1-2004

Patent Law, the Federal Circuit, and the Supreme Court, A Quiet Revolution

Glynn S. Lunney Jr
Texas A&M University School of Law, glunney@law.tamu.edu

Patent Law, the Federal Circuit, and the Supreme Court: A Quiet Revolution

Glynn S. Lunney, Jr. *

Over the last twenty years, a quiet revolution has taken place in patent law. Traditionally, patents were rarely valid, but if valid, broadly enforced. Since Congress created the Federal Circuit in 1982 and vested it with exclusive intermediate appellate jurisdiction over patent appeals, patents have become routinely valid, but narrowly enforced. This article evaluates the economic consequences of this revolution. Focusing on the reasons for, and the costs of, uniformity in patent protection, this article shows that the revolution will tend to limit the patent system's ability to ensure the expected profitability, and hence the existence, of desirable, but high cost innovation.

I. INTRODUCTION

Over the last twenty years, a quiet revolution has taken place in patent law. Before 1982, courts strictly enforced the nonobviousness requirement in an attempt to limit patents to "those inventions which would not be disclosed or devised but for the inducement of a patent."[1] At the same time, courts broadly interpreted a patent's scope under the doctrine of equivalents, so that a patent would reach any product or process that performed "substantially the same function in substantially the same way to obtain the same result."[2] Taken to-

* Professor of Law, Tulane University School of Law, New Orleans, Louisiana. I would like to thank Mark Lemley, Leslie Lunney, Kimberly Moore, Tom Nachbar, and my dissertation committee, David Malueg, Ila Alam, and Emilson de Silva for helpful comments and suggestions. As always, any remaining mistakes are my responsibility.

[1] *Graham v John Deere*, 383 US 1, 11 (1966).

[2] *Machine Co v Murphy*, 97 US 120, 125 (1877); see also *Sanitary Refrigerator Co v Winters*, 280 US 30, 42 (1929) (quoting function-way-result standard with approval);

 0-226-64593-2/2004/0011-0001$10.00

gether, these two doctrinal interpretations ensured that patents were rarely valid, but if valid, broadly enforced. However, since 1982, the nature of patent protection has gradually changed. In that year, Congress created the Federal Circuit and vested it with exclusive intermediate appellate jurisdiction over appeals from patent litigation.[3] Intended, at least by some of its supporters, to rescue patents from a judiciary often suspicious, if not overtly hostile, towards patents, the Federal Circuit has taken its role as defender of the patent system seriously. Using its exclusive jurisdiction over patent appeals, and relying on the sporadic and inherently limited nature of Supreme Court review, the Federal Circuit has rewritten the nonobviousness requirement and the doctrine of equivalents, sharply limiting their reach. Where traditionally only those innovations that represented substantial technical advances would satisfy the nonobviousness requirement, under the Federal Circuit's interpretation, even a minor technical advance will suffice. Where traditionally, the doctrine of equivalents would reach most competing variations of a patented invention, under the Federal Circuit's interpretation, even directly competing substitutes will often prove noninfringing. As a result, under the Federal Circuit we have moved from patents that were rarely valid, but if valid broadly enforced, towards patents that are routinely valid, but narrowly enforced.

Although the Court undoubtedly has some power to reign in the Federal Circuit's rewriting of patent law, the Court has so far shown surprisingly little willingness to do so. With respect to the nonobviousness requirement, even where the Federal Circuit has expressly and directly rejected contrary Court authority, the Court has (so far) entirely refused to revisit the issue. With respect to the doctrine of equivalents, the Court has recently begun to play a more active role, granting certiorari in two cases, *Warner-Jenkinson Co. v. Hilton Davis Chemical Co.*[4] and in *Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co.*,[5] in which the Federal Circuit had narrowed sharply the doctrine of equivalents. Rather than rebuke the Federal Circuit, however, the Court largely embraced the Federal Circuit's doctrinal revisions.[6] Although the Court expressly refused to eliminate the

---

Graver Tank & Mfg Co v Linde Air Prods Co, 339 US 605, 608 (1950) (quoting function-way-result standard with approval).

[3] See Federal Courts Improvement Act of 1982, Public L No 97-164, 96 Stat 25 (codified in scattered portions of 28 USC).

[4] 520 US 17 (1997).

[5] 535 US 722 (2002).

[6] Even in rejecting aspects of the Federal Circuit's approach, the Court in *Warner-Jenkinson Co* emphasized the Federal Circuit's special expertise in patent law. In discussing whether the traditional "function-way-result" or the alternative "insubstantial variation" represented a superior verbal formulation of the doctrine of equivalents, the Court stated: "[W]e see no purpose in going further and micro-managing the Fed-

doctrine of equivalents altogether, as some members of the Federal Circuit desired, the Court neither actively defended the doctrine nor restored its pre-Federal Circuit scope. Instead, the Court largely affirmed a sharply narrower role for the doctrine of equivalents.

Given its refusal to revisit the Federal Circuit's changes to the nonobviousness doctrine and its decision largely to affirm the Federal Circuit's changes to the doctrine of equivalents, the Court has likely reinforced the trend towards routinely valid, but narrowly enforced patents already evident in Federal Circuit decisions over the last ten years. The question thus becomes whether this transformation of patents is desirable. To evaluate this transformation, this article begins in Part II with a statistical summary of appellate patent litigation over the last sixty years. Despite the Federal Circuit's propatent holder reputation, this summary reveals that claims of patent infringement are no more likely to succeed since the Federal Circuit's advent. However, where historically, claims of patent infringement failed because the court ruled the patent invalid or otherwise unenforceable, claims of patent infringement fail today because the defendant's product or process falls outside the scope of the plaintiff's patent. Part III of this article then identifies the doctrinal changes that appear to have driven the apparent trends in patent enforceability. Together, Parts II and III establishes the Federal Circuit's shift towards routinely valid, but narrowly enforced patents.

After setting the stage, Parts IV, V, and VI examine the economic consequences of the Federal Circuit's transformation of patents. Under the traditional economic analysis of patents, the costs of patent protection arise directly from the tension between a regime of private rights and the public good character of an innovation's information component. The point of private property is to enable its owner to exclude others, but excluding others from a good characterized by nonrivalrous consumption is not Pareto optimal. Patents are costly, under the traditional analysis, precisely to the extent that this tension between a private right of exclusion and the nonrivalrous character of an innovation's information component reduces the innovation's social utility.

If we used the traditional economic approach to evaluate the switch to routinely valid, but narrowly enforced patents, the traditional approach would identify two, potentially offsetting, consequences. On the one hand, effectively eliminating the nonobviousness requirement would impose a social loss by extending patent protection to innovations that would have been devised and disclosed

---

eral Circuit's particular word-choice for analyzing equivalence. We expect that the Federal Circuit will refine the formulation of the test for equivalence in the orderly course of case-by-case determinations, and we leave such refinement to that court's sound judgment in this area of its special expertise." *Warner-Jenkinson Co*, 520 US at 40.

without the inducement of a patent. On the other hand, narrowing the scope of protection should yield social benefits by reducing the private rights-public good tension for patented innovations. Whether, under the traditional economic approach, the social losses would outweigh the social benefits is simply unclear.

