Rule of Reason,"* 2 J. COMPETITION L. & ECON. 215 (2006); Daniel A. Crane, *Rules Versus Standards in Antitrust Adjudication*, 64 WASH. & LEE L. REV. 49 (2007); Yannis Katsoulacos & David Ulph, *On Optimal Legal Standards for Competition Policy: A General Welfare-Based Analysis*, 57 J. INDUS. ECON. 410 (2009); and Mark A. Lemley & Christopher R. Leslie, *Categorical Analysis in Antitrust Jurisprudence*, 93 IOWA L. REV. 1207 (2008).

[60] *See* NCAA v. Bd. of Regents of Univ. of Okla., 468 U.S. 85 (1984); FTC v. Ind. Fed'n of Dentists, 476 U.S. 447 (1986); *see also* Cal. Dental Ass'n v. FTC, 526 U.S. 756, 780–81 (1999) ("As the circumstances here demonstrate, there is generally no categorical line to be drawn between restraints that give rise to an intuitively obvious inference of anticompetitive effect and those that call for more detailed treatment. What is required, rather, is an enquiry meet for the case, looking to the circumstances, details, and logic of a restraint.").

reasoning fit well with both balancing and the structured rule of reason. In either case, if there is a basis for presuming both the presence of *H* and the absence of *B*, liability would follow.

An interesting feature of these rules for present purposes is that, if a presumption of liability (however strong) is to be overcome, the primary basis for doing so involves an assessment of whether the procompetitive benefit (*B*) might be significant rather than negligible or nonexistent.[61] Under unconstrained balancing, subsection I.B.1 instructs that one should examine information in whatever order makes sense in light of what is currently known and the diagnosticity/cost ratio of further clumps of information that one might examine. In a realm where preliminary indications are that *H* is significant and *B* is likely to be zero or nearly so, expecting the defendant to offer something substantial on *B* to warrant further analysis would often appear sensible. Relatedly, merely questioning *H* would seem insufficient to change the decision in most cases given a strong presumption that *H* is present, as long as no *B* is visible on the horizon. It is also notable that, if a significant *B* does appear in the offing, the underlying basis for *H* is then open to reconsideration.[62] Thus, it appears that courts, when deciding whether a practice should be deemed illegal with only a limited inquiry, do appropriately take a more flexible approach regarding the order in which information on *H* and on *B* is considered.[63] A broader lesson of subsection I.B.1 is that such flexibility is often quite valuable, but tends to be

[61] *See, e.g.*, U.S. COLLABORATIONS GUIDELINES, *supra* note , at 10–11 ("Alternatively, where the likelihood of anticompetitive harm is evident from the nature of the agreement, or anticompetitive harm has resulted from an agreement already in operation, then, absent overriding benefits that could offset the anticompetitive harm, the Agencies challenge such agreements without a detailed market analysis." (footnotes omitted)); Timothy J. Muris, *The New Rule of Reason*, 57 ANTITRUST L.J. 859, 861 (1989) ("The first issue to address is whether the conduct is inherently suspect. . . . If the conduct is inherently suspect, then the issue becomes the existence of efficiency justifications."). The nature of the defendant's burden regarding *B* under various formulations is not entirely clear. It matters, of course, how large a *B* must be advanced (perhaps compared to the presumed or typical *H*, the actual *H* not having been assessed) and how strong the demonstration thereof must be (plausibly alleged? meeting a production burden?) to trigger a full(er) rule of reason analysis under a quick look. Part of the difficulty in this regard reflects the heterogeneity of practices. Sometimes procompetitive justifications may be fairly apparent given the nature of a restraint, but other times it may require significant evidence to overcome the presumption concerning an absence of much *B*.

[62] On one hand, to the extent that a strong initial presumption, supposedly drawn from experience, simply vanishes once a prospective *B* is advanced with some force seems hard to justify, although one should consider evidence bearing on *H* in the case at hand in light of that preexisting knowledge regarding what *H* is normally like in such cases. There is an important, particular reason that stronger belief in a significant *B* may erode confidence in *H*, concerning the aforementioned frequent interdependency between the two. If a practice usually causes only *H* and no *B*, then the absence of any evidence establishing *B* reinforces the presumption about *H*, whereas if we are convinced that a defendant's action is profitable to it because of the large *B* that it generates, we are accordingly less suspicious that the reason for its action must have been *H*, here, the anticompetitive effects. This reasoning underlies part of the Supreme Court's rationale in *Cal. Dental*. *See* 526 U.S. at 771–73 ("The case before us, however, fails to present a situation in which the likelihood of anticompetitive effects is comparably obvious. . . . Whereas [Justice Breyer in his dissent] accepts, as the Ninth Circuit seems to have done, that the restrictions here were like restrictions on advertisement of price and quality generally . . . , it seems to us that the CDA's advertising restrictions might plausibly be thought to have a net procompetitive effect, or possibly no effect at all on competition. The restrictions on both discount and nondiscount advertising are, at least on their face, designed to avoid false or deceptive advertising in a market characterized by striking disparities between the information available to the professional and the patient. . . . The existence of such significant challenges to informed decisionmaking by the customer for professional services immediately suggests that advertising restrictions arguably protecting patients from misleading or irrelevant advertising call for more than cursory treatment as obviously comparable to classic horizontal agreements to limit output or price competition." (footnotes omitted)).

[63] Another dimension of flexibility concerns how much inquiry is required before concluding that a per se or quick look mode of analysis is appropriate, in lieu of a full rule of reason assessment. *See, e.g.*, *NCAA*, 468 U.S. at 104 n.26 ("Per se rules may require considerable inquiry into market conditions before the evidence justifies a presumption of anticompetitive conduct.")

stifled if structured information protocols guide the conduct of litigation (which, as subsection I.B.2 explains, they largely do not).[64]

Before leaving this subject, it must be noted that much of what is described here, and in the next section, reflects understandable challenges in resolving antitrust cases that often involve highly complex factual disputes, including battles of experts, often regarding practices and industries with which the courts have little experience. Moreover, a federal district judge will often have never heard such a case before, so the task of deciding motions, managing discovery, conducting a trial, and reaching a decision (if it is a bench trial) will be daunting. And obviously the challenges facing lay juries are even larger, which in turn motivates greater judicial control that itself is fraught. The comparative advantage of specialized agencies and possibilities for reform of the litigation process seem particularly apposite.[65]

B. *Mergers*

A significant domain of antitrust law involves the review of horizontal mergers, which are governed by a similar framework in much of the world.[66] On its face, the law states a straightforward balancing test, asking whether a proposed merger would substantially lessen competition (focusing in this section on the formulation in the United States).[67] As implemented by antitrust agencies and courts, however, the law in action can usefully be understood, at least in some respects, as a structured decision procedure. In particular, both agency guidelines and court opinions examine *H* and *B* sequentially, and it often appears that this is more than merely for the convenience of presentation in that the consideration of *H* seems to be undertaken without regard to what might be thought about the level of *B* in a particular case.

More specifically, one can view merger analysis as adhering to the sort of three-step regimen outlined in subsection I.A.1. Regarding step 1, even though horizontal mergers

[64] A central question under these related doctrines concerns the procedural stage and the structuring of litigation. For example, if it is only determined at the end of trial whether the defendant's evidence on *B* is sufficient to warrant a full rule of reason analysis, does the plaintiff then lose unless, from the complaint through discovery and expert reports and trial, it had undertaken to do everything necessary to prevail under a full rule of reason? Or do we then start again, beginning with an amended complaint or, if the allegations were present in the alternative (which they often are), with a new round of discovery and so forth? Some commentators have recognized this problem and suggested a form of bifurcation to address it. *See, e.g.*, HOVENKAMP, *supra* note , ¶ 1914d2.

[65] *See, e.g.* SECTION OF ANTITRUST LAW, AM. BAR ASS'N, PRESIDENTIAL TRANSITION REPORT: THE STATE OF ANTITRUST ENFORCEMENT 17–19 (Jan. 2017) (suggesting that the enforcement agencies consider innovation in the conduct of adjudication in federal court, such as through restructuring litigation and using court-appointed magistrates and experts); F.M. Scherer, *Making the Rule of Reason Analysis More Manageable*, 56 ANTITRUST L.J. 229 (1987) (calling for the restructuring of litigation, such as through limitations on discovery and the use of preliminary expert reports to narrow issues).

[66] *See, e.g.*, U.S. DEP'T OF JUSTICE & FED. TRADE COMM'N, HORIZONTAL MERGER GUIDELINES (2010) [hereinafter U.S. MERGER GUIDELINES]; *Guidelines on the Assessment of Horizontal Mergers Under the Council Regulation on the Control of Concentrations Between Undertakings*, 2004 O.J. (C 31) 5.

[67] Clayton Act Section 7 refers to cases in which "the effect of such acquisition may be substantially to lessen competition, or to tend to create a monopoly." 15 U.S.C. § 18. Sherman Act Section 1, 15 U.S.C. § 1, also formally covers mergers and does not contain a specific demand that effects be "substantial." *See supra* section A (discussing Sherman Act Section 1, where the rule of reason is the governing principle and varying specifications of a structured version may but do not always mention that anticompetitive effects need to be substantial or that they substantially outweigh procompetitive ones). Language in the U.S. Merger Guidelines is inconsistent as to whether there is a substantiality requirement. *Compare* U.S. MERGER GUIDELINES, *supra* note , at 1 ("The Agencies seek to identify and challenge competitively harmful mergers while avoiding unnecessary interference with mergers that are either competitively beneficial or *neutral*." (emphasis added)), *with id.* at 2 ("The Agencies consider any reasonably available and reliable evidence to address the central question of whether a merger may *substantially* lessen competition." (emphasis added)).

generally create at least some incentive to raise prices by eliminating the direct competition between the merging parties, most mergers are not challenged because the predicted price increase is not regarded to be large enough.[68] That is, even though $H > 0$, $H \leq H^*$. One of the rationales for this hurdle is that most mergers are regarded to generate at least some efficiency benefits.[69]

Yet at step 2 efficiency defenses are rarely deemed adequate. This, in turn, suggests that, at least de facto, there is a B^* that substantially exceeds zero. Finally, at step 3, which is only reached if there are cognizable efficiencies, a balancing test is conducted to determine whether to permit or prohibit[70] the merger.[71] Even when step 3 is not stated as distinct from step 2, as long as there is a significant threshold, B^*, we have in essence a three-step protocol (depending on the manner in which B^* is set, as elaborated below).

This section will examine the basis for setting H^* and B^* as high as they seem to be and the apparent sequential siloing of the analysis of anticompetitive and procompetitive effects. As we should expect from the analysis in Part I, these questions are intimately related from the perspective of optimal system design.

More recently, an explicit rationale for step 1 of the current structure (which has been present much longer) has been offered. Because many proposed mergers are filed and they are generally regarded to generate nontrivial efficiencies—but these efficiencies are thought to be hard to quantify in individual cases[72]—it is suggested to be convenient to essentially assign prospective mergers an "efficiency credit," that is, a sort of presumptive B.[73] Once that is done,

[68] For example, from 2003–2012, among mergers sizeable enough to require reporting in the United States, second requests were issued in 3.1% of the cases and 60% of those generated some form of opposition. *See* JOHN KWOKA, MERGERS, MERGER CONTROL, AND REMEDIES: A RETROSPECTIVE ANALYSIS OF U.S. POLICY 9–10 (2015).

[69] *See, e.g.*, U.S. MERGER GUIDELINES, *supra* note , at 29 ("[A] primary benefit of mergers to the economy is their potential to generate significant efficiencies and thus enhance the merged firm's ability and incentive to compete, which may result in lower prices, improved quality, enhanced service, or new products."). Relatedly, if efficiency benefits are at least in part passed on to customers, then the net effect of the merger may not be to increase price.

[70] "Prohibit"—a term used loosely in the text—actually means, in many jurisdictions including the United States, for an agency, a decision to seek an injunction in court to stop the merger or, for the court, to grant the injunction.

[71] The common denominator is often taken to be consumer welfare. *See, e.g.*, U.S. MERGER GUIDELINES, *supra* note , at 2 ("Regardless of how enhanced market power likely would be manifested, the Agencies normally evaluate mergers based on their impact on customers."); *id*. at 30–31 ("The Agencies will not challenge a merger if cognizable efficiencies are of a character and magnitude such that the merger is not likely to be anticompetitive in any relevant market. To make the requisite determination, the Agencies consider whether cognizable efficiencies likely would be sufficient to reverse the merger's potential to harm customers in the relevant market, e.g., by preventing price increases in that market." (footnotes omitted)).

[72] *See, e.g.*, Joseph Farrell & Carl Shapiro, *Antitrust Evaluation of Horizontal Mergers: An Economic Alternative to Market Definition*, B.E. J. THEORETICAL ECON., Jan. 2010, art. 9, 1, 10 ("[M[erger-specific efficiencies are often very hard to predict, even for the firms themselves but especially for antitrust agencies and courts . . . ."); Dennis A. Yao & Thomas N. Dahdouh, *Information Problems in Merger Decision Making and Their Impact on Development of an Efficiencies Defense*, 62 ANTITRUST L.J. 23 (1993). For work addressing the incorporation of efficiencies into merger analysis, see, for example, William J. Kolasky & Andrew R. Dick, *The Merger Guidelines and the Integration of Efficiencies into Antitrust Review of Horizontal Mergers*, 71 ANTITRUST L.J. 207 (2003); Timothy J. Muris, *The Government and Merger Efficiencies: Still Hostile After All These Years*, 7 GEO. MASON L. REV.729 (1999); Robert Pitofsky, *Efficiencies in Defense of Mergers: Two Years After*, 7 GEO. MASON L. REV. 485 (1999).

[73] This idea was articulated in Frederick R. Warren-Boulton, *Merger Policy and Enforcement at the Antitrust Division: The Economist's View*, 54 ANTITRUST L.J. 109, 112 (1985) ("I should preface this discussion by saying that the very existence of 'safe harbor' Herfindahls in the *Guidelines* already implies a 'standard deduction' for efficiencies. Such a standard deduction is implicit in a policy that allows mergers that increase concentration to some extent, even without a showing of any efficiency gains. Alternatively, the parties can choose to itemize efficiencies, rather than just take the standard deduction, by presenting an explicit efficiency defense."). Much later, it was elaborated in Louis Kaplow & Carl Shapiro, *Antitrust*, *in* 2 HANDBOOK OF LAW AND ECONOMICS 1073, 1162–69 (A. Mitchell Polinsky & Steven Shavell eds., 2007), and applied, notably, in Farrell & Shapiro, *supra* note ; *see also* Joseph Farrell & Carl Shapiro, *Upward Pricing*

the step 1 examination of H considers not whether it exceeds zero but whether it exceeds the level of that credit. If we call that credit H^*, the step 1 test becomes $H > H^*$.

Under this rationale, if a case passes step 1 it would not make sense in step 2 to ask merely whether $B > 0$, because the merging parties have already been credited with a B that substantially exceeds zero. Hence, we instead ask whether $B > B^*$. Moreover, given the stated rationale, it might seem that B^* should equal H^*. Or one might set B^* somewhat above H^* on the ground that, unless efficiencies are nontrivially greater than the credit, it is not worth the effort to assess them.[74]

Note that, even though efficiencies are regarded to be important and commonplace, the implication of the efficiency credit at step 1 is that step 2 will often fail to find efficiencies because the search is for above-average efficiencies, not just any efficiencies. However, this observation alone seems insufficient to explain that efficiencies are rarely said to be found (even though routinely proffered by the merging parties). After all, if B^* measures, say, average efficiencies, one might have thought that we would have $B > B^*$ in roughly half the cases.[75] Furthermore, in order to avoid too many false negatives—failures to prohibit mergers for which $H > B$—it might be thought optimal to set H^*, and thus B^*, at a below-average level, in which case step 2 findings of $B > B^*$ would be even more common.

In any event, the appropriate magnitude of what is referred to here as B^* has not been as fully elaborated as the general notion of an efficiency credit, which focuses on step 1. Regarding the optimal setting of B^*, the analysis in subsection I.A.1, focusing on structured decision procedures as final decision rules, suggests that equating B^* and H^* has some appeal. But it was seen to be even more sensible to set B^* equal to the H determined in step 1, which converts step 2 into the ultimate balancing test.[76]

Pressure and Critical Loss Analysis: Response, THE CPI ANTITRUST JOURNAL, Feb. 2010, at 1, 5-6 (further elaborating the ideas in a spirit suggestive of some of the analysis that follows here). Due to the latter writings (and reflecting in part that Joseph Farrell and Carl Shapiro were, respectively, the chief economists at the FTC and the Antitrust Division of the Department of Justice at the time of the 2010 revisions to the U.S. Merger Guidelines), this way of thinking about the issue has become more widespread. *See, e.g.*, Hovenkamp, *supra* note , at 379–81. Nevertheless, it remains an unofficial rationale, not articulated as such in the cases or the Guidelines as a central explanation for why such a high demand is placed on the demonstration of H even though horizontal mergers, particularly larger ones (most of which are not prohibited), generate at least some nontrivial unilateral incentive to increase prices due to diminished competitive pressure (as well as sometimes making coordinated price elevation more likely). More relevant for present purposes, prior examination of the efficiency credit idea has not fully elaborated its connection to the underlying structured decision procedure nor related it to the analysis of optimal information collection, which are the focuses here.

[74] For elaboration, see note .

[75] Due to administrative costs, it may be optimal in such a framework to ignore further demonstrations of efficiencies by the merging parties unless they exceed the credited level by more than a small amount. (Likewise, anticipating an argument later in this section, it may be optimal to ignore demonstrations by the government that efficiencies are below this level unless they fall short by some notable degree.) *See* Louis Kaplow, *Optimal Insurance Contracts When Establishing the Amount of Loss Is Costly*, 19 GENEVA PAPERS ON RISK & INS. THEORY 139 (1994); Louis Kaplow & Steven Shavell, *Accuracy in the Determination of Damages*, 39 J.L. & ECON. 191, 195–98, 206–09 (1996). However, particularly for large mergers, where the costs of collecting and analyzing information may be fairly small relative to the stakes, this sort of adjustment may be secondary. More broadly, the optimal setting of an efficiency credit is not the correct framing; rather, the principles of optimal information collection sketched in subsection I.B.1 should, from the outset, guide how decisions are made both on whether and what information to collect next and how to decide the case, as elaborated below. Of central relevance regarding efficiencies in particular is the diagnosticity/cost ratio of the best information not yet collected. The greater the value of that information, the greater is the range over which the information should be collected. If instead one is constrained to a structured decision protocol, a lower efficiency credit would tend to be desirable the smaller are typical efficiencies and the less costly they are to learn more about.

[76] In any case, considered solely as a decision rule, it is not optimal to set B^* above H and, relatedly, above H^* (because then there is a range of H, $H^* < H < B^*$, in which it is possible that the wrong decision will be made—in particular, prohibiting the merger when $H^* < H < B \leq B^*$).

The analysis in Part I as a whole should lead us to be skeptical of this protocol, both the setting of significant thresholds, *H** and *B**, and the sequencing that involves first examining *H* and then, only if it exceeds some *H**, turning to *B*. In addition to infirmities as a final decision rule taking the information set as given, information is not in general optimally collected in this fashion. We also saw that there is an intimate interrelationship between optimal information collection and optimal decisionmaking.

One way to see these points is to consider how high to set *H** and *B**. Should the efficiency credit be the same in all cases? (And is it, say, $1,000,000 or $100,000,000? Or is it a common percentage, perhaps of sales? What percentage?) Should it vary by industry? (Perhaps different for hospital mergers and mergers of wireless communication systems?) Or more specifically by the merger itself? (Perhaps some horizontal mergers in a given industry have partial overlap but also substantial complementarity compared to others.)

In light of such questions, it seems difficult not to look at least somewhat at the particular merger under review. At this point, however, a reviewing agency would be examining some information regarding *B* in order to set the efficiency credit, *H**—before examining *H*. This, in turn, leads naturally to the question of how much information on *B* should be collected in the first pass. The answer should depend in rough terms on the diagnosticity/cost ratio of what additional information on *B* might be gathered, so in some cases the best answer might be very little and in others much more. Note also that, the more one sets an *H** based on actual case information about *B*, the more the three-step protocol begins to dissolve into unconstrained balancing, as will be elaborated below.[77]

But one should also compare the value of further information on *B* to that regarding *H*. That is, there is no optimal stopping point on *B*, in a vacuum. Instead, as described in subsection I.B.1, the diagnosticity/cost ratio of the next best available information on *B* should be compared to that on *H*. When the latter is higher, we should switch to *H*. Furthermore, we should not then proceed, until some endpoint, on *H*, but rather switch back to *B* when that is the subject of the next most valuable clump of information that we might collect.