Rather than follow the traditional economic approach, this article reexamines the economics of patents in an attempt to understand more clearly the role the nonobviousness doctrine and the doctrine of equivalents play within the economic structure of patent law. This reexamination identifies information, agency, and other transaction costs as the keys to understanding patent law. With perfect information, and in the absence of agency and transaction costs, the patent system would likely[7] prove both unnecessary and undesirable. In such a perfect world, a Lindahl-style contribution scheme,[8] for example could ensure optimal innovation. Moreover, if we could make such a scheme work, it would ensure optimal innovation without requiring a regime of exclusive rights inconsistent with the nonrivalrous consumption of the innovation's information component. However, in the real world, information, agency, and transaction costs undoubtedly exist. In their presence, a regime of exclusive rights may prove more efficient and effective for encouraging some types of in-

[7] Despite the usual assumption that private ownership of public goods reduces the social value of the associated public good, over the years, a few commentators have argued that property rights in certain kinds of information increase the information's social value, despite the information's public good character. See Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 JL & Econ 265, 276-77 (1977); see also Mark F. Grady & Jay I. Alexander, *Patent Law and Rent Dissipation*, 78 Va L Rev 305 (1992); F. Scott Kieff, *Property Rights and Property Rules for Commercializing Inventions*, 85 Minn L Rev 697 (2000). Essentially, these commentators argue either that information goods are subject to the same type of overuse problems that can plague commonly owned private goods or that private ownership of information can otherwise enhance desirable coordination. Although I tend to reject these arguments for reasons that others have given, see, e.g., Robert P. Merges & Richard R. Nelson, *On the Complex Economics of Patent Scope*, 90 Colum L Rev 839, 872-79 (1990) (arguing that while patent races seem wasteful, the competition they generate improves the speed with which innovation occurs and thereby leads to more innovations than alternative of single firm exploitation of an innovation); Robert P. Merges, *Commentary: Rent Control in the Patent District: Observations on the Grady-Alexander Thesis*, 78 Va L Rev 359 (1992); a full discussion of these arguments is beyond the scope of this article. For purposes of this article, I will accept the traditional view that granting exclusive rights in the information component of an innovation, while perhaps necessary to ensure the innovation's existence *ex ante*, will reduce the innovation's social value *ex post*.

[8] See Knut Wicksell, *A New Principle of Just Taxation* (1896), reprinted in Richard A. Musgrave & Alan T. Peacock eds, *Classics in the Theory of Public Finance* 72 (St Martin's Pr, 1958) (JM. Buchanan trans); Erik Lindahl, *Just Taxation—A Positive Solution* (1919) (Elizabeth Henderson trans), reprinted in Richard A. Musgrave & Alan T. Peacock eds, *Classics in the Theory of Public Finance* (St Martin's Pr, 1958). For further discussion of what has been called the Wicksell-Lindahl tax, see Jules L. Coleman, *Markets, Morals and the Law* 278-81 (Cambridge, 1988).

novative activity. Even where a regime of exclusive rights represents the best available alternative for encouraging certain types of innovation, the social value of an innovation will presumably be somewhat less if protected by a patent than if its public good aspect could have been freely and fully exploited. Yet, if providing patent protection ensures the creation of a desirable information product and does so more efficiently than the plausible alternatives, such as patent prizes or direct government subsidies,[9] the fact that the information product could have been more valuable still in the absence of the patent's protection has little practical significance.[10]

Nevertheless, the continuing tension between private rights and public goods suggests that, even where patents are the best available policy mechanism, we should provide patent protection only if, and to the precise extent, necessary to secure each individual innovation's *ex ante* expected profitability.[11] Yet, the same information, agency, and transaction costs that require the use of a patent regime in the first place also limit our ability to tailor patent protection to each individual innovation. In the face of imperfect information and potential agency and transaction costs, the historical practice of both patent and copyright law has been to provide more-or-less uniform protection for a creative product or process that satisfies a given set of more-or-less uniform prerequisites. With uniformity, expanding patent protection may increase the incentive for, and thereby ensure the existence of, additional innovative products. However, if we expand protection uniformly to all of the creative works that satisfy a given set of prerequisites, expanded protection will also apply to those innovative products that would have been produced with no or less protection ("preexisting" products). Given uniformity, determining the optimal scope of patent (or copyright) protection becomes a

[9] For a discussion of some of the advantages and disadvantages of various approaches to encouraging innovation, see Brian D. Wright, *The Economics of Invention Incentives: Patents, Prizes, and Research Contracts*, 73 Am Econ Rev 691 (1983); see also Michael Kremer, *Patent Buyouts: A Mechanism for Encouraging Innovation*, 113 Quarterly J Econ 1137 (1998); Suzanne Scotchmer & Jerry Green, *Novelty and Disclosure in Patent Law*, 21 Rand J Econ 131 (1990). Generally speaking, the principal advantage of a regime of exclusive rights is that such a regime, in addition to tying the rewards for innovation market directly to the innovation's marketplace success, tends to decentralize the decision-making process, assigning decision-making responsibility to those likely to possess the relevant, but otherwise private, information. Awarding an innovator an exclusive right to her innovation allows the would-be innovator to decide whether the expected rents available exceed her reservation cost for the innovation. It also allows an innovator to threaten to exclude a consumer from access to the innovation in order to force the consumer to reveal her true reservation price for the innovation.

[10] See R. H. Coase, *The Lighthouse in Economics*, 17 JL & Econ 357 (1974).

[11] I recognize that an innovator's reservation cost must include the cost not only of those research efforts that succeed, but also the associated research efforts that will fail.

balancing of the value gained from the additional creative output that broader protection may ensure against the value lost from the reduced ability to exploit the information component of the preexisting products. As we extend protection to a broader range of preexisting innovations, any given expansion in protection becomes more costly, dragging down the optimal level of protection.

This uniformity insight, largely missing from existing analyses of patent and copyright, suggests both that: (1) patent protection must refuse to provide protection sufficient to ensure an expectation of profit for the full range of innovative products eligible for patents, even in cases where the innovative products represent the most valuable use of society's resources; and (2) although variation in protection entails its own costs, there may be instances where tailoring protection will prove desirable in order to limit the costs of uniformity. In addition and most importantly, the costs of uniformity dictate that any system of uniform rights over information must be narrowly tailored to a particular instance where the incentives available from a market, operating against a background of private rights in tangible things alone, leave a significant gap between an innovative product's expected desirability, relative to alternative uses of the resources, and its expected profitability. Because the costs of uniformity increase with the divergence between the optimal level of uniform protection and the optimal level of individualized protection for the range of innovative products eligible for protection, we should strive to limit application of a uniform system of intellectual property rights to "similar" innovative products. To minimize the costs of uniformity, innovative products are "similar" precisely to the extent that: (1) a given set of uniform prerequisites defines when a significant gap will likely arise between the desirability of an innovative product (relative to alternate uses of the resources) and its expected profitability; and (2) a given set of uniform exclusive rights approximates the protection precisely necessary to close that gap for the range of innovative products eligible for protection.

Having identified the costs of uniformity as a key to understanding the economic structure of patent law, the article then moves in Parts V and VI to examine whether the shift towards routinely valid, but narrowly enforced patents, is likely to promote "the Progress of . . . the useful Arts." Viewed in terms of the costs of uniformity, the shift entails two central consequences. First, rewriting the nonobviousness doctrine to allow patents even for those innovations that would have been devised and disclosed in the absence of a patent both directly reduces the social value of the overprotected innovations and indirectly reduces the optimal level of patent protection. By expanding the category of preexisting innovative products to which any given uniform increase in protection will apply, effectively eliminat-

ing the nonobviousness requirement increases the costs of any given expansion in patent protection. The switch to routinely valid patents therefore drags down the optimal level of uniform patent protection and reduces the patent system's ability to encourage desirable, but costly innovation. Second, narrowing the nonobviousness requirement and the doctrine of equivalents eliminates two sources of variability in patent protection. Under the traditional interpretations of these doctrines, the legal system could use information from post-innovation developments to tailor the patent protection an innovation received to the level precisely necessary to ensure that innovation's *ex ante* expected profitability. Because such tailoring reduces the costs of uniformity, tailoring enables the patent system to expand protection for, and thereby ensure the expected profitability of, more costly innovations, without unduly overprotecting less costly innovations. By reducing the legal system's ability to tailor protection to the individually optimal level, the switch to routinely valid, patents narrowly enforced leads to a "one size fits all" protection scheme that again limits the patent system's ability to ensure more costly innovations.

Although this article does not attempt to resolve the ultimate desirability of the switch as an empirical issue, focusing on the costs of uniformity identifies, more precisely than the traditional analysis, the economic trade-off at stake in the switch. Specifically, the switch to a "one-size fits all" patent system will tend to promote the goals of the patent system if and only if the information and agency costs entailed in: (i) separating and precluding patents for those innovations that would have occurred in the absence of a patent; and (ii) individually tailoring protection to the level precisely necessary to enable an innovator to capture her reservation cost; exceed the social value of the additional innovations such tailoring efforts can ensure.

We begin with an empirical analysis of appellate patent decisions over the last sixty years and the changing nature of patent protection.

II. THE CHANGING NATURE OF PATENT PROTECTION

Since Congress created the Federal Circuit and vested it with exclusive intermediate appellate jurisdiction over patent appeals in 1982, the nature of patent protection has changed. Although often perceived as a pro-patent holder court,[12] an empirical examination of appellate

[12] See, *e.g.*, Robert L. Harmon, *Patents and the Federal Circuit* 684-740 (Bureau of National Affairs Inc, 3d ed 1994); Ian Ayres & Paul Klemperer, *Limiting Patent holders' Market Power Without Reducing Innovation Incentives: The Perverse Benefits of Uncertainty and Non-Injunctive Remedies*, 97 Mich L Rev 985, 1024 & nn.99-100 (1999) (noting that the Federal Circuit is more "pro-patent" than its predecessor courts); Rochelle Cooper Dreyfuss, *The Federal Circuit: A Case Study in Specialized*

patent litigation over the last sixty years suggests that a patent holder is no more likely to succeed on a patent infringement claim under the Federal Circuit than under the circuit courts it replaced. Yet, when we break down the reasons why patent holders fail to succeed on their infringement claims, we find two significant differences under the Federal Circuit. First, there has been a sharp reduction in the percentage of patent infringement claims that fail because the patent is found invalid or otherwise unenforceable. And, second, there has been a sharp increase in the percentage of patent infringement claims that fail because the allegedly infringing device or process is found to fall outside the patent claims. The statistics therefore suggest a shift from rarely valid, but broadly enforced towards routinely valid, but narrowly enforced, patents since 1982. We begin our examination of this shift with a statistical summary of the appellate resolutions of patent infringement litigation over the last sixty years.

Courts, 64 NYU L Rev 1, 17-20, 25-26 (1989) (describing Federal Circuit's sensitivity to patent policy and resulting pro-patent owner stance in substantive issues as well as improved availability of remedies and preliminary injunctive relief); Lawrence G. Kastriner, *The Revival of Confidence in the Patent System*, 73 J Pat & Trademark Off Soc'y 5, 13 (1991); Allan N. Littman, *Restoring the Balance of Our Patent System*, 37 IDEA 545 (1997) (noting that the Federal Circuit is overwhelmingly pro-patent) Allan N. Littman, *The Jury's Role in Determining Key Issues in Patent Cases:* Markman, Hilton Davis *and Beyond*, 37 IDEA 207, 209 (1997) ("Patent lawyers have perceived both juries and the Federal Circuit to be pro-patent."); Robert P. Merges, *Commercial Success and Patent Standards: Economic Perspectives on Innovation*, 76 Calif L Rev 805, 822 (1988) (noting Federal Circuit's pro-patent reputation); Alexander E. Silverman, *Intellectual Property Law and the Venture Capitalist Process*, 5 High Tech L J 157, 161 (1989) (stating that creation of Federal Circuit has increased power of patents); *Symposium: Early Patent Publication: A Boon or Bane? A Discussion on the Legal and Economic Effects of Publishing Patent Applications After Eighteen Months of Filing*, 16 Cardozo Arts & Enter LJ 601, 623 (noting that "the Federal Circuit is very pro-patent") (statement of Douglas Wyatt, Senior Partner, Wyatt, Gerber, Meller, & O'Rourke); David Silverstein, *Patents, Science and Innovation: Historical Linkages and Implications for Global Technological Competitive-ness*, 17 Rutgers Computer & Tech L J 261, 310-11 (1991) (noting that "statistics as well as perceptions of the patent bar" showed the Federal Circuit to be "pro-patent"); Thomas G. Field, Jr., *Zurko Raises Issue of Patentability Standards*, Natl L J at C2 (Feb 8, 1999) ("Because the invalidity rate is now lower than it was between the 1930s and the 1960s, some suspect federal circuit judges, even those not formerly on the CCPA, of being unduly 'pro-patent.' Former corporate patent counsel seem to be particularly suspect."). The perception of a pro-patent bias has become so strong at times that a number of Federal Circuit judges have felt the need to step forward and affirmatively deny the supposed bias. See Howard T. Markey, *The Federal Circuit and Congressional Intent*, 41 Am U L Rev 577, 579 (1992) ("The uninformed, unsupported, and unsupportable assertion that the Federal Circuit might somehow become biased in favor of patents has apparently by now foundered on the facts."); Randall R. Rader, *Specialized Courts: The Legislative Response*, 40 Am U L Rev 1003, 1013 (1991) ("In addition, actual practice reveals that the Federal Circuit has not become the specialized court of limited jurisdiction its detractors feared. Rather, the court hears cases in virtually every area permitted by its jurisdictional statute and shows no overt favoritism in patent disputes.").

To explore the changing nature of patent protection under the Federal Circuit, I conducted an empirical investigation of all appellate decisions arising from patent infringement[13] litigation in six pre-Federal Circuit time periods beginning with the period 1944-1946, and since January 1, 1984.[14] As has become the practice, I conducted a population, rather than a sample, study and included all intermediate appellate utility[15] patent infringement decisions that were available in the "US Court of Appeals Cases—Federal Circuit" LEXIS database for the post-Federal Circuit periods or in the "Federal Cases—Combined Courts" LEXIS database for the pre-Federal Circuit periods. The defined population included 1,492 decisions, and included both published and Rule 36 summary affirmances.