We also learned to be skeptical of the implicit assumption that information on *H* and on *B* is distinct. First, some information pertains to both. Second, when one accounts for information collection synergies, it usually makes sense to clump information by source rather than by issue. In addition, it will often make sense to proceed in parallel in light of time limits for agency review.

Interestingly, for horizontal mergers, more of the information separates into *H* and *B* than is the case in some other settings. The reason is that anticompetitive effects are concerned with the merging parties' competitive interactions, the nature of consumer demand, and so forth, whereas countervailing efficiencies often involve production technology and how well the two firms' operations might mesh. That said, there may be overlap between *H* and *B* because combining the firms' operations may have implications for how customers are served, which may influence competitive interaction,[78] and because considerations of entry that are relevant to

[77] Relatedly, it does not seem sensible to expend substantial effort, within an agency or in court, to engage in an elaborate contest about the correct level of *B* to impute, without regard to the facts of the case, rather than simply to do the best one can to assess what facts exist about the actual level of *B*.

[78] Consider, for example, a merger of two hospitals in the same region. It is often true that the closer they are geographically, the greater will be the diminution in competition but the more plausible will be productive synergies. For example, if they plan to combine two departments—eliminating one and expanding the other—this suggests that patients at the former may now use the latter, which may more broadly indicate that the two hospitals are close competitors for other services as well.

anticompetitive effects often depend on the sorts of technological considerations that may inform the analysis of efficiencies.[79]

And there are other reasons that the two issues can be interdependent, including the important generic point regarding motivation: the parties are ordinarily presumed to expect to profit by their proposed merger, so much information seemingly on only *H* or only *B* relates to both. If anticompetitive effects seem large, they can motivate a merger that generates few, if any efficiencies (or even generates diseconomies), whereas if they seem low, then the merger is more likely to be motivated by proffered efficiencies. Conversely, if efficiencies seem large, they readily motivate a merger without anticompetitive effects, whereas if they seem negligible (or there are diseconomies), then the merger is more likely to be motivated by anticompetitive effects.[80]

Taking all of these considerations together, it appears that it would often be optimal to analyze a prospective merger in a flexible manner that sometimes alternated the collection of information about *H* and about *B* and often collected information that bears both on *H* and on *B*. And, following the analysis of subsection I.B.1, all of the information in hand, at each point—whether on *H* or on *B* or on both—should be used in deciding what to do next: collect more information (and what that would be) or make a decision (to permit or prohibit the merger).

When this is done, there is obviously no sequential siloing of *H* and *B*. Nor are there distinct thresholds, *H** and *B**. Rather, along the way, an agency's estimates of both *H* and *B* are continually revised, and those estimates, the information in hand, and what can be expected to be learned from further investigation, all determine whether and how to proceed. If at some point it seems fairly certain that *H* substantially exceeds *B*, a decision not likely to be altered in light of additional information that might be collected, it is optimal to stop and prohibit the merger. If instead *B* clearly exceeds *H*, then the agency should stop and allow the merger. If the matter is closer and, in particular, any tentative view may plausibly be overturned by what one may learn from further investigation at a reasonable cost, then information collection should continue.

One suspects that, despite declarations in formal agency guidance documents or by courts to the contrary, this more flexible approach is closer to what agencies actually do.[81] In particular, if they become convinced that $H \leq B$, they are likely to stop and allow the merger. It is also notable that firms seeking to merge provide substantial information to the agencies up front, more than is formally required, and (especially in mergers that the proponents fear might be challenged) including significant information on efficiencies. That is, the decisionmaker has some information on *B* to begin with. In addition, through formal and informal means, agencies

[79] For example, if two large firms claim that their merger will generate significant economies of scale, that would tend to suggest that their increased incentive to raise prices will not readily be countered by new entrants that may need to start small for a substantial period of time.

[80] Put more precisely, the domain of possible mergers involves some joint distribution of the *H* and *B* that they would generate. Even if those distributions were independent, we are supposing in addition that only mergers that generate positive expected profits would be proposed, which in general would create dependence because *H* (anticompetitive effects, generally resulting from the ability to charge higher prices) and *B* (efficiencies) are both sources of profit. Furthermore, even if we incorporated possible managerial (agency) and behavioral phenomena, such jointness would plausibly exist and have the character presented in the text.

[81] *See* FED. TRADE COMM'N & U.S. DEP'T OF JUSTICE, COMMENTARY ON THE HORIZONTAL MERGER GUIDELINES 2 (2006) [hereinafter U.S. MERGER GUIDELINES COMMENTARY] ("The ordering of these elements in the Guidelines, however, is not itself analytically significant, because the Agencies do not apply the Guidelines as a linear, step-by-step progression that invariably starts with market definition and ends with efficiencies or failing assets. Analysis of efficiencies, for example, does not occur 'after' competitive effects or market definition in the Agencies' analysis of proposed mergers, but rather is part of an integrated approach."). The Commentary's general description appears to be borne out in its extensive section on efficiencies. *See id*. at 49–59.

press on these proffers and gather further information on whatever fronts they expect to be helpful.

Return now to the notion of an efficiency credit. The foregoing indicates that this may best be viewed less as a prespecified H^* than as something more like a stand-in for the current estimate (guesstimate) of the actual B in the merger at hand. Under this interpretation, however, the step 1 test of whether $H > H^*$ (with H^* as the efficiency credit) has morphed into a running inquiry into whether $H > B$, the unconstrained balancing test.

Suppose instead that one has not yet looked at B, or not very much, so that H^* more comports with the original credit notion. A further question to ask is why—in addition to allowing the merging parties in step 2 to establish that B is unusually high (greater than the credit)—we should not also allow the agency to establish that B is unusually low (much smaller than the credit). As mentioned, merging parties often make proffers at the outset regarding efficiencies. When this information either directly suggests that efficiencies are smaller than average or when the parties claim otherwise but are not credible, such a downward adjustment is natural. For example, if most of the proffered efficiencies can fairly obviously be achieved without the merger, why not impute a very low degree of efficiencies, that is, set a low H^*, thereby reducing the burden regarding the demonstration of H? In addition, as explained, an optimally designed information gathering process will, at points, collect further information on B along the way—either because that is the next best information to collect or because B is illuminated by information that bears jointly on H and B. That information should then be employed, including the possibility that it suggests an unusually low B and hence does not require as high of an H to justify a challenge.

This approach would not help the agency if H^* is set high initially, the $H > H^*$ test fails, and it never gets to step 2. Instead it would wish, preemptively, to be able to show that B is probably low, which would lower H^*, making step 1 easier to pass. In practice, an agency, internally, is free to do so. In court, if it was thought that the structured decision procedure stated the law, there may be a greater problem, although the agency could urge the court to appropriately set H^* at a low level, using the evidence on the low B to influence that calibration (keeping in mind that the magnitude of H^* is not actually specified in the merger guidelines or in court decisions).[82]

A final point to consider is the tendency of merger guidelines and practice to be highly skeptical of claimed efficiencies. For example, the U.S. Merger Guidelines relegate efficiencies

[82] The U.S. MERGER GUIDELINES, *supra* note , § 5.3, employ the Herfindahl-Hirschman Index ("HHI"), a derivative of market shares, which requires market definition, presented in *id*., § 4. For the courts, the most relevant case is *Philadelphia Bank*, known for its so-called structural presumption. *See* United States v. Philadelphia National Bank, 374 U.S. 321, 363 (1963) ("[A] merger which produces a firm controlling an undue percentage share of the relevant market, and results in a significant increase in the concentration of firms in that market, is so inherently likely to lessen competition substantially that it must be enjoined in the absence of evidence clearly showing that the merger is not likely to have such anticompetitive effects."). This presumption likewise speaks of market shares, which require market definition. These market share threshold tests, however, do not really speak to the pertinent competitive effects. *See* Louis Kaplow, *Market Share Thresholds: On the Conflation of Empirical Assessments and Legal Policy Judgments*, 7 J. COMPETITION L. & ECON. 243 (2011). Moreover, the market definition concept on which they rely is incoherent. *See* Kaplow, *Market Definition*, *supra* note ; Louis Kaplow, *Market Definition and the Merger Guidelines*, 39 REV. INDUS. ORG. 107 (2011). As explained in those sources, the only plausible way to even think about choosing a market definition requires having in hand an estimate of the effects, the very thing that the market definition and resulting market shares were meant to be a partial proxy for in the first place. Another difficulty with *Philadelphia Bank*'s structural presumption is that it is, as commonly understood, only a presumption. Its strength, when faced with inevitable attempts at rebuttal, is obscure. Nevertheless, all of this apparatus, to varying degrees, appears to influence courts, particularly in light of their self-perceived limitations in attempting to ascertain the pertinent effects.

to a modest section near the end, and the topic sentences introducing three key points begin as follows: "The Agencies credit only those efficiencies . . . ." "Efficiencies are difficult to verify and quantify . . . ." "Efficiency claims will not be considered if . . . ."[83] This skepticism might to some extent rationalize the efficiency credit, which, in the form described above, requires both the belief that efficiencies are commonplace and often significant[84] and also that this belief should be embodied in a large credit that usually sticks (in the sense that it is not superseded in step 2).

Recall that a key feature of the value of information noted in subsection I.B.1 concerns diagnosticity: the ability of additional information to resolve uncertainty about a matter. High diagnosticity has two prerequisites: that there is significant uncertainty to begin with, and that the information that might be collected would materially reduce that uncertainty. Regarding merger efficiencies, the former is often regarded to be true but the latter might be doubted, at least in many settings. When that is so, it may be that agencies and courts can do little better than positing a guesstimate and, typically, sticking with it.

That said, it is not clear how universal this inscrutability problem is, that it justifies a significant credit that should rarely be questioned and then in only one direction (upward, as advocated by the merging parties), or that efficiencies are nearly always so much harder to estimate than are anticompetitive effects. Starting with the first two points, certainly some assessment is often possible.[85] For example, there is often significant attention to whether efficiencies are "merger-specific," the term used for those that cannot readily be achieved by the less restrictive alternative of forgoing the merger and instead making other arrangements, often contractual.[86] There are significant difficulties, many relating to subtleties in the analysis in the field that economists refer to as contract theory and its particular application to the theory of the firm.[87] To suggest part of the problem, it is easy to imagine that just about anything done via merger might be accomplished by sufficiently elaborate contracting, but by similar logic it is not

[83] *See, e.g.*, U.S. MERGER GUIDELINES, *supra* note , at 30.

[84] This view is controversial. *Compare* Sandra Betton, B. Espen Eckbo & Karin S. Thorburn, *Corporate Takeovers*, *in* 2 HANDBOOK OF EMPIRICAL CORPORATE FINANCE 289 (B. Espen Eckbo ed., 2008) (survey indicating that the combined value of merging firms increases); *id*. at 294 ("[W]e show that the value-weighted sum of announcement-induced three-day abnormal stock return to bidders and targets is significantly positive. This conclusion holds for the entire sample period 1980–2005 as well as for each of the five-year subperiods."); *id*. at 391–99 (explaining how a number of types of evidence are inconsistent with the collusion hypothesis), *with* Lars-Hendrik Röller, Johan Stennek & Frank Verboven, *Efficiency Gains from Mergers*, *in* EUROPEAN MERGER CONTROL: DO WE NEED AN EFFICIENCY DEFENCE? 84, 112 (Fabienne Ilzkovitz & Roderick Meiklejohn eds., 2006) (concluding that "there seems to be no support for a general presumption that mergers create efficiency gains"). *See also* Kaplow & Shapiro, *supra* note , at 1152–57 (surveying empirical evidence on the effects of mergers).

[85] Indeed, the U.S. Merger Guidelines themselves suggest that this is so. *See* U.S. MERGER GUIDELINES, *supra* note , at 31 ("The Agencies have found that certain types of efficiencies are more likely to be cognizable and substantial than others. For example, efficiencies resulting from shifting production among facilities formerly owned separately, which enable the merging firms to reduce the incremental cost of production, are more likely to be susceptible to verification and are less likely to result from anticompetitive reductions in output. Other efficiencies, such as those relating to research and development, are potentially substantial but are generally less susceptible to verification and may be the result of anticompetitive output reductions. Yet others, such as those relating to procurement, management, or capital cost, are less likely to be merger-specific or substantial, or may not be cognizable for other reasons."). Furthermore, the agencies' commentary offers a number of examples from past cases. *See* U.S. MERGER GUIDELINES COMMENTARY, *supra* note , at 49–59.

[86] *See* U.S. MERGER GUIDELINES, *supra* note , at 30 & n. 13.

[87] See Kaplow & Shapiro, *supra* note , at 1164, for elaboration and citations. The Nobel Prize in economics was recently awarded to Oliver Hart and Bengt Holmström for work in this domain. *See* PRIZE COMM. OF THE ROYAL SWEDISH ACAD. OF SCI., CONTRACT THEORY: SCIENTIFIC BACKGROUND ON THE SVERIGES RIKSBANK PRIZE IN ECONOMIC SCIENCES IN MEMORY OF ALFRED NOBEL 2016 (2016).

clear why firms as such need to exist in the first place. Relatedly, firms are sometimes described as a nexus of contracts,[88] further blurring the distinction between a firm and mere contracting.

Another part of the challenge is practical, which also relates to the third point. To the extent that antitrust agencies, over time, have been staffed with teams of industrial organization economists who specialize in the analysis of anticompetitive effects, and not with teams of experts in business organization, operations, and the like, they might find it easier to assess *H* than *B*. Perhaps greater diversification of in-house staff and expanded use of industry experts to analyze particular cases would be helpful in this regard.[89] Moreover, as is familiar, even with existing expertise it is quite difficult to predict *H* in any event, further calling into question the view that it should pretty much always be assumed that *H* is more readily illuminated through further investigation than is *B*.[90] Finally, there is the potential (at least sometimes, and often enough to be worthwhile) to illuminate merger effects through the examination of the parties' internal documents, not only relating to the merger process itself but also, and perhaps less subject to manipulation, previous analysis and planning (documents prepared in the ordinary course of business). Such information may bear on *H*, on *B*, or on both.

In sum, it seems hard to rationalize a fairly rigid structured decision procedure for the assessment of horizontal mergers of the sort that appears on the surface to govern merger review. One suspects that, in practice, agencies are notably more flexible, and perhaps they could be even more so. Their professed skepticism about establishing efficiencies may in part be strategically motivated: once the agency concludes that $H > B$ in a particular case, when it goes to court it may hope that the tribunal will be skeptical of the parties' claims that *B* is much larger. Of course, it also hopes that the tribunal will be skeptical of the defendants' suggestion that *H* is much smaller.[91]

[88] *See* Michael C. Jensen & William H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, 3 J. FIN. ECON. 305 (1976).

[89] It is also often said that some of the difficulty arises "because much of the information relating to efficiencies is uniquely in the possession of the merging firms." U.S. MERGER GUIDELINES, *supra* note , at 30. This feature may justify demands that the merging parties produce such information, but in many areas of law (such as negligence in torts) this is true but is not seen either to justify a large credit or to then be reluctant to adjust it. Moreover, many sorts of efficiencies might best be gauged by broader industry experience, which might be understood by industry consultants outside the merging firms. And, ironically, much of the information on anticompetitive effects is also in the merging firms' possession, but we do not deem this to be a reason to refrain from analyzing them as best we can.

[90] The U.S. Merger Guidelines most explicit methodology, which uses the hypothetical monopolist test to define markets and then make inferences from the market shares therein, is extremely problematic. *See* sources cited *supra* note . Merger simulation techniques have become increasingly more sophisticated but require data that is not always available and may be sensitive to structural assumptions that are difficult to confirm. Internal documents and opinions of industry participants and experts can often be illuminating but, as noted in the text, these sources may often clarify efficiencies as well. The main qualification is that large buyers may have a good sense of pricing incentives but not of the internal operations of the merging parties.

Interestingly, concerns about coordinated effects (that a merger may facilitate coordinated oligopolistic price elevation) that once dominated merger analysis have receded to a substantial degree, reflecting in no small part the difficulty of predicting them in convincing ways. It is not obvious that the result is desirable, and it is ironic to reject most or all of such challenges because of an implicit imputation of a significant efficiency credit that is in large part justified (and made largely immune from refutation by the agencies) by the difficulty of determining the actual magnitude of efficiencies in a given case.

[91] *See supra* note (discussing *Philadelphia Bank*'s structural presumption, the U.S. Merger Guidelines, and problems with the use of market definition and market share to assess anticompetitive effects). Outsiders to antitrust may wonder why the agencies' guidelines (which are not regulations and do not even purport to bind the agencies themselves) are so influential on courts. Among the central reasons are judges' appreciation of their need for guidance, the absence of substantive Supreme Court merger decisions for nearly a half century (during which time the Court's direction in other areas of antitrust has shifted, casting doubt on the extent to which the old merger cases should be regarded as good law), and the broad respect the U.S. Merger Guidelines have commanded among antitrust practitioners and academics.

As discussed in the immediately preceding section on antitrust's rule of reason, non-expert judges,[92] who may be hearing the first (and last) merger case of their careers,[93] face steep challenges in deciding these complex cases. Strong presumptions are accordingly appealing to them, both in reaching decisions and in justifying their conclusions in written opinions. Ideally, we would like judges to be skeptical of but thoughtful about both sides' claims about H and B, enjoining the truly undesirable mergers but not others. That is a tall order, but, as the present analysis suggests, it is hardly obvious that it is aided rather than confounded by a structured decision procedure, particularly when neither H^* nor B^* is articulated and hence must implicitly be set by the judge in the course of deciding the case.

III. Title VII Disparate Impact

The structured decision rule for Title VII disparate impact cases[94] is contained in § 703(k), added by the Civil Rights Act of 1991:

[92] Merger cases involve the government seeking an injunction (usually a preliminary injunction that, if granted, may lead the parties to abandon the merger) and thus are decided by judges.

[93] A cursory examination of the number of federal district judges, their average tenure, and the number of merger cases in court yields a mean (over a career) of approximately 0.3 merger cases per judge. In practice, some district court judges, such as those in the District of Columbia, will hear more.

[94] This Part focuses exclusively on the structured decision procedure for disparate impact cases. It is familiar that disparate treatment cases also have a structured protocol, and one that bears some similarities while also having notable differences. Although some of the analysis developed here may well have implications for disparate treatment doctrine, the differences are substantial enough that the matter will not be pursued further. More broadly, some other areas of discrimination law feature disparate impact tests. *See, e.g.*, Texas Dep't of Housing & Community Affairs v. Inclusive Communities Project, Inc., 135 S. Ct. 2507 (2015) (holding that disparate impact claims are cognizable under the Fair Housing Act); Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11460 (Feb. 15, 2013) (codified at 24 C.F.R. 100.500 (2014)) (implementing a three-step burden-shifting framework for disparate impact claims under the Fair Housing Act). Again, some of the present analysis may be illuminating in these other contexts.

> An unlawful employment practice based on disparate impact is established under this subchapter only if—(i) a complaining party demonstrates that a respondent uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin and the respondent fails to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity; or (ii) the complaining party makes the demonstration described in subparagraph (C) with respect to an alternative employment practice and the respondent refuses to adopt such alternative employment practice.[95]

Recasting this provision in the language of this article, step 1 requires the plaintiff to demonstrate that a challenged employment practice causes a disparate impact (*H*), which, if done, shifts the burden in step 2 to the defendant to demonstrate that the practice is job related and consistent with business necessity (*B*), which the plaintiff may attempt to rebut or, in what might be viewed as a third step, demonstrate the existence of a less restrictive alternative, referred to here as an alternative employment practice. On its face, there is no balancing step. Let us now examine each of these steps more closely, including refinements contained in existing doctrine, to see how they relate to this article's analysis.

[95] 42 U.S.C. § 2000e–2(k)(1)(A) (2012). Subparagraph (C) states: "The demonstration referred to by subparagraph (A)(ii) shall be in accordance with the law as it existed on June 4, 1989, with respect to the concept of 'alternative employment practice.'" *Id.* § 2000e–2(k)(1)(C); *see infra* note (elaborating this facially opaque statement).

Step 1 requires the plaintiff to "demonstrate" H, which is understood to indicate a persuasion burden.[96] Accordingly, this structured decision procedure entails the sequential siloing of H and B that has been addressed throughout this article.

It is not apparent on its face whether step 1 requires $H > H^*$, or stated another way, whether $H^* > 0$ and, if so, by how much. Although the statute is silent,[97] which might be interpreted as erecting no threshold, a number of considerations suggest that H^* is nontrivial. First, treatises and some cases state that the disparate impact that a plaintiff must demonstrate needs to be substantial or significant.[98] Second, reference is sometimes made to the Equal Employment Opportunity Commission Uniform Guidelines's "four-fifths" "working rule," under which the disadvantaged group must be treated at least twenty percent worse than others are to

[96] *See, e.g.*, 1 BARBARA T. LINDEMANN, PAUL GROSSMAN & C. GEOFFREY WEIRICH, EMPLOYMENT DISCRIMINATION LAW 3-12 to -13 (5th ed. 2012) ("A court will consider statistical evidence offered by both the plaintiff and the defendant to determine whether, on the basis of the most probative evidence, the challenged practice or selection device has a substantial disparate impact on a protected group. The burdens of production and persuasion at this stage are on the plaintiff.").

[97] *See, e.g.*, Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 518 (2003) ("At the same time, section 703(k) leaves a great deal unsettled about the nature of disparate impact actions. The statute does not describe the degree of disparity needed to trigger disparate impact liability . . . .").

[98] *See, e.g.*, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , at 3-13 (stating that the challenged action must have "a substantial disparate impact"); 1 CHARLES A. SULLIVAN & LAUREN M. WALTER, EMPLOYMENT DISCRIMINATION: LAW AND PRACTICE 269 (4th ed. 2009) (heading their section "Markedly Disproportionate Impact"); *id*. ("But precisely how disparate the impact must be shown has never been determined."); Pamela L. Perry, *Two Faces of Disparate Impact Discrimination*, 59 FORDHAM L. REV 523, 570-74 (1991) (discussing competing perspectives by commentators and differing statements by courts regarding whether a quantitatively substantial impact should be required in addition to statistical significance). The basis for such statements is not entirely clear. For example, Barbara Lindemann, Paul Grossman, and Geoffrey Weirich claim support from the statute itself (which is silent) and from *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 425–28 (1975), a case that predates and thus is arguably superseded by the (ambiguous) statute. *Albemarle*'s most on-point language is: "This burden arises, of course, only after the complaining party or class has made out a prima facie case of discrimination, i.e. has shown that the tests in question select applicants for hire or promotion in a racial pattern *significantly different* from that of the pool of applicants. *See McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802 (1973)." *Id.* at 425 (emphasis added). Interestingly, *McDonnell Douglas*, the source of support, is a disparate treatment case, and the context of the discussion in *Albemarle* is defenses, which arise at step 2 (and the passage that is cited in *McDonnell Douglas* does not itself contain any language pertaining to whether a plaintiff must demonstrate the "significance" of anything). *See also* Griggs v. Duke Power Co., 401 U.S. 424, 426 (1971) (referring, at the outset of the opinion, to the question before the Court as involving "requirements [that] operate to disqualify Negroes at a substantially higher rate than white applicants," but failing later to repeat any variant of "substantial" or to address the matter in other ways); Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 307–08 (1977) (indicating that demonstration of "gross statistical disparities" can, alone, in a proper case, establish a prima facie case, but not commenting on whether some minimal degree of disparity is always required); Teamsters v. United States, 431 U.S. 324, 339 (1977) (similarly suggesting that significant disproportions can establish a prima facie case). Later, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , at 3-19, states that "[t]he lower courts . . . have not fashioned a uniform rule" to fill in the gap due to "[t]he Supreme Court . . . [having] given no definitive guidance," and they also cite the (also pre-1991-Act) plurality opinion in *Watson v. Fort Worth Bank & Trust*, 497 U.S. 977 (1988) (opinion of O'Connor, J.), which states: "Once the employment practice at issue has been identified, causation must be proved; that is, the plaintiff must offer statistical evidence of a kind and degree sufficient to show that the practice in question has caused the exclusion of applicants for jobs or promotions because of their membership in a protected group. Our formulations, which have never been framed in terms of any rigid mathematical formula, have consistently stressed that statistical disparities must be *sufficiently substantial* that they raise such an inference of causation. . . . Later cases have framed the test in similar terms." *Id.* at 994-95 (emphasis added) (quotations of the previously noted language from *Griggs* and *Albemarle* omitted). Rather confusingly, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , at 3-24 to -26, concludes its discussion of the issue by, in sequence: discussing a circuit court case finding that statistical significance can be insufficient when the disparity is "of limited magnitude"; citing a later circuit court case asserting that no circuit court has ever held that practical significance was required; and stating that courts do assess whether any differences are "substantial" (and citing a large number of cases to support that proposition).

prompt action.[99] Although not binding on the agency or on courts, this rule of thumb seems to have had some influence, although it is unclear how much so in more recent cases.[100] Third, it is typically required that the plaintiff demonstrate a statistically significant effect,[101] and, even if a showing of statistical significance is alone sufficient, it is a familiar property of significance tests that, for a given sample size, larger substantive effects are more likely to be found statistically

[99] 29 C.F.R. § 1607.4D ("D. *Adverse impact and the 'four-fifths rule.'* A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant, or where special recruiting or other programs cause the pool of minority or female candidates to be atypical of the normal pool of applicants from that group.").

[100] *See, e.g.*, Lindemann, Grossman & Weirich, *supra* note , at 3-22 to -24; Sullivan & Walter, *supra* note , at 278 ("The Supreme Court has neither adopted nor rejected the four-fifths rule. It has generally described the Uniform Guidelines in terms of a mechanism for allocating scarce enforcement resources, rather than as a rule of law."). Some courts reject the four-fifths rule or any independent requirement of substantiality when statistical significance is established. *See, e.g.*, Jones v. Boston, 752 F.3d 38, 48–53 (1st Cir. 2014); *see also id*. at 46 ("The Supreme Court has most recently described a prima facie showing of disparate impact as 'essentially a threshold showing of a significant statistical disparity . . . and nothing more.' *Ricci v. DeStefano*, 557 U.S. 557, 587 . . . (2009)."); *id*. at 51–52 (noting respects in which the four-fifths rule is arbitrary depending on selection rates and whether the differential is defined by reference to those who pass or those who fail the test). The *Ricci* reference cited in *Jones* regarding the requirement in step 1 is made in passing (the Court's emphasis was on the later steps); it states: "essentially, a threshold showing of a significant statistical disparity . . . and nothing more." 557 U.S. at 587. In turn, *Ricci*'s only authority for this proposition is *Connecticut v. Teal*, 457 U.S. 440, 446 (1982), which contains no more than a bland statement of a requirement of disparate impact, with reference neither to statistical disparities nor to whether anything more might be required. In the majority opinion in that case, the only reference to a word having "statistics" as a root appears later and as an aside: "See also *New York Transit Authority v. Beazer*, 440 U.S. 568, 584 (1979) ("A prima facie violation of the Act may be established by statistical evidence showing that an employment *practice* has the effect of denying members of one race equal access to employment *opportunities*") (emphasis added)." 457 U.S. at 450. My own impression from a number of more recent lower court cases and from discussions with some in the field is that the focus is largely on statistical significance. There are exceptions, *see, e.g.*, Waisome v. Port Auth. of N.Y. & N.J., 948 F.2d 1370, 1376–77 (2d Cir. 1991), and other contrary indications, *see, e.g.*, Federal Judicial Center, Reference Manual on Scientific Evidence 252 (3d 3d. 2011) ("When practical significance is lacking—when the size of the disparity is negligible—there is no reason to worry about statistical significance."). For an earlier analysis, see Paul Meier, Jerome Sacks & Sandy L. Zabell, *What Happened in* Hazelwood*: Statistics, Employment Discrimination, and the 80% Rule*, in Statistics and the Law 1 (Morris H. DeGroot, Stephen E. Fienberg & Joseph B. Kadane eds., 1986).

[101] *See, e.g.*, Lindemann, Grossman & Weirich, *supra* note , at 3-19 to -22; Sullivan & Walter, *supra* note , at 276–80; Michael J. Piette, *Module IV: Reference Guide for Analyzing Allegations of Employment Discrimination: An Economist's View*, *in* Expert Economic Testimony: Reference Guide for Judges and Attorneys 221, 237 (Thomas R. Ireland et. al eds., 1998) ("[T]he Supreme Court in a landmark case involving the selection of minorities to serve on grand juries, *Castaneda v. Partida*[, 430 U.S. 482, 486 n.17] (1977), stated '. . . if the difference between the expected value and the observed number is greater than two or three standard deviations,' then the process we are observing is unlikely to have occurred by chance alone. This 'rule' was quickly applied to the employment discrimination area in *Hazelwood School District vs. U.S.*[, 433 U.S. 299, 311 n.17] (1977)[,] and became known as the 'Hazelwood Standard.'" (emphasis omitted)); *see also* Smith v. Xerox Corp., 196 F.3d 358, 366 (2d Cir. 1999) ("Although courts have considered both the four-fifths rule and standard deviation calculations in deciding whether a disparity is sufficiently substantial to establish a prima facie case of disparate impact, there is no one test that always answers the question. Instead, the substantiality of a disparity is judged on a case-by-case basis."); Sullivan & Walter, *supra* note , at 276 ("[H]ow much impact is needed for a claim to exist . . . is addressed here, but it is complicated by the continuing confusion between statistical significance and quantum of impact."); Jennifer L. Peresie, *Toward a Coherent Test for Disparate Impact Discrimination*, 84 Ind. L.J. 773, 775 (2009) (describing lower courts' mixed use of a statistical significance requirement and the four-fifths rule and stating that "none of the circuits have a uniform standard for evaluating disparate impact cases").

significant.[102] In all, it seems that we may well have $H^* > 0$, although (aside from the four-fifths rule) there is little indication of the actual magnitude of H^*.

Regarding Title VII's structured protocol as a final decision rule, we know from the analysis in subsection I.A.1 that setting H^* significantly above zero can result in erroneous outcomes: failures to assign liability even though $H > B$, which arises when $B < H \leq H^*$. The extent of this problem depends, of course, on how high H^* is set, which is unclear. One should also keep in mind the possible screening function of setting $H^* > 0$, recalling as well subsection I.B.1's discussion of how screening may best be accomplished in a more flexible manner that is more in line with balancing.

Another important question in this domain is whether H^* in practice is set in a vacuum, which most statements seem to envision, or is instead set contextually, with an (at least implicit) eye toward the likely level of B in the case at hand. Step 1's demand is most likely to have bite at the summary judgment stage, in assessing whether a plaintiff's statistical demonstration (contained in an expert report) is sufficient to create a genuine dispute, and at the end of a trial (or after a plaintiff has presented its full case). In these settings, the judge[103] will have had some, or even complete, exposure to the defendant's case when making a decision on step 1, so such an implicit violation of the structured decision rule is plausible. To that extent, the de jure decision rule, which is both sequential and does not purport to involve balancing, may operate somewhat like a de facto balancing test of sorts (on which more in a moment).

If the plaintiff meets its burden on H, step 2 then requires the defendant to "demonstrate" B, also indicating a persuasion burden.[104] The major question concerns, in the parlance of this article, the magnitude of B^*. The statutory language is that, to avoid liability, a challenged employment practice shown to have a disparate impact must be "job related for the position in question and consistent with business necessity," drawn from a statement in *Griggs v. Duke Power Co.*[105] that, in context, seems to have been casually crafted with no particular meaning in

[102] *See, e.g.*, *Jones*, 752 F.3d at 53 ("First, the very need to show statistical significance will eliminate small impacts as fodder for litigation in many instances because proving that a small impact is statistically significant generally requires large sample sizes, which are often unavailable."); Piette, *supra* note , at 246. The centrality of tests of statistical significance in this (and other) legal settings, although reflecting an understandable borrowing from its use in the social sciences and medicine in particular, is also puzzling in a number of respects. First, there is the distinction related to the point in the text, between statistical significance and practical significance (effect size). Second, legal proof burdens, including in the Title VII setting, are typically formulated as probabilities, with the preponderance rule requiring that something be more likely than not. *See, e.g.*, 2 MCCORMICK ON EVIDENCE 484 (Kenneth S. Broun ed., 6th ed. 2006) ("The most acceptable meaning to be given to the expression, proof by a preponderance, seems to be proof which leads the jury to find that the existence of the contested fact is more probable than its nonexistence." (citing MODEL CODE OF EVIDENCE R. 1(3))). Such a probability is a Bayesian posterior probability regarding the truth of a proposition, whereas statistical significance tests ask a qualitatively different question. *See, e.g.*, David H. Kaye, *Statistical Significance and the Burden of Persuasion*, LAW & CONTEMP. PROBS. 13, Autumn 1983. In addition, when the concern is with ex ante behavior (here, the deterrence of discriminatory employment practices), none of these notions is apt. *See* Kaplow, *Burden of Proof*, *supra* note ; Kaplow, *Likelihood Ratio Tests*, *supra* note .

[103] Title VII disparate impact claims are tried to a judge—the 1991 amendments allowing for damages, and therefore a jury, being applicable only to disparate treatment cases. *See* 42 U.S.C. § 1981a(a)(1) ("In an action brought by a complaining party under section 706 or 717 of the Civil Rights Act of 1964 [42 U.S.C. 2000e–5, 2000e–16] against a respondent who engaged in unlawful intentional discrimination (not an employment practice that is unlawful because of its disparate impact) prohibited under section 703, 704, or 717 of the Act [42 U.S.C. 2000e–2, 2000e–3, 2000e–16] . . ., the complaining party may recover compensatory and punitive damages as allowed in subsection (b) . . . ."); *id.* § 1981a(c)(1) ("If a complaining party seeks compensatory or punitive damages under this section—(1) any party may demand a trial by jury . . . .").

[104] *See, e.g.*, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , at 3-13 ("If impact is established, the inquiry becomes whether the practice or selection device is 'job related for the position in question and consistent with business necessity.' The burdens of production and persuasion at this stage are on the defendant . . . .").

[105] 401 U.S. 424 (1971).

mind.[106] On its face, this language is subject to many interpretations that collectively span a broad range of possibilities[107] and may be associated with different conceptions of disparate impact law.[108] This ambiguity is often noted and does not appear to be fully resolved.[109]

One interpretation, emphasizing "job related" (which seems to get more attention than does "business necessity"[110]), takes that phrase literally and minimally, corresponding to $B^* = 0$ (or perhaps slightly more). One can, moreover, understand a way of thinking that would support such a view. Specifically, if a practice is at all job related, then its use enhances productivity[111]

[106] *See id*. at 431 ("The Act proscribes not only overt discrimination, but also practices that are fair in form, but discriminatory in operation. The touchstone is business necessity. If an employment practice which operates to exclude Negroes cannot be shown to be related to job performance, the practice is prohibited."). This language seems to state "business necessity" as the test, but then in essence to define it as whether a practice is "related to job performance." The discussion in *Griggs* that follows this announcement can readily be understood as suggesting that the defendant had made no demonstration of any job-relatedness, so even a minimal requirement was sufficient to support liability. However, the language is not sufficiently sharp to command such a reading or to show that it would have been enough for the defendant to persuasively demonstrate some relationship, even a minimal one. Subsequent language arguably suggests a minimal threshold. *See id*. at 436 ("What Congress has forbidden is giving these devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance. Congress has not commanded that the less qualified be preferred over the better qualified simply because of minority origins. Far from disparaging job qualifications as such, Congress has made such qualifications the controlling factor, so that race, religion, nationality, and sex become irrelevant. What Congress has commanded is that any tests used must measure the person for the job, and not the person in the abstract.")

[107] The failure of the statutory language to resolve this core ambiguity was intentional, reflecting the need for compromise in passing the 1991 Act. *See, e.g.*, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , at 3-38; SULLIVAN & WALTER, *supra* note , at 284–85.

[108] It is hoped that the discussion to follow makes clear that this article neither takes any stand on which view best comports with existing caselaw nor advances a position on the normative force of any view regarding the purpose of disparate impact law.

[109] *See, e.g.*, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , at 3-13 ("[T]he precise meaning of the substantive standard is not defined in the statute, and continues to be litigated . . . ."); SULLIVAN & WALTER, *supra* note , at 283–91; Primus, *supra* note , at 518 ("At the same time, section 703(k) leaves a great deal unsettled about the nature of disparate impact actions. The statute does not . . . describe how 'necessary' a practice must be for an employer to defend itself successfully on the ground of 'business necessity.' The concepts of business necessity and alternative employment device are taken from pre-1991 cases, and section 703(k) codifies that common law with all of its attendant uncertainties."); *see also id*. at 518 ("[T]here has long been a dispute over whether disparate impact doctrine is an evidentiary dragnet designed to discover hidden instances of intentional discrimination or a more aggressive attempt to dismantle racial hierarchies regardless of whether anything like intentional discrimination is present."); *id*. ("Nor do the 1991 amendments resolve ambiguities concerning the purpose of disparate impact law. Instead, they reflect the lack of consensus among those who passed the amendments about the rationale for and contours of the disparate impact standard."). In examining the cases, there seems to be more emphasis on what type of validation is required for different types of business justifications in different settings than on the magnitude of the claimed productivity boost that must be shown. *See, e.g.*, Amy L. Wax, *Disparate Impact Realism*, 53 WM. & MARY L. REV. 621, 631–35 (2011).

[110] *See, e.g.*, SULLIVAN & WALTER, *supra* note , at 286 ("These opinions, therefore, seem to emphasize job-relatedness over business necessity; indeed, the two terms may mean the same thing."). Consider the mixed depictions of step 2's requirement in a recent case in which the matter was the central question being decided: *Lopez v. Lawrence*, 823 F.3d 102, 111 (1st Cir. 2016) ("[T]his inquiry trains on whether the selection practice—here, the use of the exam—is '*valid*.' In simple terms, a selection practice is valid if it *materially enhances* the employer's ability to pick individuals who are *more likely to perform better* than those not picked." (emphasis added)); *id*. at 116 (describing the question as whether "the practice causing that [disparate] impact serves an *important need* of the employer"); *id*. at 116–17 (indicating that the plaintiffs lose the issue because they do not "claim that the exams are not *materially better predictors* of success [on the job] *than would be achieved by random selection* of those officers to be promoted" (emphasis added)). Other courts as well often focus on validity, *see, e.g.*, Gulino v. N.Y. St. Educ. Dep't, 460 F.3d 361, 383–85 (2d Cir. 2006), which seems to address the stringency of the requisite demonstration rather than what it is that must be demonstrated (whether a slight benefit suffices or something larger, perhaps much larger, is required).

[111] The concept of productivity involves an oversimplification because productivity can rise for discriminatory reasons that are disallowed, a complication set to the side here. (Suppose, for example, that white workers could be shown

at least somewhat. Hence, a profit-motivated employer (or cost-minimizing nonprofit organization) would use the practice if it entertained no thoughts about discrimination. We can compare the case in which an organization procures its paperclips from whatever company charges the lowest price, preferring one brand to another even if the savings are mere pennies—without giving a second's thought as to which company might employ fewer minorities, have its headquarters in an even-numbered zip code, or be headed by a CEO born under a favored sign of the zodiac.

To elaborate, this perspective might be associated with a view that disparate impact doctrine under Title VII is implicitly—despite official statements to the contrary[112]—about a sort of imputed, objective intent to discriminate.[113] As long as $B > 0$, a profit-maximizing employer would choose the challenged practice, thereby failing to support an inference of intent to discriminate. Put another way, the practice leads to the hiring (or promotion, and so forth) of more qualified employees, even if only slightly more so. This would be true even if the disparate impact was large: perhaps the use of a test results in hiring 1000 fewer minority workers and merely saves the employer (on average) a few dollars on each hire.[114] Note further that, if $B < 0$, the employer is leaving money on the table in causing the disparate impact, which is why some might refer to this basis for liability as involving, at least implicitly, an objective imputation of intent.[115]

to have worse esprit and thus be less productive because they would be annoyed by minorities, or that sales may be lost because customers would not like dealing with minority sales personnel.)