After identifying cases in the relevant population, I initially identified each case as one of three results: (1) "success;" (2) "failure;" or (3) "non-final" resolutions where a patent holder neither succeeded nor failed. "Success" was defined as a decision where a patent holder obtained preliminary or permanent injunctive relief, or damages, on any patent claim at issue in the litigation.[16] "Failure" was defined as a decision where the appellate court finally resolved all claims of patent infringement *and* no claim of patent infringement in the case succeeded. The final category consisted of non-final decisions, where a patent holder did not succeed in obtaining the relief sought, but the claims of infringement were not finally rejected by the court. Rather, the appellate court reversed the ruling of the district court on one aspect or another, and remanded the case for further proceedings.[17] For the second category of cases, I also determined whether the claims at issue in the litigation failed because: (1) the claims were found invalid

[13] Appeals from Patent and Trademark Office decisions as to whether to issue a patent were excluded from the study.

[14] The author used a search of "core-terms (patent and infring!)" and an appropriate date restriction to identify initially the relevant cases. This search tended to exclude appeals from decisions of the Patent and Trademark Office refusing to issue a patent, but to obtain all patent infringement cases, whether the central issue was infringement or validity. To supplement this initial search, an additional search of "core-terms (patent and obviousness)" with an appropriate date restriction was conducted.

[15] Rulings involving design or plant patents were excluded from the study.

[16] Under this definition, a patent holder was considered to have succeeded even if the patent holder did not obtain relief on all patent claims at issue and even if some patent claims at issue were held invalid. Although other definitions of "success" might be used, compare Glynn S. Lunney, Jr., *E-Obviousness*, 7 Mich Telecommun & Tech L Rev 363 (2001) (using a claim-by-claim analysis of patent litigation to evaluate the changing role of obviousness), I felt that this approach appropriately recognized the fact that patents are drafted with multiple claims specifically to account for the possibility that some of the claims might be found invalid or not infringed in a subsequent case.

[17] If such an initially non-final resolution subsequently came before an appellate court a second time, then it would again be classified into one of the three identified categories depending on the result in the second appeal.

or otherwise unenforceable[18]; or (2) the allegedly infringing product or process did not fall within the scope of the patent claims at issue, either literally or under the doctrine of equivalents. A summary of the data, by period, is included in Appendix I.

Figure 1 presents the percentage of cases in which a patent holder succeeded on a claim for patent infringement in each of the six pre-Federal Circuit time periods, and for successive two-year periods from 1984 through 2001. To facilitate comparison between the pre- and post-Federal Circuit time periods, Figure 1 also presents the average[19] success rates for patent holders in the pre- and post-Federal Circuit periods.

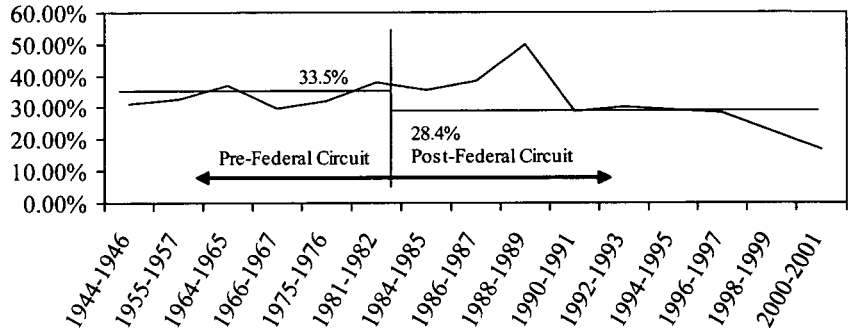

Figure 1. Percentage of All Population Cases[20] in Which a Patent Holder Succeeded on a Patent Infringement Claim

[18] The "average" rates for the pre- and post-Federal Circuit periods are calculated as the arithmetic average of the six pre-Federal Circuit time periods and the nine two-year post-Federal Circuit periods, respectively.

[19] Under the author's definition of success, cases where the appellate court remanded the case for further proceedings were not considered a success for the patent holder. These non-final resolutions were however included in the count of total cases, and therefore reduced the calculated success rate. Nevertheless, even under alternative definitions, the basic conclusion—that success rates have not increased significantly under the Federal Circuit—remains unchanged, as do the general trends in the data. For example, if we excluded nonfinal decisions from our sample, then the average pre-Federal Circuit success rate would rise to 36.8 percent and the average post-Federal Circuit success rate would rise to 35.6 percent. Even with the nonfinal decisions excluded from the sample, the overall success rate remains statistically unchanged. In addition, even with nonfinal decisions excluded, the trends in the data remain the same: The success rate under the Federal Circuit begins generally higher in the Federal Circuit's early years from 1984-1989, reaches its peak in 1988-1989, and then declines steadily from 1990-2001.

[20] This category includes those cases where a patent was held invalid for a failure to satisfy the statutory requirements of sections 101, 102, 103, 112 of the Patent Act, or because the court held the patent invalid or unenforceable under a judge-made doctrine, such as inequitable conduct or patent misuse.

As Figure 1 reflects, patent holder overall success rates have fallen from an average of 33.5% in the periods before, to 28.4% after, the creation of the Federal Circuit. Moreover, most of this fall has occurred since the 1988-1989 period. If we focus on the years 1984-1989, at the outset of the Federal Circuit's tenure, patent holders succeeded in 41.4% of the appellate cases, with patent holder success rates peaking in 1988-1989 at 50%—well above the patent holder's average success rate in the pre-Federal Circuit era. If, on the other hand, we focus on the results since 1989, patent holder success rates under the Federal Circuit fell to an average of 24.4% for 1990 to 2001, trended generally lower throughout this period, and reached their lowest level (16.7%) in the most recent period under study (2000-2001).

Although not statistically significant,[21] the drop in success rates following the creation of the Federal Circuit contradicts, or at least, does not support, the usual portrait of the Federal Circuit as a pro-patent holder forum.[22] Yet, because of the self-selection involved in appellate litigation, the fact that the success rate does not change significantly with the Federal Circuit's advent is not altogether surprising. Commentators usually suggest that parties involved in litigation will settle those cases where both sides can accurately predict the likely outcome.[23] Particularly where one party or the other is very

[21] If we treat each of the six pre-Federal Circuit and each of the post-Federal Circuit time periods as discrete data points, with a normal distribution around their mean, then we can calculate a confidence interval around the average success rate for the pre- and post-Federal Circuit eras. At a 95 percent, two-tailed, level of confidence, the confidence interval was ±3.7 percent for the pre-Federal Circuit periods and ±7.4 percent for the post-Federal Circuit periods. Because the confidence intervals for the pre- and post-Federal Circuit periods overlap, the difference between the average success rates in the two periods is not statistically significant at a 5 percent confidence level. For a similar approach to hypothesis testing, see John R. Allison & Mark A. Lemley, *Empirical Evidence on the Validity of Litigated Patents*, 26 AIPLA QJ 185, 194 n.20 (1998) (noting that a population sample can be treated as a subset of randomly distributed observations from a larger superpopulation of possible cases for appellate resolution).

[22] For example, in testimony before the Federal Trade Commission during its recent round of hearings on Competition and Intellectual Property Law and Policy in the Knowledge-Based Economy, Professor Mike Scherer stated:

> [S]tatistically it used to be, before the Federal Circuit came into existence, about two-thirds of patents that were litigated were found either invalid or not infringed or both. Two-thirds of the cases, the patent holder lost. That has nearly reversed since the Federal Circuit.

Testimony of F. M. Scherer, Hearing of the Federal Trade Commission on Trends in Federal Circuit Jurisprudence, July 10, 2002, at 33-34 (transcript available at http://www.ftc.gov/opp/intellect/020710trans.pdf) (last visited December 14, 2002). For an expression of similar perceptions of the Federal Circuit as a pro-patent forum, see the sources cited in note 12.

[23] See, e.g., George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J Legal Stud 1, 5 (1983) (mathematical model demonstrates that individual

likely to win on appeal, the usual expectation is that they will settle accordingly. The opportunity to settle usually leaves for appellate decision those cases where there are relatively evenly balanced arguments on each side, leading to the expectation of similarly balanced results. While an unexpected change in patent law may temporarily shift the balance in favor of one party or the other, parties should quickly adjust their expectations, and settlement offers or demands, accordingly. As a result, even when the law changes unexpectedly, we should nevertheless expect a fairly rapid return to the norm.

However, rather than a 50-50 norm, the patent appellate success rates appear to center around a 30-70 norm. Although there are a number of possible explanations for this,[24] the availability of injunctive relief probably best explains the 30-70 norm. Unlike a monetary remedy, which can be directly discounted by the patent holder's chance of success to determine an appropriate settlement figure, discounting the possibility of injunctive relief for settlement is not so straightforward. Although the parties can attempt to value the possibility of injunctive relief in monetary terms, both uncertainty and idiosyncratic considerations are likely to complicate that effort. More importantly, patent holders and alleged infringers calculate the value of an injunction from radically different perspectives. For the patent holder, the value of an injunction consists in the additional rents earned if the would-be competitor is successfully excluded from the market. In contrast, an alleged infringer values the possibility of injunctive relief based upon the rents that he expects to earn if successful in entering the market. Because the market will become more competitive after entry, successful entry will reduce the total rents available. The alleged infringer, if allowed to enter, will thus not only split the producer surplus pie with the patent holder, but will also reduce the size of the pie available. As a result, the alleged infringer's

---

maximizing decisions of the parties will create a strong bias for a 50% success rate for plaintiffs or appellants at trial). But see Theodore Eisenberg, *Testing the Selection Effect: A New Theoretical Framework with Empirical Tests*, 19 J Legal Stud 337 (1990) (criticizing Priest-Klein methodology and rejecting their 50% hypothesis as a description of all civil litigation).

[24] These alternative explanations would include the possibility that patent holders sometimes pursue weak claims of infringement in order to establish or maintain a litigious reputation. Alternatively, litigants may not be fully informed regarding their chances of success. For example, the parties involved may not be as fully informed as the available information would permit regarding their respective likelihood of success. Plaintiffs' attorneys, for example, may overestimate their chances of success, choosing to rely on the Federal Circuit's general pro-patent holder reputation to ensure their victory rather than undertaking a more realistic appraisal. Such explanations, while possibilities, are not entirely satisfactory, however. For such an explanation to account for the observed change in success rates, we would not only have to assume that the parties involved are relatively uninformed, but that they became relatively less informed after the advent of the Federal Circuit, and particularly so since 1989.

expected share of the rents available after entry will necessarily prove smaller, and depending on the precise market conditions at issue, potentially much smaller than the patent holder's expected loss should entry occur.[25] As a result, depending on the market conditions expected if entry occurs, the parties may prove unable to find mutually acceptable settlement terms even in cases where both parties agree that the patent holder has an objectively small chance of success.[26]

Because these and other considerations, in combination with the parties' ability to mutually select cases for appellate resolution, will tend to dictate an expected norm for patent holder success rates,[27] we

[25] To take one possible example, if the parties expect to engage in Bertrand competition by offering identical products and competing for customers on prices if the alleged infringer is allowed to enter the market, then both parties will expect price to fall to marginal cost if entry occurs. Under such competition, neither the patent holder nor the alleged infringed will earn any producer surplus if entry occurs. The alleged infringer could therefore offer nothing more than the expected litigation costs in return for dismissal of the patent infringement action. Even if the patent holder has an objectively small chance of successfully excluding the alleged infringer from the market, the minimum amount the patent holder would accept to dismiss the infringement action may exceed the alleged infringer's maximum settlement offer. While Bertrand competition is not inevitable in every market where entry occurs, it illustrates the type of sharp reduction in expected total rents that may leave an alleged infringer unable to offer a price sufficient to purchase from the patent holder even an objectively small chance of obtaining injunctive relief.

[26] As a theoretical matter, the parties always have room for settlement if they can agree on the patent holder's chance of success. On the one hand, the parties have room for a settlement that would allow the alleged infringer to enter the market in return for a payment from the alleged infringer if: the rents lost by the patent holder if entry is allowed multiplied by the expected chance of success are less than the rents that the alleged infringer expects to earn from entry. On the other hand, the parties have room for a settlement in which the alleged infringer agrees not to enter the market in return for a payment from the patent holder if: the rents lost by the patent holder if entry is allowed multiplied by the expected chance of success are greater than the rents that the alleged infringer expects to earn from entry. For that reason, at least in theory, there should always be room for a settlement of one type or the other. However, in my experience, patent holders are seldom willing to settle on terms that require them (as they see it) to pay to enforce their patents. Although a full consideration of the issue is beyond the scope of this article, presumably such a refusal to pay can be justified as a perfectly rational desire to avoid creating a parade of would-be competitors, each demanding a similar payment not to enter.

[27] The relevant question for our purposes is whether the considerations that may drive a patent holder to pursue a claim on which she is unlikely to succeed have both: (i) changed; and (ii) changed at the same times as the observed changes in success rates. If the considerations that may drive a patent holder to pursue an objectively weak claim have changed and at the same times as the observed changes in success rate, then the reduction in average patent holder success rating under the Federal Circuit may not be due solely to some underlying doctrinal change. Nevertheless, as we shall see, see text accompanying notes 35-131, the statistical picture is consistent with the Federal Circuit's doctrinal changes. As a result, even if the statistics alone do not establish a causal relationship, together with these other considerations, they present a strong cir-

must take care not to read too much regarding patent enforceability into appellate litigation "success" statistics alone. However more or less willing to enforce patents the Federal Circuit may be relative to the circuit courts that it replaced, we should expect parties to adjust for that willingness in their settlement negotiations, leaving for appellate resolution only those cases where the benefit of an appeal for both parties exceeds the cost of foregoing the alternative of private settlement. To the extent that a thirty percent success rate appears to represent the point at which both parties prefer appeal to settlement before the advent of the Federal Circuit, we should expect the appellate success rate to remain roughly constant after the Federal Circuit's advent, *ceteris paribus*.[28]

Nevertheless, if we focus on the reasons why patent claims have failed, we find two sharp differences between the Federal Circuit and the circuit courts that it replaced. First, the Federal Circuit is far less likely to reject a claim of patent infringement on the grounds that the patent at issue is invalid or otherwise unenforceable. Second, the Federal Circuit is far more likely to reject a claim of patent infringement on the grounds that the patent at issue was not infringed.

Figure 2 presents the percentage of cases in which claims of patent infringement failed because the patent(s) at issue was found invalid or otherwise unenforceable, both on average and for the particular six pre- and nine post-Federal Circuit periods under study.