[112] *See* Griggs v. Duke Power Co., 401 U.S. 424, 431–32 ("The Act proscribes not only overt discrimination but also practices that are fair in form, but discriminatory in operation. . . . We do not suggest that either the District Court or the Court of Appeals erred in examining the employer's intent; but good intent or absence of discriminatory intent does not redeem employment procedures or testing mechanisms that operate as 'built-in headwinds' for minority groups and are unrelated to measuring job capability. . . . Congress directed the thrust of the Act to the consequences of employment practices, not simply the motivation."). There are a number of reasons that a focus on discriminatory intent may be disclaimed, in addition to the obvious one that the statute does not even implicitly invoke intent in any sense: targeting intent may intensify opposition, make judges reluctant to find discrimination, or lead to confusion in thinking that subjective intent must be demonstrated (which, with organizational defendants, can be problematic, in addition to difficulties of proof even when the focus is on a specific individual's intention). *See also* Primus, *supra* note , at 519 ("The conception of disparate impact doctrine as an evidentiary dragnet is ambiguous about whether it seeks to discover hidden deliberate discrimination or hidden subconscious discrimination; furthermore, the idea of subconscious discrimination is itself subject to more than one interpretation. The alternative idea, that the doctrine aims to dismantle racial hierarchies irrespective of present intentional discrimination, might mean that it aims to integrate the workplace. But it could also mean, less ambitiously, that it aims to integrate the workplace only to the extent that existing hierarchies can be dismantled through the elimination of irrational business practices. Moreover, the self-perpetuation of hierarchies is often related to subconscious discrimination, such that attempting to separate the two problems risks oversimplification.")

[113] This interpretation, as well as the one offered below involving balancing—both of which are sketched here in somewhat caricatured form—correspond to competing understandings that have been articulated in prior work. *See, e.g.*, Primus, *supra* note , at 498–99 ("To oversimplify for the moment, some readings of the prohibition on disparate impact see it as an evidentiary device aimed at ferreting out present discriminatory states of mind, while others see it as concerned with the lingering structural consequences of discrimination practiced in the past." (footnotes omitted)).

[114] Often this point will not hold as a practical matter, depending on how the statement is interpreted. If, say, a job requirement is only slightly job-related but very substantially reduces the eligible pool, then the employer would lose more from the pool reduction than it gains from the productivity advantage. Hence, a profit-maximizing employer would not add a job requirement unless it had a significantly greater productivity effect. (This simple point is obvious in any hiring process; myriad factors that are plausibly job-related are not made requirements for application, only those that are particularly significant.) These points are better viewed from the perspective presented below, wherein one can imagine an employer running a regression that takes into account all relevant factors and assigns each factor a weight that corresponds to its influence on marginal productivity.

[115] Employers may often be unaware of the actual consequences, particularly if both B and H are small. Moreover, the case in which $B = 0$ is unresolved, in that it suggests employer indifference. However, when there is a

Note that, under this view, there is no comparison of H and B. Or, in the spirit of the implicit rationale, what we have been calling H is not really taken to be a measure of social harm as such. Rather, the first step indicates whether or not there is discrimination (a dichotomous inquiry), and, if there is, only that which has no productivity justification is deemed to be impermissible.

At the opposite end of the spectrum, one could emphasize "business necessity" and interpret this language in a literal fashion that sets B^* extremely high—how high depending on the business. For example, if some huge corporation has billions of dollars in annual profits, then any practice that did not reduce productivity by more than that—pushing the company into bankruptcy, let's say—is not strictly "necessary" and hence does not meet the dictates of step 2. Accordingly, in this example, billions might have to be spent to hire one additional minority employee. Of course, at smaller and less profitable entities, the demand would be lighter, but still possibly large. Although linguistically plausible, this interpretation does not seem to have gained traction, either in particular statements or in practice.

There is a great chasm between this "necessity" interpretation under which B^* may be huge and the preceding "job related" interpretation under which $B^* = 0$. For any $B^* > 0$, the analysis in subsection I.A.1 is applicable. Of particular interest is how B^* relates to H^* and, even more so, how it relates to the H established in step 1. Although renditions of disparate impact's second step do not ordinarily refer to step 1, it is entirely imaginable that, when reaching step 2, a judge who has heard evidence on H and already reached a decision on step 1 will have that H in mind when applying step 2's ambiguous standard. Such a linkage, in turn, might result in a mode of decisionmaking that involves at least an implicit comparison of H and B, tantamount to balancing.

More precisely, if $B^* = H$, then step 2's test of whether $B > B^*$ is equivalent to asking whether $B > H$. Moreover, if H is taken to constitute a measure of social harm, this would make sense. As we know from subsection I.A.1, if B^* is instead taken to exceed H, then we may (under this view) mistakenly assign liability in some cases: specifically, when $H < B \leq B^*$. And if B^* is taken to be less than H, then we may mistakenly fail to assign liability in some cases: when $H > B > B^*$. Interestingly, this latter sort of error at step 2 did not arise in subsection I.A.1's examination of our stylized structured decision procedure because step 3 came to the rescue. That is, when step 2 passed in such a case, the result was not to assign no liability and stop, but rather to proceed to step 3's balancing; in that instance, step 2 was rendered redundant. Here (setting to the side for the moment consideration of less restrictive alternatives), we do not have that saving feature, so setting $B^* < H$ is problematic under the currently contemplated view of the law's purpose.

Let us now reflect on what this interpretation of step 2—wherein B^* is implicitly set equal to H (or at least is positively related to H[116])—means with regard to the purpose of Title VII disparate impact doctrine. Here, society is willing to incur some cost in order to avoid

nontrivial negative impact on the size of the pool, there usually would be a cost to the employer (although a proper notion of B would be defined net of such effects). *See supra* note .

[116] To elaborate, much of what is discussed in the text to follow applies to any $B^* > 0$, because then there is some tradeoff admitted between productivity costs to the employer and avoiding disparate impact on minority employees. Moreover, any comparison between H and B allows a balancing interpretation. Because the units of H and B are not, on their face, comparable, some translation is required. In this respect, the key question is whether the (implicit) B^* is increasing in the court's estimate of H from step 1. If it is, then "as if" balancing is involved. *See* Kaplow, *supra* note , at 1049–55. As noted, on its face step 2's test makes no reference to the degree of disparate impact, H, so a literal interpretation, even supposing that we do not take B^* to be 0 or huge, is that it is at some intermediate level, independent of H, which possibility was considered in the preceding paragraph.

disparate impact as such. Under the first view noted above, it was mentioned that productivity gains of a few dollars per employee would be seen as sufficient justification for hiring 1000 fewer minority workers. Under the "necessity" view, even billions of dollars might have to be sacrificed in order to hire one more. Now, we are contemplating an intermediate position, which posits a tradeoff between avoiding disparate impact and requiring employers to incur costs in terms of reduced productivity. Such a view of the doctrine's purpose might be associated with language in *Griggs* about barriers that "'freeze' the status quo"[117] by creating "'built-in headwinds' for minority groups."[118] Note that, even if the doctrine does not formally admit any dependence of the stringency of step 2 (here, taken to be embodied in *B**) on the magnitude of what is found in step 1 (regarding *H*), it seems plausible that a judge who embraced this view would at least to some degree be influenced by *H*, being more inclined to find that step 2 passed when step 1 was a close call and less inclined to find step 2 satisfied when the disparate impact found in step 1 was large.[119] And it also suggests, as mentioned above, that a judge deciding a close case at step 1 may be influenced by a sense of the *B* that would be shown at step 2, particularly to the extent that information on *B* is already in hand.

In examining different possibilities regarding the magnitude of the *B** threshold in step 2, we can see that quite different notions of the purposes of disparate impact doctrine are brought into focus. It seems possible that part of the ambiguity on the meaning of the enigmatic requirement of job relatedness and business necessity reflects a reluctance to address contentious issues openly. If step 2 really does embody balancing, then the need to make statements about how much *B* must be sacrificed to reduce *H* may be uncomfortable. In any event, whatever may be the conceptual, practical, and political challenges,[120] once any tradeoff is admitted, consistency in decisionmaking requires that decisions be made *as if* balancing is undertaken.[121]

Turn now to the final step, under which a defendant who has met its burden under step 2 (whatever that may be) may nevertheless lose if there exist alternative employment practices, understood to refer to less restrictive alternatives that have less (or no) disparate impact but nevertheless generate *B*. Current doctrine does not seem to be entirely clear on whether equal effectiveness is required.[122] In this context, the appropriate understanding of this step in

[117] 401 U.S. at 430 (emphasis omitted).

[118] *Id*. at 432; *see, e.g.*, Primus, *supra* note , at 523–25.

[119] *See, e.g.*, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , at 3-41 ("Some courts have suggested that the degree of disparate impact caused by the challenged practice can affect the sufficiency of the employer's proof of justification."); *id*. at 3-47 ("Courts have held that the greater the disparate impact of a test, the higher the correlation required, and vice versa. Similarly, as the disparate impact increases, a stronger showing is necessary of the importance of the criterion to successful job performance." (footnotes omitted)); *id*. at 4-70 ("[I]f the employer's practices result in a high degree of exclusion and have a low degree of business utility, they are more likely to be found unlawful; if they result in a low degree of exclusion and have a high degree of business utility, they are likely to be found lawful.").

[120] *See* Kaplow, *supra* note , at 1047–55 (elaborating many of the challenges and how they are best understood and addressed). The latter reference is meant to suggest a possible reluctance to articulate balancing explicitly even when undertaken. In that regard, there is the further possibility that the balancing interpretation of disparate impact law may raise constitutional problems of the sort associated with affirmative action, which is the focus of Primus, *supra* note .

[121] *See supra* note (discussing how this perspective applies to any intermediate choice of *B** as long as the critical level rises with the *H* found at step 1). There is also a cost-effectiveness interpretation of balancing: for a given aggregate cost to employers, consistent decisionmaking embodied in "as if" balancing maximizes the hiring of minorities; conversely, a given reduction in aggregate disparate impact will be implemented at the lowest cost to overall productivity under a consistent balancing framework.

[122] For example, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , states: "To rebut the employer's proof of business necessity, a plaintiff can demonstrate that the employer refused to implement an *effective* alternative practice or selection device that would have a lesser disparate impact." *Id*. at 3-13 (emphasis added). The question is how "effective" the practice must be. *See also id*. at 3-45 ("Plaintiffs *may* also have to show that the proposed alternative is *substantially equally* valid." (emphasis added)). The treatise SULLIVAN & WALTER, *supra* note , at 294, states that "courts that have

principle depends on B^* in step 2 and the associated rationale for disparate impact doctrine. Put conversely, a stand on whether equal effectiveness is required has implications for step 2's content and the doctrine's purpose.

Suppose first that $B^* = 0$, associated with the view that step 2's job-relatedness requirement is satisfied by any, even fairly minimal, productivity boost associated with the challenged employment practice. In that case, consistency seems to imply that an alternative employment practice must be equally effective. If it is not, then its adoption would be associated with a reduction in B. Put another way, suppose that an employer had initially utilized the proffered alternative employment practice and subsequently switched to the practice being challenged in the case at hand. By hypothesis, this switch would generate a positive B and hence, if $B^* = 0$, the new practice would be justified given the underlying standard.

Now suppose instead that $B^* = H$, associated with the implicit balancing view. In that event, subsection I.A.2 tells us that the less restrictive alternative is properly analyzed under the second balancing test or, equivalently, the delta/delta test. Accordingly, equal effectiveness is a sufficient condition for the plaintiff to prevail at this step but not a necessary one. Indeed, the case of equal effectiveness has no special significance except that in some situations it may be clear that such a requirement is met, making the decision (to assign liability) easy.

The focus thus far has been on the disparate impact structured decision procedure as a final decision rule. Turn now to how it performs with regard to the collection and analysis of information, supposing that it actually were to serve as the information protocol, largely bracketing for the moment the actual conduct of U.S. civil litigation. That is, let us apply the analysis developed in subsection I.B.1 to the present context.

Here, we again have a sequentially siloed regimen that first considers all information on H, only then turning to B. As always, even if information did naturally clump separately by issue, this sequencing tends not to be optimal. Most obviously, under the first view of step 2, under which $B^* = 0$, evidence that $B > 0$ would suggest that efforts can be terminated early—

addressed the issue have required equal efficacy," but the offered support, *see id.* at 294 n.236, consists of two cases, the first a dissenting opinion (a feature not noted) that uses the language "equally effective" in passing (and in turn cites for support a case that does not contain such language), and a second that offers dicta that merely says that the practice "would also serve the employer's legitimate business interest" (and cites in support *Albemarle*, which language likewise does not specifically indicate whether the alternative must be equally effective). Nevertheless, a number of courts require equal effectiveness, although a small additional cost or inconvenience is often ignored in undertaking this assessment, and the "equality" demand is sometimes relaxed or modified in other ways. *See, e.g.*, Lopez v. Lawrence, 823 F.3d 102, 111 (1st Cir. 2016) (stating step 3 as: "do the plaintiffs show that the employer has refused to adopt an alternative practice that *equally or better* serves the employer's legitimate business needs, yet has a lesser disparate impact?" (emphasis added)); *id.* at 119 (referring to "equal or greater validity" and an "equally or more valid test"); Jones v. Boston, 845 F.3d 28, 35 n.3 (1st Cir. 2016) ("Similarly, while it may be within the scope of inquiry to consider the putative costs of the Officers' proposed alternative, . . . a reasonable jury could find that there would have been no *material* cost differential . . . ." (emphasis added)); Johnson v. Memphis, 770 F.3d 464, 472 (6th Cir. 2014) ("the plaintiff must demonstrate: (1) the availability of alternative procedures that serve the employer's legitimate interests and (2) produce '*substantially* equally valid' results, but with (3) less discriminatory outcomes" (emphasis added)) (citing the 1978 EEOC Guidelines, the source of the aforementioned four-fifths rule, for the quoted language); Allen v. Chicago, 351 F.3d 306, 314 (7th Cir. 2003) (requiring that those promoted under the proposed alternative be "*substantially* equally qualified" (emphasis added)). The Supreme Court's fairly recent statement of this requirement in *Ricci v. DeStefano*, 557 U.S. 557, 578 (2009), however, is more murky, referring to "an available alternative employment practice that has less disparate impact and serves the employer's legitimate needs." This ambiguity in the 1991 Act is attributed to the unusual legislative history in which compromise required an explicit refusal to address key issues. As indicated at the outset of this part, *see supra* note , the Act specifically refers to preexisting case law on the question, but it was understood that those cases did not address the matter. *See, e.g.*, LINDEMANN, GROSSMAN & WEIRICH, *supra* note , at 3-42 to -46; SULLIVAN & WALTER, *supra* note , at 292 (emphasizing that "the phrase *alternative employment practice* did not exist in any Supreme Court case before it was used in *Wards Cove*," making it odd that the statute limits interpretation of the term to pre-*Wards Cove* caselaw).

subject to both confirmation and the possibility that the defense might be negated by an equally effective alternative employment practice. (Otherwise, one might struggle interminably and needlessly about the plaintiff's statistical demonstration at step 1.) Under the balancing view, it is by now familiar that the appropriate means of gathering and analyzing information may well involve alternation, depending on the diagnosticity/cost ratio of the information that remains to be collected.[123] This point was alluded to above in discussing the decision rule, when noting that a judge assessing step 1 might naturally peek ahead at *B* and that one deciding step 2 may have *H* in mind when considering the appropriate threshold.

In addition, information often clumps by source rather than by issue, so there can be significant synergy loss in sequencing—or, if information is collected by source, decision precision is needlessly sacrificed if information on *B* is in hand but ignored. Moreover, much information by its nature pertains to both *H* and *B*, rendering separation incoherent. Consider some illustrations of these general points in the present context.

To begin, whether or not view one prevails (under which there is a sort of implicit, objective intent inquiry), intent will be probative and inevitably links the two issues in a now-familiar manner.[124] The more it appears that $B = 0$, the more suspicious we may be that a discriminatory purpose motivates the employment practice, making step 1's $H > H^*$ test more likely to be satisfied. And a very low *H* suggests an alternative explanation, which may well involve a positive (and larger) *B*.

It is also true that much of the evidence and the methodology utilized in establishing a plaintiff's statistical demonstration of disparate impact at step 1 is intertwined with business justifications that are supposedly not reached until step 2. Consider the familiar point that a plaintiff in a discriminatory hiring challenge needs to make its demonstration regarding *H*, at step 1, using an appropriate definition of the employment pool. Including children would be inappropriate, as would including unlicensed individuals for an occupation that requires a license (such as an airline pilot).[125] Defendants often argue at step 1 that the plaintiff's pool is too broad,

[123] Paralleling the discussion in section II.B regarding the difficulty of proving efficiencies that may be generated by mergers, in the employment context it is often difficult—and, of relevance, more difficult than showing disparate impact—to assess the contribution of various traits to productivity. For example, what evidence could an employer who failed even to interview a mere high school graduate for a position as a research scientist point to in demonstrating job relatedness? A prior randomized controlled trial in which it hired a large sample of such individuals who then failed to produce substantial results? Even considering less extreme examples, suppose that a factor in determining bonuses or promotions is an employee's degree of cooperation with others. Validating the contribution to productivity of different degrees of cooperativeness surely is not easy. A great challenge with disparate impact cases concerns the question of what should count as sufficient validation. On one hand, myriad factors that plausibly relate to productivity are routinely taken into account in screening resumes, interviewing employees, setting bonuses, and making promotion decisions. If any that were not rigorously validated were impermissible (that is, if they had disparate impact, which many probably do), the costs to productivity could be serious. On the other hand, it is well known that "we've always done it that way" can be a conscious or subconscious cover for discrimination and, in any event, longstanding practices with no discriminatory motivation can survive despite possibly substantial contradictory evidence that could readily be collected but is not. Consider MICHAEL LEWIS, MONEYBALL: THE ART OF WINNING AN UNFAIR GAME (2003). Related lessons also appear throughout PHILIP E. TETLOCK & DAN GARDNER, SUPERFORECASTING: THE ART AND SCIENCE OF PREDICTION (2015), and in the chapter on when we can trust experts in DANIEL KAHNEMAN, THINKING, FAST AND SLOW, ch. 22 (2011).

[124] Indeed, in proclaiming that intent is not the test, *Griggs* stated: "We do not suggest that either the District Court or the Court of Appeals erred in examining the employer's intent . . . ." Griggs v. Duke Power Co., 401 U.S. 424, 431.

[125] *Compare* Carpenter v. Boeing Co., 456 F.3d 1183, 1197 (10th Cir. 2006) (requiring that the plaintiff's statistical analysis be "based on data restricted to qualified employees" or that, if such data be unavailable, based on a reliable proxy instead), *with* Stagi v. Nat'l R.R. Passenger Corp., 391 Fed.Appx. 133, 147–48 (3d Cir. 2010) (requiring that additional factors that might explain the disparate impact be advanced as part of the defendant's demonstration of business justification, in step 2). *See generally* Anderson v. Zubieta, 180 F.3d 329, 342 (D.C. Cir. 1999) ("It is true that in

for example, by omitting some qualification such as a high school education, whereas the plaintiff's challenge may be that this very qualification is the cause of the disparate impact. As some important literature has explained, adding a variable to the pertinent regression equation tends to reduce the coefficient on, say, race—moving the disparate impact toward zero and potentially nullifying its statistical significance—precisely when the added variable (a high school education, in our example) itself has a disparate impact on race.[126] Adding this variable would be correct if it was job related but would actually confirm the plaintiff's case if it was not. Yet job-relatedness (*B*) is a question that is supposed to be deferred to step 2.

As a heuristic for thinking about step 1's demonstration regarding *H* (which involves notable oversimplification[127]), it is helpful to consider a regression equation—for hiring, wages, or promotions, as the case may be—in which all possible explanatory variables are included.[128] One can then ask how the magnitudes of the resulting coefficients differ—if at all, and in a statistically significant way[129]—from the corresponding magnitudes in a similar regression equation where what is "explained" (the dependent variable) is workers' productivity, normally

order to eliminate the most common nondiscriminatory explanation for a disparity—lack of qualifications—a plaintiff's prima facie case must take into account the 'minimum objective qualifications' for the position at issue. . . . But that does not mean a plaintiff must take account of every qualification recited by the employer, nor even of every 'objective' qualification. Rather, what the case law means by 'minimum objective qualifications' are those objective qualifications that can be shown to be truly required to do the job at issue."); Wax, *supra* note , at 630–31 ("Threshold requirements of any kind can end up screening out minority applicants. Thus using such requirements to define potential candidates is itself vulnerable to challenge under the disparate impact rule. Unfortunately, the Supreme Court has set no clear standard for identifying the population against which workplace disparate impact should be assessed and the lower courts vary in their approach. This aspect of disparate impact doctrine is in serious disarray." (footnote omitted)).