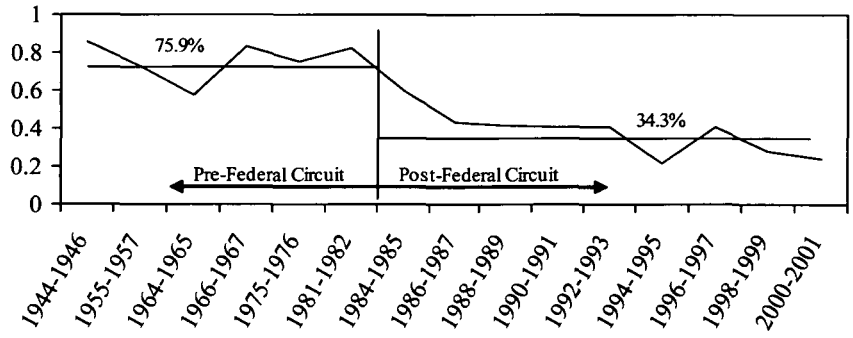

Figure 2. Percentage of Failure Results Due to Ruling that Patent Was Invalid or Otherwise Unenforceable

---

cumstantial case that the Federal Circuit is changing patent law to reduce, on average, patent enforceability.

[28] While the mutual selection of cases for appellate resolution limits our ability to interpret success rates as evidence of enforceability directly, to the extent that mutual selection dictates an expected norm, the extent to which success rates vary from the expected norm are direct evidence of the parties' abilities to predict appellate resolution accurately. For a discussion of this issue, see text accompanying notes 191-196.

As Figure 2 reveals, patent invalidity is significantly[29] less likely to be the reason why a claim of patent infringement fails under the Federal Circuit. Before the Federal Circuit, invalidity accounted for fully three-quarters of the cases in which claims of patent infringement failed. In contrast, after the Federal Circuit's creation, invalidity accounted for just more than one-third of the failure results. If we examine more carefully this reduction in invalidity results, we find a sharply reduced role for the nonobviousness requirement as the principal reason. In the pre-Federal Circuit era, a ruling that the patent claim(s) at issue was invalid due to obviousness accounted for 64.8 percent of the failure results. In contrast, under the Federal Circuit, a failure to satisfy the nonobviousness requirement accounted for only 14.6 percent of the failure results.

Figure 3, on the other hand, presents the percentage of cases in which a claim of patent infringement fails because, as either a legal or factual matter, the allegedly infringing device or process was found to fall outside the scope of the patent claims at issue.

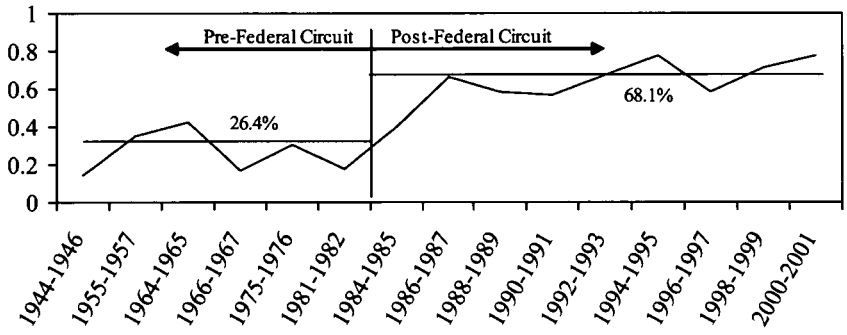

Figure 3. Percentage of Failure Results Due to Ruling that Patent Was Not Infringed

As Figure 3 reflects, an inability to establish infringement is significantly[30] more likely to be the reason why a claim of patent infringement fails under the Federal Circuit. In the pre-Federal Circuit era, a failure to establish infringement accounted for roughly one-

[29] Treating the failure due to invalidity results for the six pre-Federal Circuit and nine post-Federal Circuit periods as discrete data points, and assuming a normal distribution of the data, the 95 percent confidence intervals were ±12.0 percent for the pre-Federal Circuit periods and ±9.4 percent for the post-Federal Circuit periods.

[30] Treating the failure due to non-infringement results for the six pre-Federal Circuit and nine post-Federal Circuit periods as discrete data points, and assuming a normal distribution, the 95 percent confidence intervals were ±13.4 percent for the pre-Federal Circuit periods and ±10.2 percent for the post-Federal Circuit periods.

quarter[31] of the cases in which claims of patent infringement failed. However, under the Federal Circuit, an inability to establish infringement accounted for more than two-thirds of the failure results.

Although (again) we must be careful of drawing inferences too readily from appellate patent litigation results because of the self-selection problem,[32] the timing and extent of the changes in the relative importance of invalidity and non-infringement in explaining why patent claims fail leave the indelible impression that the Federal Circuit is deliberately and systematically changing the nature of patent law. As the pre-Federal Circuit statistics reflect, appellate courts before the Federal Circuit actively policed the prerequisites for patent protection, strictly enforcing the nonobviousness requirement in particular to ensure that a discovery represented a substantial technical advance before it received a patent monopoly. However, once a discovery satisfied the stringent requirements for a patent, appellate courts recognized a correspondingly broad patent scope so that the patent effectively excluded most would-be competitors. At a purely subjective level, the single resolution most representative of this era was a finding by the district court (usually after a bench trial) that the patent was invalid, but if valid, infringed, subsequently affirmed on appeal on the grounds of patent invalidity. In contrast to this traditional approach, the Federal Circuit has made it far easier for a patent holder to repel challenges[33] to her patent's validity. By eviscerating the nonobviousness requirement, the Federal Circuit has substantially reduced the level of creativity required to establish a valid pat-

[31] Some cases involved multiple infringement claims, some of which failed due to invalidity and some of which failed due to an inability to establish infringement. The percentages in Figures 2 and 3 do not therefore necessarily sum to one.

[32] As I have explained elsewhere, one of the principal costs to a patent holder of bringing an infringement action is the possibility that her patent will be declared invalid. Because the Federal Circuit has sharply reduced the chance of such a result, patent holders may find infringement claims with objectively smaller chances of succeeding on the infringement element. Lunney, 7 Mich Telecommun & Tech L Rev 363 at 374 n.43, 384 (cited in note 16). This may account for some of the increase in the percentages of failed claims that fail due to an inability to establish that the allegedly infringing device or process falls within the patent's right to exclude, both directly and by reducing the asymmetric stakes otherwise present where the patent holder faces a significant chance of patent invalidity. Nevertheless, given the magnitude of the change in non-infringement results and the expense entailed in patent litigation, it seems unlikely that patent holders' decisions to begin bringing relatively weak infringement claims accounts for much of the changes in failure results reflected in Figures 2 and 3.

[33] Although the phrasing "repel challenges" may seem curious, under the Patent Act of 1952, a patent, once issued, is presumed valid. 35 USC § 282 (2002). Therefore, the burden is on the alleged infringer to prove the patent invalid. See id. As a result, a patent holder need not prove her patent valid. A patent holder need only repel those challenges to her patent's validity that the alleged infringer raises.

ent. At the same time, while the Federal Circuit has routinely upheld patents even for minor advances, the Federal Circuit has also limited patents to a correspondingly narrow scope. Again, at a purely subjective level, the single resolution most representative of the Federal Circuit era was a ruling, as a matter of law, that the patent was valid, but not infringed.

III. EXPLAINING THE STATISTICS: DOCTRINAL CHANGES UNDER THE FEDERAL CIRCUIT

When we look beyond the statistics, we find a number of doctrinal changes that have driven this shift towards narrower, but more readily valid, patents. We begin with the doctrinal changes that the Federal Circuit has made with respect to validity, before moving to the doctrinal changes the Federal Circuit and the Court have made with respect to infringement.

A. The Federal Circuit's Doctrinal Changes to the Validity Inquiry and the Court's Silent Acquiescence

The formal statutory requirements for obtaining a patent have remained essentially unchanged since the enactment of the Patent Act of 1952. In order to obtain a valid patent, an individual must invent a "useful process, machine, manufacture, or composition of matter"[34] that represents a novel[35] and nonobvious[36] advance over existing technology (or "prior art"). In addition, the inventor must also file a patent application before the invention has been on sale, in public use, patented, or described in a printed publication for one year.[37] The patent application must describe the invention in sufficient detail to enable a person of ordinary skill in the art to practice the invention without undue experimentation and "shall conclude with one or more claims particularly pointing out and distinctly claiming the invention."[38] After the Patent and Trademark Office ("PTO") has examined the application and determined that it satisfies the requirements of the Patent Act, a patent will issue.[39] Once issued, a patent is presumed valid[40] and, in any subsequent litigation, the burden of

[34] 35 USC § 101 (2003).
[35] 35 USC § 102(a) (2003).
[36] 35 USC § 103 (2003).
[37] 35 USC § 102(b) (2003).
[38] 35 USC § 112, ¶ 2 (2003).
[39] 35 USC §§ 131, 151 (2003).
[40] 35 USC § 282 (2003).

proof is on the alleged infringer to establish, by clear and convincing evidence, the patent's invalidity.[41]

Yet, if the formal statutory requirements have remained unchanged, the Federal Circuit has used its exclusive jurisdiction over patent appeals to re-interpret some of these statutory requirements to ensure more readily the validity of litigated patents. Of these doctrinal changes, two appear most significant to the sharp reduction in invalidity results reflected in Figure 2.

First, the Federal Circuit has relentlessly enforced the presumption of validity for issued patents. For example, before the Federal Circuit's advent, courts had weakened the presumption of validity where an alleged infringer presented evidence of prior art that was not before the PTO when the patent issued. Where the pertinent prior art was not presented to the PTO, there would seem to be no basis for deferring to the PTO and its acknowledged technical expertise. As a result, before the Federal Circuit's advent, courts stated that "even one prior art reference not considered by the Patent Office can suffice to overthrow the presumption."[42] While the burden of persuasion remained on the alleged infringer to demonstrate invalidity, the quantum of proof required to demonstrate invalidity became "less stringent."[43]

After assuming the exclusive responsibility for patent appeals, the Federal Circuit moved quickly to reject any weakening of the presumption of validity. In 1983, in *American Hoist & Derrick Co. v. Sowa & Sons, Inc.*,[44] the Federal Circuit acknowledged that there was "no reason to defer to the PTO" once an alleged infringer introduced evidence of prior art not considered by the PTO. "But," the court continued, such evidence "has no effect on the presumption [of validity] or on who has the burden of proof."[45] Nor "does the standard of proof

[41] See, e.g., *American Hoist & Derrick Co v Sowa & Sons, Inc*, 725 F2d 1350 (Fed Cir 1983); *Stratoflex, Inc v Aeroquip Corp*, 713 F2d 1530, 1534 (Fed Cir 1983).

[42] *Henry Mfg Co v Commercial Filters Corp*, 489 F2d 1008, 1013 (7th Cir 1972); see also *Baumstimler v Rankin*, 677 F2d 1061, 1066 (5th Cir 1982).

[43] *Dickey-John Corp v Intl Tapetronics Corp*, 710 F2d 329, 337 (7th Cir 1983). As the Fifth Circuit explained:

> Where the validity of a patent is challenged for failure to consider prior art, the bases for the presumption of validity, the acknowledged experience and expertise of the Patent Office personnel and the recognition that patent approval is a species of administrative determination supported by evidence, no longer exist and thus the challenger of the validity of the patent need no longer bear the heavy burden of establishing invalidity either "beyond a reasonable doubt" or "by clear and convincing evidence."

Baumstimler v Rankin, 677 F2d 1061, 1066 (5th Cir 1982).

[44] 725 F2d 1350 (Fed Cir 1983).

[45] *American Hoist & Derrick Co*, 725 F2d at 1359-60.

change."[46] While evidence of prior art that the PTO did not consider can establish a patent's invalidity, the burden remained on the challenger to establish invalidity by clear and convincing evidence.[47]

Second, in addition to relentlessly enforcing the presumption of validity, the Federal Circuit has also reduced sharply the extent of the technological advance required to sustain a patent. Under sections 102(a) and 103 of the Patent Act,[48] an invention must represent a novel and nonobvious advance over the prior art. Under section 102(a), an invention is new or novel unless "all of the elements and limitations of the claim are found within a single prior art reference."[49] For the prior art disclosure to anticipate a claimed invention, "[t]here must be no difference between the claimed invention and the reference disclosure, viewed by a person of ordinary skill in the field of invention."[50] In contrast, under section 103, we ask whether the claimed invention "would have been obvious to one of ordinary skill in the art . . . in view of the teachings of the prior art as a whole."[51] Because the nonobviousness requirement compares the claimed invention to the prior art as a whole, rather than to a single reference, and because it can bar patentability even if there are differences between the claimed invention and the prior art, nonobviousness has traditionally represented the principal substantive hurdle for patentability. This is clearly reflected in our statistical analysis of appellate patent results. As discussed, on its own, a failure to satisfy the nonobviousness requirement accounted for 64.8 percent of the cases in our pre-Federal Circuit sample where claims of patent infringement failed.

Application of the nonobviousness requirement in cases involving so-called combination patents perhaps best reflected the stringency of the requirement before the Federal Circuit's advent.[52] Because we determine nonobviousness against "the teachings of the prior art as a whole,"[53] if the prior art encompassed each element of a claimed invention, then the prior art would seem to encompass the claimed in-

[46] Id at 1360.; see also *Stratoflex, Inc v Aeroquip Corp*, 713 F2d 1530, 1534 (Fed Cir 1983).

[47] *American Hoist & Derrick Co*, 725 F2d at 1360.

[48] 35 USC § 103 (2003).

[49] *Scripps Clinic & Research Found v Genentech, Inc*, 927 F2d 1565, 1576 (Fed Cir 1991) (citations omitted).

[50] Id.; see also *Carella v Starlight Archery & Pro Line Co*, 804 F2d 135, 138 (Fed Cir 1986).

[51] *Stratoflex, Inc*, 713 F2d 1530 at 1537.

[52] See, e.g., *Sakraida v Ag Pro, Inc*, 425 US 273 (1976); *Anderson's-Black Rock v Pavement Salvage Co*, 396 US 57, 61 (1969); *Great A & P Tea Co v Supermarket Equip Corp*, 340 US 147 (1950).