[126] The most extensive legal treatment of this idea is Ian Ayres, Testing for Discrimination and the Problem of "Included Variable Bias" (2010) (unpublished manuscript), https://www.law.upenn.edu/live/files/1138-ayresincludedvariablebiaspdf (also featuring the *Griggs* high school diploma requirement as an example). *See id.* at 16 (heading one section "A 'Business-Justification' Approach to Disparate-Impact Testing"); *id.* at 33 ("Regardless of what substantive standard is adopted for determining what qualifies as a business justification (and hence what should be included in the unjustified [disparate impact regression] specification), the application of the standard will turn on facts and or reasoning that are external to the regression itself."). A preceding line of economics literature developing this and related ideas includes Burton G. Malkiel & Judith A. Malkiel, *Male-Female Pay Differentials in Professional Employment*, 63 AM. ECON. REV. 693 (1973); David E. Bloom & Mark R. Killingsworth, *Pay Discrimination Research and Litigation: The Use of Regression*, 21 INDUS. REL. 318 (1982); Mark R. Killingsworth, *Analyzing Employment Discrimination: From the Seminar Room to the Courtroom*, 83 AM. ECON. REV. PAPERS & PROC. 67 (1993); John Yinger, *Evidence on Discrimination in Consumer Markets*, J. ECON. PERSP., Spring 1998, at 23, 26–29. (This literature features a further subtlety concerning endogeneity: for example, even if men and women in a given job classification receive equal pay, if there is differential hiring or promotion into that classification, that difference might constitute discrimination.) The interdependency idea is also briefly but sharply stated by Michael Piette, *supra* note , at 254 ("Lastly, and perhaps most important under the heading of legitimacy, is the problem of tainted independent variables. Suppose a regression analysis includes a variable for education that, in a race case, is a key determinant of salary differences between black and white employees in a clearly different job group. Regression analysis indicates a high t-statistic on education and an insignificant t-statistic on the race coefficient. Given that in almost all groups, white employees have received more formal education than black employees, it would appear that education goes a long way towards explaining salary differences between black and white employees. The burden is on the employer, however, to demonstrate separate from the regression, that education was required and affected performance, and hence directly determined salary. To the extent that education is not related to job performance, it is an inappropriate variable to use in a regression. Excluding key variables and including irrelevant variables have the same impact.") (quoted in Ayres, *supra*, at 22).

[127] Among other things, this presentation abstracts from considerations of functional form, interactions among independent variables, and endogeneity.

[128] The discussion that follows focuses on steps 1 and 2, abstracting from alternative employment practices. In principle, they could be incorporated as well, in a manner that reflects the previous discussion of this final step and how the analysis depends on the view of the purpose of the doctrine and correspondingly the interpretation of step 2.

[129] *See supra* note (discussing issues concerning the role that statistical significance plays in disparate impact cases).

associated with step 2's assessment of *B*.[130] Disparate impact might be said to arise (notably, under view one) when a variable is given different weight in the employer's decision, say, on hiring, than its effect on productivity, and in a direction that indicates the pertinent disparate impact.[131] For example, giving more weight to a high school education than is indicated by its contribution to productivity, when the disadvantaged class of prospective employees is less likely to have a high school education, would indicate a disparate impact.

In this formulation, disparate impact is indicated by the *difference* between the weight of the challenged practice in predicting employment outcomes (such as hiring decisions) and the weight it has in predicting productivity.[132] (The corresponding statement under the balancing view appears in the margin.[133]) We thus have a setting in which the interrelationship between *H* and *B* in determining liability is central. One could sequence the analysis of the two regression equations, looking first at the employment regression and then the productivity regression. But the first, standing alone, indicates disparate impact in only the most minimal fashion. Step 1 would be satisfied in principle whenever an employer gave weight to any factor correlated with the pertinent status in a direction that produces disadvantage: for example, an employer requiring a high school diploma for the position of research scientist, or work experience to be considered for a managerial position. One suspects that, in practice, a judge would likely find such cases unconvincing at step 1, which is to say that there may be some sort of implicit look-ahead to step 2. The notion that one might exercise judgment in this regard—taking into account more or less information (or, when fairly obvious, something more akin to judicial notice) regarding *B*, relying on hunches about what is plausible and the ease of assessing the matter (the

[130] The presentation in the text (a "heuristic") is strictly a thought experiment. Actually implementing the productivity regression in particular would be a daunting task. *See supra* note (on the challenges of demonstrating various traits' contributions to productivity).

[131] The suggestion in Ayres, *supra* note , at 34–36, that one might "cap" a coefficient in the employment regression is in similar spirit. An issue, however, is that Ayres's suggestions focus on the plaintiff's prima facie case, at step 1, whereas the evidence required to assess the cap, which is a measure of the productivity of the factor, does not formally come up until step 2. Of course, this reinforces the broader point, here and in Ayres, that business justification (*B*) is intertwined with disparate impact (*H*). Relatedly, Ayres argues more broadly that "disparate impact tests should only include controls for attributes that are plausibly business justified." *Id.* at 5–6 (emphasis omitted); *see id.* at 17. As explained in the text to follow, this sort of suggestion is problematic. The focus is on step 1. Suppose there that, when the variable in question is added to the regression, it eliminates the impact of race, so the plaintiff's demonstration of disparate impact seems to fail. If the variable being *plausibly* job related is sufficient, then the plaintiff loses under the formal doctrine, even if it can shown (and may already have been shown, as noted below) that this plausible justification, when examined closely (including rebuttal evidence) fails. This point suggests a correction to Ayres's formulation, in which the step 1 assessment includes essentially the full step 2 inquiry as well. Much of Ayres's analysis suggests that he well appreciates this point, but is perhaps reluctant to advance an implication that, however logical, seems to sharply conflict with the doctrine. Hence, this fudge. (As mentioned in the text to follow, it may well be that judges use a similar fudge, but, particularly when deciding a case at the end of a trial, with a full appreciation of what has been proved regarding *B*.)

[132] A feature of this regression-based presentation is that it illustrates how, in principle, the question should not be understood as whether the employer *considers* a factor (that has a differential incidence on the minority group) that, say, under view one, is *entirely unpredictive*, but rather whether the employer *gives more weight* to the factor than is appropriate in light of its impact on productivity. If a factor is measured, say, by a test, it may be that the test is a sensible way to measure the pertinent factor, but that too much weight is placed on the test. One might say that the alternative employment practice indicating the appropriateness of assigning liability would be the use of the very same test but giving it less weight in the decision. *See supra* note (discussing pool size and the weight given to various factors in hiring).

[133] Under the balancing view, one would be concerned instead with the corresponding ratio. In the text's example, the question would be whether the relative contribution to productivity was sufficient to justify the resulting degree of disparate impact, a test that could fail (resulting in liability) even if there was no disparate impact under the first view. (If there is disparate impact under the first view, there will be under the second, a fortiori. A corresponding way to put this point is that, once we know that $H > 0$ and, moreover, that $B \leq 0$, there can be no doubt that $H > B$.) Note further that, when the relevant test involves, in essence, a ratio, it is incoherent to engage in a separate, sequential analysis that asks, first, how large is the numerator, without regard to what the denominator might be.

diagnosticity/cost ratio)—is more in the spirit of optimal information collection than is an attempt to adhere rigidly to the sequential separation apparently mandated by the doctrine. Doing so informally and against the grain of a structured decision procedure can generate confusion and error while also reducing transparency.

To close, consider more explicitly the question raised in subsection I.B.2 concerning the extent to which litigation conforms either to the dictates of optimal information collection or to the protocol reflected in existing doctrine.[134] Absent explicit bifurcation that begins at the outset of a case, this sequenced rule does not economize on discovery costs.[135] As already noted, step 1 itself is pertinent mainly at summary judgment and at the end of trial (or in a motion for judgment as a matter of law at the close of the plaintiff's case). And even at summary judgment, the question is only whether there is a genuine dispute as to disparate impact, *H*, although in practice this can still generate terminations, notably when a plaintiff expert's statistical demonstration falls notably short.[136] In accord with the foregoing, such assessments of step 1, whether at summary judgment or later, are difficult to disentangle from step 2's test regarding *B*. Furthermore, some or all of the information pertaining to the latter may be available when deciding on step 1. Another interesting feature is that, since disparate impact cases involve bench trials, we have the interesting situation in which a judge is being to asked at summary judgement whether a reasonable factfinder—him- or herself—might reasonably find for the plaintiff on the issue. It may thus seem less surprising, and less inappropriate, for a judge to be more venturesome in resolving issues at this stage (including by peeking ahead at information regarding *B*, which is strictly part of step 2).[137] At the end of the plaintiff's case at trial, the situation is similar.[138]

At the conclusion of a trial, the only consequence of adhering to the structured rule, if it has nontrivial thresholds, is to sometimes generate the wrong outcome even given the evidence

[134] Class certification questions, of particular relevance in Title VII disparate impact litigation, are set to the side here (as they are throughout this article).

[135] For example, in a recent decision reviewing a grant of summary judgement for the defendant on step 1, the court made clear that essentially all of the information pertaining to step 2 and step 3 had already been adduced through the process of discovery. *See* Jones v. Boston, 752 F.3d 38, 55 & n.19 (1st Cir. 2014) ("In view of the size of the record, though, and the fact that the district court judge who has presided over this case has not yet parsed that record to assess business necessity or its rejoinder, we decline to do so in the first instance. . . . In declining to decide the issues in the first instance, we do not suggest that the district court must reopen the record to allow further discovery or expert reports. The district court retains its customary discretion to manage the case, and we expect that it will give due weight to the fact that each party has already had ample time to put its best foot forward.").

[136] In disparate impact cases, it is also plausible for a defendant to win at summary judgment on step 2—and without deciding (or even rejecting) a defendant's claim that step 1 fails, which is to say, taking the issue out of order—particularly when B^* is taken to be essentially zero, as it is under the first view. The reason is that a defendant's purported business justification, even when aggressively challenged, may clearly survive at least somewhat (that is, even granting the plausible rebuttal). When $B^* = 0$, this would be enough (assuming further that there is no genuine dispute about a possible alternative employment practice).

[137] Suppose, for example, that a judge is unsure whether a plaintiff's regression results supporting disparate impact should be accepted in light of a defendant's claim that the equation improperly excluded some variable that, when included, eviscerates the result of significant discrimination. The prior discussion suggests that such a judge may well be influenced by how powerful is the defendant's showing (and plaintiff's rebuttal thereof) regarding whether that variable substantially relates to job performance or does not at all—to the extent that much of this material on *B* is already in the record

[138] Confined to the trial itself, the judge will only have heard the plaintiff's case and not the defendant's rebuttal (except via cross-examination), whereas, at summary judgment, the judge will typically have been presented with the defense expert's report and other information to provide the context for the defendant's argument that the plaintiff's evidence is insufficient to create a genuine dispute regarding *H*, in step 1.

before the judge.[139] Yet, as mentioned, one suspects that a judge would have at least some tendency to look ahead, at evidence on *B*, when thinking about how to decide step 1, formally limited to *H*, and to keep *H* in mind if reaching step 2 and contemplating what should implicitly be taken as *B** when deciding whether the evidence on *B* is sufficient to satisfy step 2. Even with no explicit balancing step, one might expect de facto balancing—that is, if the judge implicitly adopts the balancing view rather than view one, under which $B > 0$ is sufficient to assign no liability (after accounting for alternative employment practices). Given the law's ambiguity concerning the magnitude of the thresholds *H** and *B**, it does not seem difficult to craft such an opinion in a manner that is consistent with the official doctrine and makes no mention of balancing. And a judge may be inclined to behave in these ways even without thinking explicitly in balancing terms.[140]

IV. CONSTITUTIONAL LAW

A. *Strict Scrutiny*

Strict scrutiny in U.S. constitutional law employs a structured decision procedure that bears some resemblance to those considered in this article.[141] First, it asks whether the challenged government action infringes a qualifying fundamental right or involves a suspect classification. If not, the strict scrutiny inquiry ends (although analysis may proceed under lower tiers of review, which are not examined here).[142] Second, if it does, the government must demonstrate that the infringement can be justified by the advancement of a compelling interest. If not, the government loses.[143] Third, if it can, then there is an assessment of whether the government's action is narrowly tailored to the justification. If so, the government action stands, but if not, it is invalid.

[139] In the present context, like those considered earlier in this article, we can also see that there is no real sense in which the burden shifts during the trial (or earlier on); even a ruling for the plaintiff on a defendant's motion for judgment as a matter of law at the end of the plaintiff's case only indicates that the plaintiff has enough evidence on *H* to create a genuine dispute, not that the plaintiff has succeeded on step 1, which determination must await the defendant's direct rebuttal in any event. (One could imagine a defendant beginning its own case by first presenting all of its direct rebuttal on *H*, before introducing its evidence on *B*, and then requesting a final decision on step 1. This, in turn, would suggest that any further rebuttal from the plaintiff be presented at that time. It may indeed sometimes be efficient to employ bifurcation, although interdependencies between *H* and *B*, which as explained can be quite important in the Title VII disparate impact context, militate against such bifurcation.)

[140] The difference between the conscious construction and the subconscious or implicit motivation constitutes another respect in which "as if" balancing may be said to occur.

[141] *See, e.g.*, ERWIN CHEMERINSKY, CONSTITUTIONAL LAW: PRINCIPLES AND POLICIES 567 (5th ed. 2015); Richard H. Fallon, Jr., *Strict Judicial Scrutiny*, 54 UCLA L. REV. 1267, 1315–16 (2007) ("However the purposes of strict scrutiny are characterized, there are three crucial steps in applying the formula: (1) identifying the preferred or fundamental rights the infringement of which triggers strict scrutiny; (2) determining which governmental interests count as compelling; and (3) giving content to the requirement of narrow tailoring.").

[142] Discussion below of steps 1 and 2 will return to the relationship between tiers of review and the interpretation of strict scrutiny's structured decision procedure. Although this section analyzes only strict scrutiny, some of what is said bears on the lower tiers of review.

[143] Decades ago, Gerald Gunther famously stated that, in light of the stringency of actual application, this test was "strict in theory and fatal in fact." Gerald Gunther, *The Supreme Court, 1971 Term—Foreword, In Search of Evolving Doctrine on a Changing Court: A Model for a Newer Equal Protection*, 86 HARV. L. REV. 1, 8 (1972). Subsequent decisions, particularly involving affirmative action, suggest a more moderate view. *See generally* Fallon, *supra* note , at 1303–05 (discussing the notion that strict scrutiny may be understood as entailing a prohibition of any infringement short of an interest in averting catastrophic consequences).

Let us now relate this protocol to this article's framework. Step 1 asks whether the infringement, which causes an *H*, qualifies by reference to some *H**. Step 2 considers whether the government's interest, *B*, in pursuing its action is sufficiently compelling to justify the infringement: is $B > B^*$? Step 3 is a species of less restrictive alternatives analysis (and is frequently discussed in those terms). On its face, there is no balancing step. In these latter respects, this formulation may seem closest to the rule just considered for disparate impact cases under Title VII.

When we examine strict scrutiny's three steps more closely, further similarities with and differences from this article's stylized three-step decision procedure emerge. Step 1 poses the most striking contrast. Until now, the $H > H^*$ inquiry involved primarily a quantitative assessment.[144] For strict scrutiny, at least on its face, step 1 seems to be qualitative and, in particular, categorical. That is, there appears to be a gradually evolving list of rights or interests, the infringement of which is deemed to count, which is to say, to trigger strict scrutiny.[145] For them, there seems to be an on/off inquiry as to whether they are infringed. For those not on the list, strict scrutiny is not triggered, no matter how great the infringement. Under this interpretation, *H** is not a quantitative threshold but rather constitutes some *set*, and the test is whether the *H* involves a type of harm that is an element in that set.[146] Subsection I.A.1 explained how, as a final decision rule, a quantitative trigger was undesirable because insisting that $H > H^*$ (with *H** a quantitative threshold) implied that there would be no liability even in some cases in which $H > B$—specifically, those in which $H^* \geq H > B$.

Use of a categorical trigger might be motivated by the considerations noted in subsection I.A.2 on rules designed to constrain balancing, which, as mentioned, are particularly pertinent to some constitutional questions. Specifically, there exist concerns that government actors might undervalue or even negatively value certain rights when the actors are motivated to entrench themselves or certain minorities are underrepresented or despised.[147] There is an obvious connection with many of strict scrutiny's triggering rights: freedom of speech and the press, particularly regarding political speech, connect to the former concern; many forms of discrimination relate to the latter.

In this regard, the categorical trigger can be regarded as jurisdictional: there is a need for close court review when the government acts in ways that raise questions about whether it can be trusted. Although the government decisionmakers being second-guessed typically have greater institutional competence in many respects, there are concerns that some actions may reflect improper motives rather than superior expertise.[148] Subsection I.A.2 notes, however, that this

[144] Recall that, in Part III on Title VII disparate impact, there was some ambiguity regarding whether this referred to the magnitude of the disparate impact or the degree of confidence with which *some* disparate impact could be established (and it was further noted how the latter, in practice, is related to the former).

[145] *See, e.g.*, CHEMERINSKY, *supra* note , at 567; Fallon, *supra* note , at 1268–69, 1316–21.

[146] One might also interpret step 1 as including not only the task in a typical case of checking whether the allegedly infringed right is on the preexisting list but also the larger endeavor of deciding which rights are on the list and defining or adjusting various boundaries.

[147] Analogously, one could ask why it makes sense to have courts review private actors' decisions in antitrust and Title VII. The answer is primarily that private actors are ordinarily trusted to advance their own interests, and the laws in question are designed to curb private self-interest when it comes at the expense of a greater social interest (that private actors do not credit on account of externalities; that is, the harm in question is to others, and the market does not operate in a fashion that is believed to result in the private actors' indirectly bearing the harm).

[148] *See, e.g.*, Alexander Aleinikoff, *Constitutional Law in the Age of Balancing*, 96 YALE L.J. 943, 984–85 (1987) (explaining, in spite of his generally critical stance toward balancing by courts in constitutional cases, that: "A better argument for the balancer is that the Court improves the balancing process by giving weight to interests that the legislature tends to ignore or undervalue. . . . First, it reinforces representation, ensuring that the interests of unpopular or underrepresented groups are counted and counted fairly. Second, it protects constitutional rights and interests that are

justification for judicial supervision does not necessarily imply that, once undertaken, it should proceed other than by balancing. If balancing makes sense in principle, and if moreover we trust the courts,[149] then judicial review of other, less trusted government officials may most sensibly be done through balancing once review is triggered.[150] The consideration of how strict scrutiny's step 2 and step 3 operate will address the extent to which that is done and whether any deviations may be justified by the concerns giving rise to judicial review in the first place.

Even if step 1 is understood in largely categorical terms, we should consider whether it might also be taken to have a quantitative dimension (a point that will loom large in any event when we consider step 2 and step 3). First, the magnitude of the typical level of *H* could influence the tier of scrutiny that is applied.[151] To the extent that it does, we have *H*s of greater magnitude tending to require stronger justifications (levels of *B*) in order to be warranted. Looking ahead to step 2, even if *B** were constant within a given tier, notably, strict scrutiny, if a higher tier has a larger *B**, then it is as if higher *H*s can only be offset with greater *B*s. This phenomenon is akin to a crude sort of balancing—more like a step function than a continuous tradeoff.[152] Note that if one made the further move of taking a sliding-scale approach to intermediate scrutiny,[153] then that step function might be partially smoothed.

sometimes forgotten in the hurly-burly of politics."); Richard H. Fallon, Jr., *The Supreme Court, 1996 Term, Foreword: Implementing the Constitution*, 111 HARV. L. REV. 54, 76 (1997) ("In the face of reasonable disagreement among the citizenry and between courts and legislatures, the two-tiered framework produced by the conjunction of suspect- and nonsuspect-content tests manifests a judicial aspiration to trust institutions of political democracy except in circumstances in which the democratic process is manifestly untrustworthy."). This interpretation is consonant with the use of a highly deferential rational basis test in settings in which there is not deemed to be sufficient basis for distrust of the primary government decisionmaker's actions.

This overall orientation is associated with footnote 4 in *United States v. Carolene Prods. Co.*, 304 U.S. 144, 152 n.4 (1983): "It is unnecessary to consider now whether legislation which restricts those political processes which can ordinarily be expected to bring about repeal of undesirable legislation is to be subjected to more exacting judicial scrutiny under the general prohibitions of the Fourteenth Amendment than are most other types of legislation. . . . Nor need we enquire whether similar considerations enter into the review of statutes directed at particular religious . . . , or national . . . , or racial minorities . . . : whether prejudice against discrete and insular minorities may be a special condition, which tends seriously to curtail the operation of those political processes ordinarily to be relied upon to protect minorities, and which may call for a correspondingly more searching judicial inquiry." *See generally* JOHN HART ELY, DEMOCRACY AND DISTRUST: A THEORY OF JUDICIAL REVIEW (1980) (emphasizing a procedural, representation-based justification for judicial review in constitutional cases).

[149] Considerations other than mistrust (narrowly construed) might sometimes favor restrictions on balancing, such as the possibility that bestowing special status on an interest might help to ensure that it is given due weight in light of temptations to sacrifice it to seemingly pressing but less important considerations. *See, e.g.*, Frederick Schauer, *Commensurability and Its Constitutional Consequences*, 45 HASTINGS L.J. 785 (1994).

[150] For elaboration, including discussion of when it may not favor balancing by the reviewing court, *See* Kaplow, *supra* note , at 1055–61.

[151] *See, e.g.*, Schaumburg v. Citizens for Better Env't, 444 U.S. 620, 636 (1980) (invalidating an ordinance regulating charitable solicitation on first amendment grounds due to overbreadth where the ordinance imposed a "direct and substantial limitation on protected activity" and did not serve a sufficiently strong government interest); Michael Dorf, *Incidental Burdens on Fundamental Rights*, 109 HARV. L. REV. 1175 (1996) (discussing the Supreme Court's tendency to apply closer scrutiny to incidental burdens that are substantial). Many infringements of the freedom of speech are subject to strict scrutiny, whereas content-neutral time, place, and manner restrictions receive intermediate scrutiny. *See, e.g.*, Heffron v. Int'l Soc'y for Krishna Consciousness, Inc., 452 U.S. 640 (1981); Grayned v. Rockford, 408 U.S. 104 (1972). Consider as well that anti-littering ordinances undoubtedly impose some burden on leafleting and newspapers, and recycling regulations (for example, those requiring a separate bin for newspapers) may also impose disproportionate burdens that do place more than a minuscule even if not a very large burden on the press.

[152] The connection is more crude than may be apparent if the *B** in step 2—for a given tier of scrutiny—is fixed, that is, independent of the magnitude of the infringement. Then, for example, a massive infringement of a right subject to intermediate scrutiny may survive because the government's interest barely exceeds a moderate *B** (that is far less than the *H* associated with the infringement), and a very slight infringement of a right subject to strict scrutiny may be struck down because the government's interest barely falls short of a huge *B** (that is far greater than the *H* associated with the

Second, consider some particular realms in which it seems most plausible that strict scrutiny would also have a quantitative dimension to its trigger. Government actions unduly burdening the free exercise of religion were once subject to strict scrutiny (and again are by statute, for federal government actions),[154] and when that is done it seems difficult to avoid giving some quantitative content with respect to what counts as an undue burden.[155] And a different outcome in *Washington v. Davis*[156] would have made step 1 in constitutional disparate impact discrimination cases much like that under Title VII. If taking that path, a screening function involving some sort of quantitative threshold at step 1 might seem appealing in light of the large number of potential cases (and some adjustments at step 2, depending on how it is understood, may be required as well).[157] By contrast, some qualitative triggers—such as content-based regulation of speech, burdens targeted at businesses in the news sector, or benefits and burdens targeted by race—might be seen as involving infringements that are rarely appropriate and thus subject to an easily triggered step 1 (no quantification required) accompanied by steps 2 and 3 that are tough on the government.

If step 1 is satisfied, step 2 then permits the government to justify the infringement by demonstrating that it advances a compelling interest. In this article's formulation, we inquire into whether $B > B^*$. At this point, it is familiar that the threshold, B^*, lies at the heart of understanding what step 2 really does. In this application, B^* (and, accordingly, B) may be given a quantitative interpretation, a qualitative (categorical) interpretation, or both. Regarding the latter possibility, we might consider only Bs of a certain type and, if they are present, further demand that they exceed some quantitative threshold. If there was only a qualitative test, then any B, however small, would justify any infringement, however large, as long as the (perhaps minuscule) B was of the right type.

Let us begin by considering the interpretation that B is purely quantitative. Specifically, assume that most[158] government interests count and the central question is whether, in magnitude, $B > B^*$. This version is now familiar. First, suppose that B^* is set independently of the H in a given case. Then, if $B^* > H$, there will exist cases in which the government's action is struck down even though $B > H$, specifically, when $B^* \geq B > H$.[159] If $B^* < H$, there will be cases

infringement). There could be a much larger H in the former case and a much larger B in the latter case, but the former infringement would be held valid and the latter invalid. That is, this step-function interpretation of levels of scrutiny, with a particular interpretation of how step 2 depends on the level of scrutiny, has some crude properties of balancing but nevertheless deviates substantially.

[153] *See, e.g.*, CHEMERINSKY, *supra* note , at 568.

[154] Strict scrutiny was ended by *Employment Division v. Smith*, 494 U.S. 872 (1990), which in turn was superseded by the Religious Freedom Restoration Act of 1993, 42 U.S.C. § 2000bb (2000), and in turn was held invalid as applied to the states in *Boerne v. Flores*, 521 U.S. 507 (1997). *See, e.g.*, CHEMERINSKY, *supra* note , at 1318–36.

[155] The points made in the notes on disparate impact, just below, are also apt regarding burdens on the free exercise of religion.

[156] 426 U.S. 229 (1976). *See, e.g.*, CHEMERINSKY, *supra* note , at 740–42.

[157] If, as will be discussed, there may be a qualitative limitation on what government interests count, and if moreover those excluded (such as cost savings) are the justifications for a huge portion of actions that have incidental disparate impact, the implication may be to invalidate much government action. Accordingly, what qualifies at step 1 may bear, under some views, on what interests should be allowed at step 2.

[158] In the simplest case, all government interests would count. But presumably there are at least some limitations. For example, a restriction on the press could not be justified by the benefit of helping incumbents win reelection by stifling dissent, and racial discrimination could not be justified by a purported benefit of harming those of the targeted race. *See* Cass R. Sunstein, *Naked Preferences and the Constitution*, 84 COLUM. L. REV. 1689 (1984). This type of limitation is present in the previous applications: in antitrust, higher profits from being better able to gouge consumers do not count in B, and under Title VII, certain employer benefits are not regarded to count (*see supra* note).

[159] One might be skeptical about the existence of such cases because of the importance of the rights that trigger strict scrutiny. However,, this is a matter of the proper weights, not whether balancing is appropriate. Indeed, the very

in which the government's action is permitted even though $H > B$, which arises when $H > B > B^*$. Unlike the stylized structured decision procedure in subsection I.A.1 and the structured rule of reason in antitrust—but like the protocol for disparate impact cases under Title VII—we do not avoid these latter errors on account of the more stringent balancing test that comes afterwards because there is no such step. (We might still be saved by a less restrictive alternative, considered momentarily—but not necessarily, because none may exist.) Under our stylized rule, step 2 was merely redundant in this case: when $B^* < H$, step 2 imposed a weaker version of the ultimate balancing test that must be passed in any event.

When B^* is set independently of H, we will ordinarily be in one of these two situations.[160] Subsection I.A.1 and subsequent analysis has shown how the only way to avoid these errors is to set the B^* in step 2 equal to the H determined in step 1. Then the step 2 test of whether $B > B^*$ becomes an inquiry into whether $B > H$, a pure balancing test. Moreover, as long as B^* is (at least de facto) a positive function of H, step 2 analysis means that decisions are made *as if* balancing is undertaken.[161]

Pursuing this matter further, under strict scrutiny we have the feature that, if step 1 is interpreted in an entirely categorical fashion—it is triggered by any infringement, no matter how small, of a qualifying right—then H was not quantified in step 1. Hence, under the presently contemplated interpretation of step 2, that prior conclusion should be understood merely as postponing by a step the need to quantify H. And, except in cases in which the government can offer no plausible B of any magnitude, this subsequent quantification will indeed be necessary.

Furthermore, if the B proffered at step 2 purports to justify the action, then the ordinary understanding of the notion of justification is that the B must be sufficient to render the action appropriate. If by a "compelling interest" it is meant that the interest must be compelling enough to warrant the infringement in the case at hand, then we have something akin to balancing at step 2. This conclusion is avoided if "compelling" is understood in a vacuum—to refer to some B^* such that $B > B^*$ is necessary and sufficient to permit the infringement without regard to H, that is, no matter how massive or minuscule the infringement might be.

Consider next the possibility that B^* is qualitative rather than quantitative.[162] (Implications for the combination case, with categorical limitations on what interests count, but a further quantitative requirement when the government does advance a qualifying interest, will be apparent.) Focusing on the purely qualitative interpretation, we can immediately see the

existence of a step 2 contemplates that sometimes infringements will be justified. Relatedly, it is not the case that all infringements of a given right, however important, are equal. A law restricting the ability to publically identify who is a spy (whether justified or not) is surely less weighty an infringement than one restricting the ability to refer by name to any government official. And, taking what many regard to be a sharp example, torture (*see, e.g.*, Fallon, *supra* note , at 1304 (discussing Charles Black's use of this example)), surely there are definitional issues concerning matters of degree (including at the low end, in defining at what point the infliction of discomfort or fear qualifies as torture) and concerning numbers (is it just an individual who is known to have planted a nuclear bomb in a city? or is every citizen to be drawn and quartered until a perpetrator is identified?). Or consider a less colorful but nevertheless important example of, on one hand, a pollution regulation that raises, in particular, the cost of newsprint to one that imposes burdens on the production and disposal of newsprint that are sufficiently large that they destroy the newspaper industry.

[160] The exception will be when, by happenstance, the H in a particular case just equals the B^* that is set without regard to the facts of the case. Note, however, that when H happens to be close to B^*, the resulting potential for error—in one direction or the other—will be small both because errors are unlikely and because the magnitude of the mistakes (the difference between H and B) will be small.

[161] *See* Kaplow, *supra* note , at 1049–55.

[162] For example, Fallon, *supra* note , at 1316 (emphasis added), states that strict scrutiny's step 2 involves "determining *which* governmental interests count as compelling." *See also id.* at 1321 ("the Supreme Court has frequently adopted an astonishingly casual approach to identifying compelling interests").

difficulties by applying the prior analysis.[163] First, when the government's interest is deemed not of a type that counts—say, savings in administrative costs—then any infringement that passes step 1, no matter how small, cannot be justified by a *B* of the wrong type, no matter how large. So, if it costs half of GDP to administer a non-infringing version of a regulation, and the degree of infringement is quite small, there would be a violation, essentially shutting down the government and society. Second, when the interest does count, then even massive infringements would be justified by the tiniest *B*, as long as it was of the right type. (Note that this latter problem is avoided if one combines a qualitative, categorical interpretation of step 2 with an appropriately fashioned quantitative requirement,[164] but the former problem is not.)

Another, related difficulty concerns whether a qualitative distinction is sustainable with respect to most government interests that may be in play. Suppose that advancing national security or public safety counts but saving money does not. As the preceding example involving half the GDP dramatically illustrates, large avoided costs obviously translate into substantial impacts on security and safety, among other things. A moment's reflection indicates that this relationship holds for smaller savings. If the national government has to spend a mere hundred million dollars a year more, one might expect, over the long run, that somewhat smaller funding would be available for security and safety. Although the probability and magnitude of budgetary impacts on these other government interests undoubtedly scale accordingly, this fundamental relationship holds.[165] Accordingly, it seems unavoidable to suggest that, at least at some point, costs count, but then we would have bridged both the qualitative and quantitative divides.

The permeability of qualitative categories regarding government interests can also be seen in other ways. Suppose that the government bans leafleting because some leaflets end up as litter, which must be removed. That interest would ordinarily be categorized as involving

[163] An additional, central question concerns the determination of which interests count and why. *See id*. at 1321–25; Stephen E. Gottlieb, *Compelling Governmental Interests: An Essential But Unanalyzed Term in Constitutional Adjudication*, 68 B.U. L. REV. 917 (1988); *id*. at 932–37 ("Unfortunately, while decisions of the Supreme Court and opinions of various members of the Court have frequently described or treated governmental interests as compelling, few have explained why. Several opinions have simply denied the existence, relevance, or weight of particular governmental interests without further attempts at justification. Many opinions referring with approval to a compelling governmental interest have provided no derivation whatsoever of that interest. Other opinions have referred only to other cases that themselves provide no derivation. . . . Thus, with few exceptions, the Court has failed to explain the basis for finding and deferring to compelling governmental interests." (footnotes omitted)). Richard Fallon further suggests that some Justices may be more or less inclined to count the interest advanced by the government based on the extent to which they agree with the decision to classify the infringement as one that should trigger strict scrutiny or, in any event, the importance of that type of infringement. *See* Fallon, *supra* note , at 1322–23. One can view this flexibility in the qualitative determination of compelling interests as a sort of implicit balancing.

[164] The universality of the quantitative dimension does not seem to be widely appreciated, although it is sometimes recognized in particular instances. *See, e.g.*, Fallon, *supra* note , at 1324 ("Finally, because diversity is inherently a matter of degree, the question emerges whether the government's interest should be defined as one in achieving diversity per se, or whether, instead, it should be regarded as one in attaining particular levels or increments of diversity? In other words, is there a compelling interest in moving from one level of diversity (that is more than zero) to another, higher level?" (footnote omitted)); Gottlieb, *supra* note , at 950 (in discussing *Moore v. Sims*, 442 U.S. 415 (1979), states: "Given these questions, it is not child abuse in general but the specific risk of abuse, if any, that might result from requiring a prompt hearing, that must be weighed against the potential for injuries caused by delay during an unjustified removal."). There is a difference between arguing that "[n]ot all compelling interests are equal," *id*. at 970, and appreciating that, for any particular interest that might be regarded as compelling, there exists the quantitative question of the extent to which it is advanced by the infringement. An interest of half the weight is still more compelling if it is advanced ten times as much.

[165] One could posit that all of the added costs will come from, say, the parks budget, none from defense, or that all will be funded by higher taxes. But unless courts will take over the operation of all government taxing and spending in perpetuity, the actual impacts will (on an expected basis) be more diffuse, and spread across all functions with some probability.

administrative inconvenience—that is, expense, not safety. But what of the fact that some leaflets inevitably blow into the street? Their removal (multiplying by the tens of thousands of government workers who dispose of litter) involves a small but strictly positive statistical risk of injuries. Likewise for traffic officers who need to be deployed when there are peaceful demonstrations.

These types of examples can also be run in reverse. Suppose that the government does establish a clear link between an infringement and national security or public safety, and suppose further that the magnitude of this interest is *B*. In most instances, there exists a way (and often many ways) to spend money to produce an offsetting impact on the same type of *B*.[166] Perhaps the interest is to keep military movements secret by reducing certain types of leaks. In that case, some additional expenditure on cybersecurity or other shielding might produce an equivalent enhancement, leaving the overall *B* of this type unaffected. So security and safety—of the magnitudes usually invoked—are really just about money after all.[167]

Qualitative, categorical distinctions of this sort seem difficult to sustain, conceptually and normatively, and a strong quantitative focus, which requires consideration of the magnitudes of both *B* and *H* (the latter normally seen as confined to step 1), seems difficult to avoid. These points raise significant questions about the extent to which strict scrutiny cases that reach step 2 may involve balancing after all, as well as about the implications if they do not. Relatedly, the analysis suggests the possibility that courts often engage in balancing, consciously or subconsciously, even while crafting opinions that do not suggest that balancing is taking place. Put another way, strict scrutiny doctrine may operate as if balancing is undertaken to a greater extent than is already recognized.[168] This possibility also raises the familiar question of whether greater explicitness, which focuses analysis and enhances transparency, would be preferable.

[166] This variation may belong better at step 3, as a sort of less restrictive alternative: if the government wants to promote security by *B*, then instead of infringing the right, it could instead have raised the budget. What matters is that the outcome would be the same, and the qualitative, categorical boundary would have been crossed.

[167] To take another example, suppose that Medicare—categorically or through a multi-factor risk assessment—made coverage for routine screening for melanoma a function of race, recognizing the starkly higher incidence on whites. Suppose further that this decision was made unanimously by a multi-racial panel of experts, applying the normal criteria used for other features of Medicare coverage. One might deem there to be a compelling public health interest in providing costly, routine screening for whites. But, of course, one could have screened everyone (perhaps an additional five million a year, at a cost of, say, $50 per screening, for an annual cost of $250,000,000). At that point, the benefit is no longer health but merely money. And, cycling back to where we began, one might counter-argue that, if all had to be covered, the budgetary impact would require trimming other coverage, which would translate the money back into a public health cost. (Note: superficial research by the author identified an undated source on an NIH website indicating that such screening is not covered by Medicare. Google searches and using the search function on Medicare's site (the "is my test covered" search box) yielded no prompt, dispositive answer.)

[168] *See, e.g.*, Fallon, *supra* note , at 1306–08; *id.* at 1307–08 ("In maintaining that strict scrutiny is sometimes applied as a balancing test, I do not mean to imply that it is always so applied or will bear no other interpretation. On the contrary, by juxtaposing the weighted balancing version of the test with the nearly categorical prohibition and illicit motive versions, I mean to suggest that a balancing interpretation is discordant with what the Court or its Justices have said and done in numerous cases. In addition, balancing applications frequently draw outraged protests from dissenting Justices who contend that the Court has betrayed the staunch commitment to preserve individual rights that the strict scrutiny test rightly embodies. My limited claim is that the Court sometimes applies a version of strict scrutiny that is little more than a balancing test." (footnote omitted)); *id.* at 1336 ("As I have tried to show, the catastrophe-avoidance and weighted balancing versions of the [strict scrutiny] test frequently require a seldom acknowledged proportionality-like judgment of whether marginal increments in the avoidance of risks or marginal reductions in the incidence of harms sufficiently justify infringements of fundamental rights in light of available, but typically less efficacious, alternatives."); *see also* Aleinikoff, *supra* note , at 963–72 (emphasizing the degree of balancing employed in constitutional law); Fallon, *supra* note , at 77–83 (describing balancing tests in constitutional law); Jud Mathews & Alec Stone Sweet, *All Things in Proportion? American Rights Review and the Problem of Balancing*, 60 EMORY L.J. 797, 799 n.4 (2011) (discussing the

If the government does offer a sufficiently compelling interest, we move to step 3 on narrow tailoring, which involves a species of less restrictive alternatives analysis.[169] A central question, the answer to which remains somewhat murky,[170] is whether a qualifying alternative—notably, a more narrowly tailored regulation—must be equally effective (or very close) and, if not, how one then determines the outcome at step 3. Paralleling the discussion of this step in Title VII disparate impact cases, it would seem that the requisite analysis would depend on the underlying conception of the overall test and, in particular, the nature of step 2.

If step 2 is taken to have no quantitative element, such that any *B* (no matter how small) of a qualifying type suffices to justify any infringement, then the corresponding analysis of less restrictive alternatives would seem to require equal effectiveness. As explained before, if an alternative is less effective, then, relative to that baseline, the original action boosts *B*. And if any positive *B* is enough to justify an infringement, then the original action would appear to be justified from this perspective. (Relatedly, like with Title VII disparate impact, if step 3 does require equal effectiveness, this carries the implication that step 2 must not have a quantitative dimension.[171])

It is familiar that courts and commentators, when speaking of equal effectiveness, often invoke the notion in rough terms, that is, contemplating that the less restrictive alternative may often be a bit less effective but nevertheless close enough to equally effective to warrant striking down the restriction. For example, narrow tailoring may involve some loss, perhaps due to the forgone prophylactic effect of a broader restriction[172] or because it may be harder in practice to prove that the narrower restriction has been violated (even when it truly was). To that extent, then, some tradeoffs are being made. And, as usual, once that is contemplated, it is natural to ask why we would not tolerate somewhat more sacrifice in *B*, as long as the concomitant reduction in *H* was even larger.[173]

majority and dissent's strong disagreement in *District of Columbia v. Heller*, 478 F.3d 370 (2008), about whether balancing is frequent or rare).

[169] The choice of the term "narrow tailoring" is naturally understood as byproduct of objections to overbreadth: if a restriction is overly broad, the solution is the particular less restrictive alternative of a more narrowly tailored version. But it is not clear that step 3 really means to be limited by this label. Consider a restriction that passes steps 1 and 2 and is not overly broad in the literal sense: perhaps it requires all individuals in domain *X* to do act *Y*, where *Y* is dichotomous and, moreover, the government interest from making an *X* do *Y* is the same for each individual (and not subject to economies of scale). But suppose that the government could instead require each *X* to do *Z* and that this alternative generates the same *B* but much less *H*. Taken literally, the original restriction passes step 3's narrow tailoring requirement and thus is valid, but if step 3 is understood more broadly as a less restrictive alternatives inquiry, then step 3 fails, so the original restriction is invalid.

[170] *See, e.g.*, Vicki C. Jackson, *Constitutional Law in an Age of Proportionality*, 124 YALE L.J. 3094, 3118 (2015) ("Not surprisingly, the U.S. case law on 'less restrictive means' sometimes obscures the distinction between 'less restrictive means' that are as effective and those that are not, in part because of the absence of any separate analysis of 'proportionality as such.'"); *id*. at 3118 n.11 (citing conflicting cases).

[171] And similar logic applies to the next case: if step 3 does involve balancing, this suggests that, if part of a coherent whole, then step 2 must involve balancing.

[172] *Cf.* Fallon, *supra* note , at 1272 ("The Court's employment of the terms 'necessity' and 'narrow tailoring' conceals a further ambiguity: If a challenged statute is necessary to promote a compelling governmental interest in the sense that nothing else would do as well, should the statute still be invalidated if it is not narrowly tailored in the sense that it employs admittedly overbroad, prophylactic restrictions? An example . . . would come from a prophylactic measure designed to protect national security in a context in which no more narrowly tailored restrictions on individual rights would so effectively reduce the risk of a calamitous terrorist strike. Astonishingly, after roughly forty years of experience with the strict scrutiny formula, the Court seems never to have resolved the question of when, if ever, overinclusive prophylactic statutes could be upheld on the ground that they are necessary to promote compelling interests.").

[173] It is hard to avoid the interpretation that, at least sometimes, judges will describe an alternative as equally effective when they know full well that it is not, in order to avoid speaking in explicitly quantitative terms about *H* and *B*, as well as *H′*and *B′*.

If step 2 does have a quantitative dimension, then the second balance—or, equivalently, the delta/delta test—is appropriate for assessing less restrictive alternatives. (Accordingly, the inputs are *H* and *B*, and also *H*′ and *B*′.) The intuition is that, for an alternative to be less restrictive, it must reduce *H* (from *H* to *H*′), and whether we should require this depends on whether the reduction in *B* (from *B* to *B*′) is smaller, making the sacrifice in the government's interest (ΔB) worth the reduction in the degree of infringement (ΔH). A closely analogous statement can be made using the language of narrow tailoring, as elaborated in the margin.[174] Put more sharply, this view holds that, on one hand, we would prefer a less restrictive alternative that eliminated most or all of the original *H*, even at some (perhaps slight) sacrifice in *B*, but we would prefer the original restriction if the best alternative reduces *H* modestly but eliminates most or all of *B* (where, moreover, we know from step 2 that $B > H$ for the enacted version). It is recognized that, accordingly, some balancing may occur under strict scrutiny at this step,[175] but it is less often appreciated that this phenomenon is derivative of step 2 having balanced the *H* and *B* from the original restriction (even if that balance was submerged).

Viewed in its particulars and as a whole, strict scrutiny doctrine does depart importantly from the stylized structured decision procedure introduced in subsection I.A.1 and, in varying degrees, from the other applications considered earlier in this article. Nevertheless, this article's framework sharpens our understanding of strict scrutiny's three-part test, often in same way that the analysis illuminated other structured protocols.

In addition, each of the issues identified in subsection I.A.2's discussion of queasiness about balancing seems particularly apt with respect to strict scrutiny. Because of the nature of the interests involved, regarding both the character of the infringements and of the government's interests, it is hardly surprising that courts are reluctant to engage in either quantification[176] or

[174] Taking the sometimes metaphorical invocation of narrow tailoring quite literally, we can imagine a restriction that, when maximal, involves the largest *H* and *B*. As the restriction is gradually narrowed—in an optimal fashion so that the initial narrowing eliminates from the restriction's reach the part of its domain with the greatest ratio of *H* incurred to *B* generated (equivalently, the greatest difference between ΔH and ΔB)—at first the combined effect is advantageous, but as the coverage shrinks ever further, the marginal tradeoff becomes worse and worse, until at some point it is optimal to stop. (Of course, if the restriction has little value even at its core, the optimal stopping point would involve no restriction whatsoever.) In this sense, narrow tailoring—or less restrictive alternatives analysis more broadly—can be understood as explicitly posing a marginal, delta/delta type of inquiry with respect to the design—and court review—of restrictions. *Cf.* Fallon, *supra* note , at 1330–31 ("In determining whether a particular degree of statutory under- or overinclusiveness is tolerable, a court must judge whether the damage or wrong attending an infringement on protected rights is constitutionally acceptable in light of the government's compelling aims, the probability that the challenged policy will achieve them, and available alternative means of pursuing the same goals. . . . In assessing whether this consideration should be controlling, it may therefore be important to take note of whether a less restrictive alternative exists that would achieve almost as much risk reduction while infringing less on protected rights. Once again, it thus seems impossible to think sensibly about compelling governmental interests and the narrow tailoring requirement as if they were sequentially isolated components of a bifurcated two-step inquiry—or as if every compelling interest were equally compelling or every infringement of a triggering right equally disturbing."); Jackson, *supra* note , at 3117–18 (discussing a case in which the Israeli High Court of Justice, in undertaking proportionality analysis, seemed to require that a less restrictive alternative be equally effective but then reintroduced more of a delta/delta framework in the final, proportionality step).

[175] *See, e.g.*, Fallon, *supra* note , at 1325 ("To put the question this way might seem to collapse the narrow tailoring prong of strict scrutiny into the compelling interest element of the test. As reformulated, the question essentially becomes whether there is a compelling governmental interest in achieving as much reduction in the risk or incidence of harm as a challenged regulation is likely to achieve.").

[176] Consider the sort of factfinding that would be required if a court were to take seriously the challenge of quantifying risks to public safety or national security, rather than retreating to dichotomous pronouncements about whether the pertinent threats in a particular case are "compelling," which judgments in turn seem to be predicated on something akin to de facto judicial notice.

direct comparisons[177]—at least explicitly, for we have seen that, particularly at step 2 and step 3, both are hard to avoid and may not actually be eschewed, at least at some level. The decisions are made *as if* balancing is undertaken.[178] As well, categorical rules sometimes play a role when review is predicated on mistrust of other government actors, including in a constitutional context, and step 1 seems to reflect such considerations.

Furthermore, subsection I.B.1's lessons on optimal information gathering, which have not been the focus in this section, also have some relevance. A familiar point is that, when there is an obvious less restrictive alternative, it may be helpful to start, and end, the analysis there rather than exert excessive effort in making hard calls at step 2. Although strict scrutiny is often stated as a sequential inquiry, there seems to be some tolerance for flexibility when appropriate.[179] Other features of strict scrutiny can be viewed through this lens, notably, when some infringements are deemed, at step 1, not to be infringements for reasons that have more to do with the government's compelling interests that are supposedly deferred to step 2. For example, it is familiar that the speech involved in undertaking bribery or other forms of criminal conspiracy, such as price-fixing, may be regarded as unprotected even though the legal prohibitions are unquestionably content-based.[180]

[177] *See, e.g.*, Bendix Autolite Corp. v. Midwesco Enters., 486 U.S. 888, 897 (1988) (Scalia, J., concurring in the judgment) ("This process is ordinarily called 'balancing,' . . . but the scale analogy is not really appropriate, since the interests on both sides are incommensurate. It is more like judging whether a particular line is longer than a particular rock is heavy."); Fallon, *supra* note , at 1270 (stating, in reference to strict scrutiny: "To count as a solution to the problem, a doctrinal structure needed, among other things, to impose discipline, *or at least the appearance of discipline*, on judicial decisionmaking and thus to escape the taint both of *Lochner*esque second-guessing of legislative judgments and of flaccid judicial 'balancing.'" (emphasis added)). Relatedly, some commentators strongly question the extent to which such constitutional decisions should be made through balancing. *See, e.g.*, BERNHARD SCHLINK, ABWÄGUNG IM VERFASSUNGSRECHT (1976); Aleinikoff, *supra* note .

[178] *See, e.g.*, Fallon, *supra* note , at 80 ("This strong criticism is quite mistaken if 'balancing' is conceived, as it should be, as a metaphor for (rather than a literal description of) decision processes that call for consideration of the relative significance of a diverse array of potentially relevant factors. Understood in this way, the term 'balancing' does not signify that decisionmaking necessarily proceeds by reducing all relevant considerations to a single metric, assigning them quantitative values, and then weighing them against one another with the precision of a scale. If this misleading picture is rejected and 'balancing' is viewed as a metaphor for multifactor decisionmaking, the 'incommensurability' objection becomes either too strong or too weak. It is too strong to be credited at all—because too inconsistent with the deepest assumptions of practical reasoning—if it suggests that, when different kinds of considerations bear on a decision, there can be 'no basis in our knowledge of value' to say that one decision is rationally preferable to another." (footnotes omitted)); Jackson, *supra* note , at 3156-57 ("Even absent a common metric, however, judgments about the relative priority of two values can be rational. An example is 'large-small trade-offs' involving a small sacrifice of one value for a large gain in another. It is a mistake to understand balancing in mathematical terms: rather, 'proportionality as such' balancing should entail a reasoning process about the priority of one constitutional value as it relates to another in a particular setting." (footnote omitted)); Kaplow, *supra* note , at 1049–55; *see also* Aleinikoff, *supra* note , at 972 ("We rarely hear objections that legislatures are unable to value and compare competing social interests. Furthermore, we expect courts to make exactly these kinds of judgments in crafting common law doctrine.").

[179] *Cf.* Fallon, *supra* note , at 1333 ("In contrast with this bifurcated sequence, I have suggested that the effort to identify compelling interests and to determine the adequacy of regulatory tailoring is likely to involve fluid, two-way traffic in which assessments of ends and means occur simultaneously—at least in cases in which challenged governmental regulations, viewed realistically, will at best merely reduce risks or incidences of harm more or less effectively than would other regulations.").

[180] *See* Frederick Schauer, *The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience*, 117 HARV. L. REV. 1765 (2004). Of course, these domain restrictions might be justified not only by looking ahead to step 2's *B* but also, at step 1, on the ground that they do not involve the type of *H* that counts, or, indeed, any *H*. It is also natural to consider the reverse: perhaps a significant reason that some infringements have been deemed to trigger strict scrutiny (including many forms of facial discrimination and content-based regulation of speech) is that a look ahead to step 2's consideration of government justifications suggests that they will rarely be sufficient and often nonexistent (confining attention to legitimate interests, that is). Such thinking, for example, underlies the per se illegality of price-

Potential interactions among the steps are even greater to the extent that intent is made part of the inquiry. Notably, the nature of the government's purported justification and the plausibility of proffered less restrictive alternatives may illuminate intent, which for facially neutral laws that discriminate may be part of the requisite analysis at step 1. Likewise, if instead disparate impact were sufficient, then as we saw in Part III there would be additional ways that attempts to sequentially silo step 1 and step 2 would be counterproductive.

B. *Proportionality Analysis*

Proportionality analysis is employed in varying ways in a number of jurisdictions and is proposed by some as a replacement for strict scrutiny—and the other tiers of review—in the United States.[181] These approaches are united in that they have a core (and a name) that focuses explicitly on balancing in some sense.[182] The brief treatment here will focus on instantiations of proportionality analysis that, at least on their surface, involve a structured inquiry.[183] These typically conclude with a final step involving proportionality (balancing[184]) as such, but precede it with other steps that, if they have independent bite, would generate some outcomes that differ from those under unconstrained balancing. Under another interpretation, the stated steps may be more of a checklist that reminds judges and communicates to primary actors certain features involved in balancing that might otherwise be given insufficient attention. Under that

fixing under antitrust law and helps to explain why even small infringements trigger strict scrutiny (which is often fatal), notably, when they are blatant.

[181] *See, e.g.*, AHARON BARAK, PROPORTIONALITY: CONSTITUTIONAL RIGHTS AND THEIR LIMITATIONS (Doron Kalir trans., 2012); NIELS PETERSEN, PROPORTIONALITY AND JUDICIAL ACTIVISM: FUNDAMENTAL RIGHTS ADJUDICATION IN CANADA, GERMANY AND SOUTH AFRICA, ch. 1 (2017); Jackson, *supra* note ; Mathews & Stone Sweet, *supra* note ; *see also* ROBERT ALEXY, A THEORY OF CONSTITUTIONAL RIGHTS 44–110, 394–414 (Julian Rivers trans., 2002) (discussing respects in which German constitutional rights norms are amenable to balancing that takes the form of proportionality analysis).

[182] Note that the term "proportionality," if applied literally to an infringement of degree H and a government interest of magnitude B, means that the restriction will be approved if and only if B exceeds some constant multiplied by H, which is the constant of proportionality. If this constant is not 1, we could restate the units of H (or of B) to convert the balancing-like formulation into a pure $H > B$ test. For example, if the factor is 5, so that the literal proportionality requirement is that a restriction is upheld if and only if $B > 5H$, we can transform our measure of H so that each original unit equals 5 converted units. (One could perform nonlinear transformations, so that if the concept of proportionality means more loosely that the minimally requisite B is increasing in H, but perhaps more or less steeply at different levels of H, one could, through a monotonic but nonlinear transformation, convert the units of H to conform to a simple balancing test.) To this reader, discussions of proportionality analysis do not clearly indicate whether "proportional" literally means proportional or is used in a looser sense that merely indicates this positive relationship that is associated with some manner of making tradeoffs. (My best guess is the latter.)

[183] Not all do. *See, e.g.*, Jackson, *supra* note , at 3098–99.

[184] *See, e.g.*, R. v. Oakes, [1986] 1 S.C.R. 103, 139 (Can.) ("Although the nature of the proportionality test will vary on the circumstances, in each case courts will be required to balance the interests of society with those of individuals and groups."); *id*. at 140 ("Even if an objective is of sufficient importance, and the first two elements of the proportionality test are satisfied, it is still possible that, because of the severity of the deleterious effects of a measure on individuals or groups, the measure will not be justified by the purposes it is intended to serve. The more severe the deleterious effects of a measure, the more important the objective must be if the measure is to be reasonable and demonstrably justified in a free and democratic society."). Although this final step is not always described as literally involving the act of balancing the competing interests—*see, e.g.*, Jackson, *supra* note , at 3099-100 ("While this step is sometimes referred to as involving 'balancing,' the 'proportionality as such' question in structured proportionality doctrine differs from 'balancing' tests that tend to focus primarily on quantification of net social good . . . ."); *cf*. Mathews & Stone Sweet, *supra* note , at 803 ("Most judges . . . would not characterize balancing in such blunt, utilitarian terms.")—it is difficult to see what difference is intended, that is, other than the presence of the preceding, structured decision procedure, which is the focus of the present analysis. (Keep in mind that placing a high weight on one side of the balance does not change the qualitative character of balancing. *See supra* note .)

interpretation—or for proportionality analysis that has no structure and simply asks the bottom-line question of whether the government's interest, *B*, is large enough to justify the magnitude of the infringement, *H*—we have pure balancing, an approach that has already been considered at length and was implicitly compared to strict scrutiny in the preceding section.

For concreteness, consider the following five-step structured protocol, which seems close to depictions of constitutional review in Canada[185] as well as to some other formulations.[186] For purposes of brevity, this statement will combine a description of the rule on its own terms with an unapologetic translation into the parlance of this article. Step 1 asks if the infringement, causing *H*, is sufficient to trigger review, by reference to some H^* (which, as will be discussed, might be interpreted to be categorical in whole or in part). If not, the challenger loses. If it is, step 2 asks the government to show a legitimate purpose. As usual, this may be viewed as a test of whether $B > B^*$ (which also raises the qualitative/quantitative question, and in particular we will be interested in the possibility that $B^* = 0$). If not, the government loses. If it succeeds, step 3 asks if the connection between the infringement, *H*, and the government interest, *B*, is rational. Again, if not, the government loses. If it is, step 4 asks if the interest is advanced in a manner that involves minimal impairment (or various other phrasings that suggest an inquiry into less restrictive alternatives).[187] If not (that is, if there does exist a less restrictive alternative), the government loses. If the impairment is minimal, we then proceed to step 5, which involves proportionality (balancing) as such.

As a preliminary comparison with strict scrutiny—which suggests much of the analysis to follow—observe the following: The first steps of each may be regarded to be similar. (The main difference may be that proportionality analysis sweeps more broadly, covering, for example, infringements that would be subject to intermediate scrutiny in the United States, but if the subsequent demands for justification would be correspondingly lower in such cases, substantial correspondence would remain.) The second step (perhaps combined with the third) might be seen as similar to step 2 under strict scrutiny. The fourth step of proportionality analysis, on minimal impairment, would be matched to step 3's less restrictive alternatives inquiry under strict scrutiny. Finally, proportionality's fifth step, balancing, is absent under strict scrutiny—

[185] *See Oakes*, 1. S.C.R. at 134–40; Jackson, *supra* note , at 3099–101, 3111–14.

[186] *See, e.g.*, PETERSEN, *supra* note , at 2 (describing the German Constitutional Court's first step as determining whether a right has been restricted and the second step as whether there is adequate justification, with the latter step having four components: whether the purpose is legitimate, the measure is rationally connected to the purpose, there exists no less restrictive alternative that is equally effective, and finally proportionality in the strict sense); Mathews & Stone Sweet, *supra* note , at 802–03 (describing: a preliminary (unnumbered) stage in which "the judge considers whether a prima facie case has been made to the effect that a government act burdens the exercise of a right"; "The first stage . . . mandates inquiry into the 'suitability' of the measure under review. The government must demonstrate that the relationship between the means chosen and the ends pursued is rational and appropriate, given a stated policy purpose."; "The second step—'necessity'—embodies what Americans know as a 'narrow tailoring' requirement. At the core of necessity analysis is a least-restrictive-means (LRM) test, through which the judge ensures that the measure at issue does not curtail the right more than is necessary for the government to achieve its goals."; "The third step—balancing *stricto sensu*—is also known as 'proportionality in the narrow sense.' In the balancing phase, the judge weighs, in light of the facts, the benefits of the act (already found to have been narrowly tailored) against the costs incurred by infringement of the right, in order to decide which side shall prevail." (footnotes omitted)).

[187] For example, Germany's formulation refers to the less restrictive alternatives inquiry using the language of whether the restriction is "necessary," *see, e.g.*, Elisabeth Zoller, *Congruence and Proportionality for Congressional Enforcement Powers: Cosmetic Change or Velvet Revolution?*, 78 IND. L.J. 567, 582 (2003), which is reminiscent of the "reasonably necessary" language sometimes used under antitrust's rule of reason, but has the potential drawback that literal interpretations of "necessary" can be extreme and thus potentially misleading, a point discussed with regard to the business necessity formulation of an employer's justification in Title VII disparate impact cases.

although, as we have seen, under some interpretations it may be understood to arise under that rule's step 2 and step 3.

Now let us consider explicitly the five steps of proportionality analysis, which will allow us to see the extent to which this rough characterization may be apt and what other interpretations are possible. Proportionality analysis's step 1 seems to be an on/off categorical test, in which case the earlier discussion under strict scrutiny would be applicable. However, it appears that in many jurisdictions, the list is longer and, in particular, at least some alleged infringements are assessed quantitatively as well. This possibility was also addressed above, for example, when noting the previous application of strict scrutiny to burdens on the free exercise of religion caused by facially neutral regulations and also the possibility that disparate impact claims may have been allowed in cases of discrimination allegedly caused by facially neutral restrictions.

Step 2, which involves some sort of $B > B^*$ inquiry, raises a number of now-familiar issues. Under one view, this is a purely qualitative test—whether the government's purpose is legitimate or otherwise of a nature that counts at all and, if it is, we only require $B > 0$.[188] Problems under this interpretation were discussed at length with respect to strict scrutiny, but a key difference here is that there will be balancing later, in step 5. As a consequence, step 2 may have little bite but practices with $B < H$ will be struck down in step 5. However, as explained previously, if the interests that count, qualitatively, are substantially circumscribed, then many laws may be struck down even though $B > H$, where the B here interprets the notion of acceptable interests more broadly. Recall also the troubles associated with advancing significant categorical limitations on government interests.

Suppose instead that step 2 is quantitative. This possibility is suggested by statements familiar in Canadian law that the infringement must be "demonstrably justified" and that it must serve a "pressing and substantial" government interest.[189] It is difficult to see how one can assess whether the government's proffered purpose is pressing or substantial without engaging in any quantification. Likewise, the requirement that the infringement be justified seems to indicate that it be, well, justified. As explained, justification ordinary denotes a reason sufficient to warrant the act in question.[190] Regardless of what may be the best interpretation in Canada or elsewhere, let us consider this case explicitly.

When step 2 is indeed quantitative, we have a similar (but not identical) diagnosis as with strict scrutiny. If $B^* > H$, then some restrictions will be condemned even though $B > H$. On the other hand, if $B^* < H$, we do not have (as with strict scrutiny or Title VII disparate impact) that some restrictions will be permitted even though $H > B$ because, at step 5, that very question will be asked and, if the answer is affirmative, the restriction will be condemned at that point. As explained previously, this case in which step 2 is nonbinding may be regarded as irrelevant. But it would be pointless—if proceeding sequentially—to struggle over whether or not $B > B^*$ if that is a close call whereas it is obvious that $B < H$ in any event. That is, we will sometimes end up avoiding an easy balance at step 5 by performing a more difficult comparison at step 2.

[188] This characterization is more consistent with that offered in PETERSEN, *supra* note , at 74 (suggesting that the inquiry is confined to whether the purpose advanced is "legitimate"), than that described earlier in the text here and in the next paragraph, which refers to the formulation that seems to be used in Canada (although this is one of the jurisdictions that Petersen considers).

[189] *See* R. v. Oakes, [1986] 1 S.C.R. 103, 135, 138–39 (Can.).

[190] Consider one of the interpretations of Canada's step 2 offered in Jackson, *supra* note , at 3100 (emphasis added): the step is said to ask "whether the government's purpose is *sufficiently important* to serve as a basis for limiting the right *at all*." If one focuses on "sufficiently important," a quantitative interpretation is suggested, but if one focuses on "at all," one might regard B^* to equal zero.

Finally, it is familiar that, once B^* is viewed as having a quantitative dimension, it makes the most sense to set $B^* = H$ because doing so uniquely avoids the preceding two problems. But when that is done, step 2 amounts to asking whether $B > H$,[191] which is precisely the balancing that purports to be located in step 5.[192] Note further that the government will only fail with any frequency at step 2 if B^* is nontrivial (or if, under the other interpretation, there is a significant categorical limitation on what types of *B*s count), which means that, if step 2 is being done sensibly, we have indeed balanced, just without admitting it.[193] (This point is consequential for the additional reason that sometimes proportionality analysis is advanced on the ground that it makes decisionmaking more transparent.[194]) As mentioned, this point would be moot if $B^* = 0$ (or nearly so) and, moreover, there are no significant categorical limitations on *B*. But then this point is largely immaterial precisely because step 2 does not matter very often.

Step 3 will be passed over quickly here because of the suggestion that, in at least some proportionality jurisdictions, this step rarely binds.[195] Note further that, to the extent it is quite undemanding but does sometimes bind, the need for balancing in step 5 will have been avoided[196] only when the balance would have been easy.[197]

Step 4, the inquiry into minimal impairment, raises many of the issues regarding inquiries into less restrictive alternatives that were developed in subsection I.A.2—and, in this instance,

[191] This assessment, just as under strict scrutiny, means that *H* must be quantified at step 2—well before step 5—even if it did not have to be quantified at step 1.

[192] More broadly, as discussed in connection with strict scrutiny, once B^* is, at least implicitly, made a positive function of *H*, we again have an instance in which decisions are made as if under balancing.

[193] It is also possible, as noted, to set $B^* < H$: the lower is B^*, the less often step 2 strikes down a government action, but the more often we do a subsequent comparison, ultimately (if step 5 is reached), but with B^* in essence elevated to *H*.

[194] *See, e.g.*, Jackson, *supra* note , at 3142–44; Mathews & Stone Sweet, *supra* note , at 804 (Proportionality analysis (PA) "is a highly formalized argumentation framework, the basic function of which is to organize a systematic assessment of justifications for government measures that would burden the exercise of a right. A government must explain such acts, which PA subjects to the highest standard of judicial scrutiny. In doing so, PA enhances the transparency of rights review, not least by making explicit the justifications for limiting rights the court has either accepted or rejected and at precisely what stage of the analysis."); *but see id.* at 807 (stating that, through the use of proportionality analysis, "judges can bring a *semblance* of determinacy to balancing by subjecting it to a fixed procedure" (emphasis added), suggesting perhaps that it is through deception rather than transparency that this mode of judicial decisionmaking would be legitimated); *id.* ("PA bestows a *sheen* of politico-ideological neutrality on a court" (emphasis added)); *id.* at 810 ("An opponent of PA may well conclude that, at best, PA is little more than fancy, doctrinal window dressing for what is, in fact, generic law making by any other name."); *see also* PETERSEN, *supra* note , at 68 (arguing that when a court engages in explicit balancing, in the final step of the proportionality inquiry, transparency is at its greatest, a point that suggests that courts wishing to disguise their balancing may undertake it implicitly at earlier steps in the analysis); *id.* at 189 (same); *id.* at 150–53 (arguing that two prominent German constitutional decisions that invoked categorical prohibitions merely disguised important balancing and failed to examine competing considerations carefully).

[195] *See also* PETERSEN, *supra* note , at 168 ("According to the German interpretation, a measure is already rationally connected to a purpose if it marginally contributes to the promotion of the latter. A severe restriction of an individual right would thus pass the rational connection stage even if it has only a minimal positive impact. However, the low effectiveness has to be taken into account at the balancing stage." (footnote omitted)); Jackson, *supra* note , at 3117 ("Canadian cases rarely turn on this third step . . . ."). Alternatively, sometimes an action may be regarded as lacking a rational connection to a purpose because of the clear availability of a less restrictive means, in which case one might usefully combine consideration of these two steps. *See* PETERSEN, *supra* note , at 74. However, my impression from reading the literature on proportionality analysis that examines particular cases is that this step binds more often than is acknowledged (as suggested, for example, by the reference in the next footnote).

[196] *See, e.g.*, PETERSEN, *supra* note , at 105 (describing two recent cases in which the Canadian Supreme Court invalidated laws for lack of a rational connection where there is some suggestion that balancing was implicitly undertaken).

[197] It could only be difficult if the infringement was likewise small. But, repeating a refrain begun in subsection I.A.1, when difficult balances are avoided by short-circuiting the process, the inevitable result is that sometimes the final outcomes will be erroneous.

the analysis may more resemble the application to antitrust's rule of reason than that for strict scrutiny, just above. Suppose, for present purposes, that balancing has not taken place at step 2, taking the five-step rubric on its face and setting to the side the analysis just offered. (If step 2 does involve full balancing, then the analysis offered for strict scrutiny, under the quantitative understanding of its step 2, would be applicable.)

Because proportionality analysis is, at the core, a balancing regime, it seems natural to focus on the quantitative version of less restrictive alternatives analysis,[198] which involves the second balance or, equivalently, the delta/delta test.[199] Under the presently maintained assumptions, we have the strange predicament considered previously: to answer step 4's question, we need to know H, B, H', and B' and to undertake analysis that involves the full richness of the basic balancing test (whether $H > B$), but we have not yet reached the balancing stage. Of course, one could and naturally would incorporate the core balancing inquiry, but then stating step 5 as a final step would be a misnomer,[200] and any suggestion that step 4 may avoid the need for balancing would be substantially misleading.[201] In addition, sometimes the less restrictive alternatives analysis will be difficult—because it is a close call or because collecting information on H' or B' is particularly challenging—yet the underlying balance (supposedly deferred to step 5) may lead to clear condemnation, notably, when B is obviously less than H to begin with. (Recall that we are here supposing that this core balancing was not performed previously, at step 2.[202])

Step 5, proportionality as such, has already been discussed at the outset of this segment, explaining that it seems akin to balancing, and explicitly so. A central observation about step 5 involves elaboration of some of the previous framing comments on proportionality analysis. On one hand, it is often suggested that the intermediate steps—steps 2 and 4 in particular—have bite, deciding many cases and, moreover, avoiding the need to undertake balancing very often. As we have seen, the former (that many cases are decided, against the government, at steps 2 or

[198] That said, this step is often described as requiring that the alternative means be "equally effective," PETERSEN, *supra* note , at 2 (describing the protocol in the German Constitutional Court), although in practice (as we have seen elsewhere) this may be relaxed. *See id*. at 11 ("Often, these alternative measures are not quite as effective as the adopted measure. However, the court does not deem the difference in effectiveness sufficiently important to justify the more severe restriction of the individual right.").

[199] If not—and if, in particular, any $B > 0$ was sufficient at step 2—then step 4 could be seen as requiring equal effectiveness. Under that view, when step 5 is reached, either less restrictive alternatives analysis would be repeated, but at that point using the balancing version, or less restrictive alternatives would be off the table (step 4 having been passed by the government), so an infringement would be valid as long as $B > H$, even though there exists a less restrictive alternative that would be superior under the balancing version of that inquiry.

[200] This points raises questions about observations like Vicki Jackson's that "Canadian cases rarely turn on this [final proportionality] step, generally finding laws unconstitutional on minimal impairment grounds." Jackson, *supra* note , at 3117. If the minimal impairment inquiry, as suggested in the text, itself involves a form of balancing and requires the court to determine all that is necessary to undertake the final step's balancing, this suggestion seems curious. The alternative is that Canadian courts require equally effective alternatives (which would be inconsistent with the foundation of proportionality analysis) and, moreover, find them routinely to be so (which may involve wishful thinking or dissembling). Some recognize that the analysis of less restrictive alternatives under proportionality analysis may well involve some implicit balancing, *see, e.g.*, PETERSEN, *supra* note , at 130–34, but such comments do not suggest the degree of potential entanglement of the two steps indicated in the text here.

[201] As explained previously, sometimes difficult balances will indeed be avoided, notably, when the second balance (or delta/delta test) is easy but the underlying balance is a close call. But there are also the opposite cases, noted in the text. If the steps were regarded as combined (collapsed), then one could do whichever was easiest in a given case, which one suspects tends to occur in any event.

[202] Under this set of interpretations, neither H nor B may have had to be quantified previously. (If step 2 sets $B^* > 0$, but B^* does not depend on H, then B will have had to be quantified there.) But, to perform the second balance (or the delta/delta test), we do need to know both H and B (and more), so to forgo asking whether B is obviously less than H at this step seems senseless.

4) can arise in two ways. First, one or both of those steps may implicitly involve the full balancing test, asking whether $H > B$. Then, even though step 5 is not reached, balancing has occurred.

Second, step 2 may often be decisive (against the government) but without having balanced H against B. In that event, some balancing is avoided. But this involves three situations: H exceeds B substantially, in which case the avoided balance would have been easy; B exceeds H substantially, in which case we have avoided an easy balance and reached the wrong outcome (this can arise when B^* is significantly above H); and H and B are closer together, so a difficult balance is avoided, but this is done by assigning liability even though this may well involve an incorrect outcome. If this is how proportionality analysis operates, and it often avoids reaching step 5, the results are quite problematic, indeed, from the core perspective embodied in the proportionality framework itself.

On the other hand, it may be that one or both of steps 2 and 4 are rarely binding. If we set $B^* = 0$ at step 2, or close to that, then step 2 only takes out cases in which step 5's balance would have been easy and come out the same way.[203] The presence of step 2 may nevertheless be regarded as valuable in a different manner (other than in how it affects outcomes) by reminding courts and signaling to government actors the importance of justifying serious infringements. Although conceivable, this view is difficult to reconcile with setting $B^* = 0$, because then the message sent by this step is that virtually anything that advances a legitimate purpose justifies serious infringements (with any stiffer message relegated to the balancing step, which would exist even without the preceding structured inquiries).[204]

In reflecting on my own modest exposure to the literature on proportionality analysis, I am sometimes left with the impression that some proponents implicitly wish to have their cake and eat it too.[205] They wish to claim that various of the steps are important—they bind, and may help avoid difficult balances—and also that the framework is really, at its core, all about proportionality.[206] The analysis is disciplined, suggesting that it is often decisive, but it is not

[203] As discussed, it may be immaterial whether the underlying $H > B$ test is conducted in step 4 or in step 5, or is undertaken by one who notices an obvious violation at step 4 and then skips ahead to step 5. Nevertheless, as suggested, this entanglement of the steps despite insistence on their separation may generate confusion and undermine transparency.

[204] Courts adopting this view may, on one hand, set $B^* = 0$—or perhaps really engage in balancing at step 2—but nevertheless announce through their opinions that infringements must be "demonstrably justified," by advancing "pressing and substantial" interests, which government actors (who read the words but do not understand the actual operation of proportionality analysis) take to impose a heftier step 2 requirement. Alternatively, there may be less disconnect if infringements ordinarily survive balancing only when B is large, and the only mismatch is that such language is attached to step 2 when it really operates through balancing, nominally at step 5 but perhaps often implicitly being undertaken at step 2.

[205] Regarding not only those who write about proportionality analysis but also scholars of strict scrutiny, Title VII disparate impact, and antitrust, I also frequently get the sense that it is believed that less restrictive alternatives analysis rescues us from the need to engage in difficult balancing. *See, e.g.*, *supra* note (discussing this phenomenon in connection with antitrust's rule of reason). As explained, however, proper analysis of less restrictive alternatives requires knowing not only H and B but also H' and B', and itself involves a balancing test—which sometimes, to be sure, may be easier than determining whether $H > B$, but it may as often be harder.

[206] For example, Jud Mathews and Alec Stone Sweet argue: "Moreover, [proportionality analysis] comprises a multi-stage balancing framework; that is, judicial balancing is not restricted to the final balancing-in-the-strict-sense stage, but takes place within each of the tests. And the tests are sequenced in order of increasing stringency, so that courts insert themselves into the legislative process no more than is necessary to defend rights." Mathews & Stone Sweet, *supra* note , at 805; *see also* Jackson, *supra* note , at 3100–01 ("In this way, if the means chosen are not suitable or necessary to advance the government's interest, the case can be resolved at one of these stages: the courts need not reach the 'proportionality as such' question unless there is a genuine conflict between the government's interest and the interests protected by the right. . . . In this way, courts are not 'substituting' their judgment for that of the legislature."). Many aspects of such statements are mysterious. Their structured framework might be caricatured as asking: Is the restriction a

discussed that it may accordingly deviate from the outcomes that would be reached under a direct inquiry into proportionality as such. Regardless of what is actually true of various proponents' intentions and understandings, the suggestion here is that the attempt to define each step's requirements more explicitly—by matching them against the stylized structured decision procedure outlined in subsection I.A.1 and analyzed from a number of angles in Part I—greatly illuminates different possibilities and sharpens our appreciation of what turns on different interpretations.

CONCLUSION

This article compares unconstrained balancing to structured decision procedures. Viewed abstractly and generally,[207] structured decision procedures suffer from two sets of infirmities. As final decision rules, they sometimes fail to assign liability even though the harm (*H*) of the defendant's action exceeds the benefit (*B*), and they sometimes assign liability even though *B* exceeds *H*. As guides to information gathering, they significantly violate every central principle of optimal information collection and rest on key predicates—that information on harm and benefit are conceptually and practically distinct—that are false in many applications.

The core of the article uses the general framework to examine three areas of law: antitrust (rule of reason and mergers), Title VII disparate impact, and strict scrutiny (and proportionality analysis) in constitutional law. In each instance, this methodology casts new light on each step of these doctrine's structured protocols, even though they differ from each other and from the stylized template.

Regarding step 1's requirement that $H > H^*$, in every application there was significant uncertainty about the magnitude of the H^* threshold. Since that threshold is one of the two features that defines structured decision rules and distinguishes them from balancing, this ambiguity is telling. If H^* is indeed significant, the general problem that step 1 results in no liability in cases in which *H* is nevertheless greater than *B* is indeed present. And if H^* is negligible, then step 1 is largely irrelevant (other than perhaps as a screening device). The only setting with a plausible justification for the existence of a weighty step 1 was in constitutional law, regarding the qualitative dimension that circumscribes a reviewing court's jurisdiction to domains in which review seems appropriate because of the degree to which the government actors being challenged may be untrustworthy.

Step 2's requirement that $B > B^*$ likewise exhibits tremendous ambiguity regarding the magnitude of B^*, the other central defining feature of these structured rules, and also creates the possibility of errors if B^* is set above *H*. Interestingly, for those structured decision rules that do not include a final balancing inquiry, the fuzziness surrounding the requisite B^* makes it possible for implicit balancing to take place at this step, and this is precisely so if B^* is determined

slam-dunk loser? If not, is it a clear loser nevertheless? If not, is it a loser on-balance? First, it is obvious that, whenever one of the former questions is answered affirmatively, the last question would be easy. Second, as explained, one of the former questions could be quite difficult even though the last one is easy. Third, it is unclear how it is less intrusive into the legislative process to invalidate based on one of the earlier questions—both in light of the foregoing points and because doing so (that is, in an earlier step) is surely more insulting to the legislature. Some authors also claim that the rigor of this process offers a stark contrast to open-ended, unprincipled balancing. *See, e.g.*, Mathews & Stone Sweet, *supra* note , at 804. But it is hard to see how this is so. (Their multi-step framework does call for analysis of less restrictive alternatives, but so do most proposed alternatives, including unconstrained balancing, properly understood, as explained in subsection I.A.2.)

[207] This more conceptual perspective is developed extensively in Kaplow, *supra* note .

contextually, so as to equal *H* in the case at hand. And under doctrines that do have an explicit balancing step, it may not reached or be largely moot because balancing has already occurred at this point.

Less restrictive alternatives requirements have proven to be confusing and problematic in many of the doctrines considered here, and in very similar ways. This article's core framework indicates how such analysis should, in principle, be conducted and clarifies many issues regarding these applications. Proper consideration of less restrictive alternatives on its face subsumes the very balancing that is deferred or omitted under these doctrines. In addition, puzzlement in some areas of the law regarding whether a less restrictive alternative must be equally effective is addressed by relating this question to the underlying legal test, notably, the degree of *B* that is otherwise required to justify a practice. Considering the analysis of less restrictive alternatives as an integral part of the overall decision rule (which, under structured rules has been highly obscure, often in unappreciated ways), rather than in a vacuum, is the best way forward.

If and when a decisionmaker reaches the final, balancing step of those structured decision rules that have one, the analysis is the same as under unconstrained balancing. Of course, one may well not reach this stage due to a prior, erroneous decision along the way or because balancing has implicitly occurred at an earlier step. In addition, the often-advanced notion that difficult balances are avoided by structured decision rules is revealed to be misleading. Many avoided balances would have been easy, indeed, often easier than the decisions that must be made at earlier steps but are unnecessary under unconstrained balancing. Moreover, whenever difficult balances are avoided, this is because of early, dispositive decisions that are as likely to be erroneous as correct. Decisionmakers who are particularly concerned about correct outcomes may consciously or subconsciously breach structured protocols, including by reverse engineering, to reach outcomes more in accord with those under unconstrained balancing.

To the extent that information gathering is taken to be guided by structured decision procedures, the general infirmities of such methods are indeed evident. This dimension was elaborated particularly with regard to merger analysis in antitrust law and the assessment of disparate impact under Title VII. To varying degrees, all of the defects of the structured approach were manifest, and some recognized challenges in the doctrine came into better focus by applying this article's framework. More broadly, U.S. civil litigation conforms neither to optimal nor structured procedures, although judges attempting to improve case management could make headway by drawing on the principles of optimal information collection that are elucidated here.

Stepping back, we can see that systematic application of this article's stylized structured decision procedure to each of the areas of law that employs a structured protocol pays off. When the match is close, the lessons carry over directly. When there are differences, their potential significance becomes apparent. And when it is hard to tell because of doctrinal ambiguity, the stylized template highlights the uncertainties and identifies the implications of different possible interpretations.

In each area of law, it is remarkable how much has been overlooked or underappreciated due to the failure of precision regarding the central features of existing structured decision procedures. It is not that courts and commentators disagree with the important criticisms advanced in this article but rather that they do not seem to be aware of the questions. At the most fundamental level, each existing or proposed protocol expressly deviates from unconstrained balancing—an entirely familiar notion—yet it has not been thought necessary to articulate just what the differences are and why, in light of them, one should favor any element of these structured substitutes.

In the applications examined here, the difficulties of quantifying the pertinent harms and benefits and, often, comparing them to each other, are substantial. The resulting queasiness about quantification and comparison may well motivate existing doctrine and help to explain why it is that some areas of law employ these structured decision procedures whereas others (such as the negligence test in tort law) do not. Nevertheless, the challenges are hidden rather than avoided, and much mischief with regard to procedure and outcomes results. Although some may view opacity as a feature rather than a bug, transparency generally promotes the quality of decisionmaking, the accountability of decisionmakers, and the sound development of the law.