[53] *Stratoflex, Inc*, 713 F2d at 1537.

vention as a whole. Perhaps for that reason, where a patent claimed a combination of preexisting elements, with each element well-known in the prior art, the Court would presume that the combination was obvious. Only where the combination produced an unexpected or "synergetic" result would the Court find that a combination of preexisting elements satisfied the nonobviousness requirement.[54] For example, in 1976, the Court considered the validity of a patent covering a water flush system to remove cow manure from the floor of a dairy barn.[55] Noting that the patent consisted of a combination of preexisting elements, such as the storage of water in tanks or pools, the Court held that the combination could not "properly be characterized as synergistic" because it did not "'result in an effect greater than the sum of the several effects taken separately.'"[56] "[T]his patent simply arranges old elements with each performing the same function it had been known to perform. . . ."[57] Although the Court acknowledged that the patent holder's flush system "produc[ed] a desired result in a cheaper and faster way, and enjoy[ed] commercial success," these secondary considerations were not sufficient to demonstrate nonobviousness. Because the claimed invention "'did not produce a "new or different function,"'" the flush system lacked the requisite synergy required for combination patents and was therefore obvious.[58]

Again, the Federal Circuit acted quickly, rejecting the Court's "synergism" requirement for combination patents in another 1983 decision, *Stratoflex, Inc. v. Aeroquip Corp.*[59] In refusing to follow the Court's approach, Chief Judge Markey, writing for the panel, stated that:

> A requirement for "synergism" or a "synergistic effect" is nowhere found in the statute. . . . The reference to a "combination patent" is equally without support in the statute . . . [and] is moreover meaningless. Virtually all patents are "combination patents," if by that label one intends to describe patents having claims to inventions formed of a combination of elements.[60]

[54] See *United States v Adams*, 383 US 39 (1966) (ruling that patent was nonobvious where combination of prior art elements "'wholly unexpectedly' has shown 'certain valuable operating advantages over other batteries' while those from which it is claimed to have been copied were long ago discarded").

[55] *Sakraida*, 425 US at 280-81.

[56] Id at 282 (quoting *Anderson's-Black Rock v Pavement Salvage Co*, 397 US 57, 61 (1969)).

[57] Id

[58] Id at 282-83 (quoting *Anderson's-Black Rock*, 397 US at 60).

[59] 713 F2d 1530 (Fed Cir 1983) (Markey, Ch. J); see also Dreyfuss, 64 NYU L Rev at 9-10 (cited in note 12).

[60] *Stratoflex, Inc*, 713 F2d at 1540.

While Chief Judge Markey's assertion regarding the lack of express statutory language is accurate, any number of patent law doctrines derive from Court decisions without express statutory support, including, for example, the clear and convincing evidence standard for overcoming the presumption of validity[61] and the experimental use gloss on section 102(b)'s public use bar.[62] Similarly, although his "virtually all" assertion overstated the traditional applicability of the combination patent doctrine, its general applicability was the point. Given that nonobviousness is judged from the perspective of the prior art as a whole, the ability to combine prior art references was and is an integral aspect of the nonobviousness inquiry.

Having rejected the Court's presumption that different prior art references can be freely combined for purposes of determining obviousness, the question becomes: When can references be combined? Although the *Stratoflex, Inc.* panel rejected the Court's approach, it did not articulate its own standard. That standard would come the following year, in *ACS Hospital Systems, Inc. v. Montefiore Hospital*,[63] when the Federal Circuit stumbled[64] upon a formulation that would substantially restrict the practice of combining prior art references in order to establish a patented invention's obviousness. Under the *ACS Hospital Systems, Inc.* panel's approach, prior art references may be combined to demonstrate the obviousness of a patented invention "*only* if there is some suggestion or incentive to do so" in the prior art.[65] Although the Federal Circuit has sometimes permitted

[61] Although the Patent Act provides for a presumption of validity, 35 USC § 282 (2003), the Patent Act does not state that the presumption can be overcome only by clear and convincing evidence. That evidentiary standard derives from the Court's decisions in *Coffin v Ogden*, 85 US (18 Wall) 120, 124 (Oct term 1873), and *The Barbed Wire Patent*, 143 US 275 (1892).

[62] In *City of Elizabeth v American Nicholson Pavement Co*, 97 US 126 (1877), the Court considered whether the patent holder's use of a road paving method to construct a toll road more than six years before he filed his patent application barred the patent under the corresponding public use provision then in force. The Court held that because the inventor's use was "by way of experiment, and in order to bring the invention to perfection," *City of Elizabeth*, 97 US at 134, it was not a public use within the meaning of the statutory bar.

[63] 732 F2d 1572 (Fed Cir 1984).

[64] Although the panel cited a handful of decisions in footnotes 13 and 14 of its opinion as support for its statement of the "teaching or suggestion" test for combining references, *ACS Hospital Sys, Inc*, 732 F2d at 1577 nn. 13, 14, none of the cited cases recite or otherwise appear to support such a test. Whether the *ACS Hospital Sys., Inc* panel intended to formulate a new test or was simply incapable of summarizing the then-existing standards is unclear.

[65] *ACS Hospital Sys, Inc*, 732 F2d at 1577. For more recent applications of the doctrine, see *Robotic Vision Systems, Inc v View Eng'g, Inc*, 189 F3d 1370, 1377 (Fed Cir 1999) ("The party seeking a holding of invalidity based on a combination of two or

"implicit" suggestions to satisfy the standard,[66] the Federal Circuit's approach reverses the key presumption in these cases. Where all of the elements were known, the Court's "synergism" approach presumed that any given combination was obvious, unless there was some reason that suggested otherwise. The Federal Circuit, on the other hand, presumes that any given combination is nonobvious, unless there is some suggestion in the prior art otherwise.

In addition to rejecting the Court's approach to combining prior art references, Chief Judge Markey was also dissatisfied with the Court's approach to the so-called "secondary considerations." In *Graham v. John Deere Co.*,[67] the Court had acknowledged that "[s]uch secondary considerations as commercial success, long felt but unsolved needs, failure of others, etc., . . . may have relevancy" "as indicia of obviousness or nonobviousness."[68] Nevertheless, while the *Graham* Court stated that these factors "might be utilized," such secondary considerations were insufficient in *Graham* itself to "tip the scales of patentability" where the invention as whole otherwise appeared obvious.[69] Until the advent of the Federal Circuit, the various circuits universally read *Graham* as requiring obviousness to be determined primarily based upon a three-part inquiry, consisting of: (1) defining the prior art; (2) identifying the differences between the claimed invention and the prior art; and (3) determining the level of ordinary skill in the art; with the secondary considerations relegated to a subsidiary role. In keeping with their reading of *Graham*, the various circuits consistently held that it was not reversible error for a trial court

---

more prior art teachings must show some motivation or suggestion to combine the teachings."); *Micro Chem, Inc v Great Plains Chem Co, Inc*, 103 F3d 1538, 1546 (Fed Cir 1997) ("A determination of obviousness must involve more than indiscriminately combining prior art; a motivation or suggestion to combine must exist.") (citation omitted); *In re Laskowski*, 871 F2d 115, 117 (Fed Cir 1989) (reversing PTO rejection of patent application for band saw wheel because although prior art contained each element set forth in the patent claims, the prior art did not contain any suggestion to combine the elements in the manner set forth in the claims); see also *In re Oetiker*, 977 F2d 1443 (Fed Cir 1992) (Nies, C. J., concurring) ("While there must be some teaching, reason, suggestion, or motivation to combine existing elements to produce the claimed device, it is not necessary that the cited references or prior art specifically suggest making the combination.").

[66] See, e.g., *Riverwood Intl Corp. v Mead Corp.*, 212 F3d 1365, 1366 (Fed Cir 2000) ("In addition, where obviousness is based on particular prior art references, there must be a showing of a suggestion or motivation to combine the teachings of those references, though it need not be expressly stated.").

[67] 383 US 1 (1966).

[68] *Graham v. John Deere, Co*, 383 US 1, 17-18 (1966).

[69] Id at 36; see also *Sakraida v Ag Pro, Inc*, 425 US 273, 282-83 (1976); *Anderson's-Black Rock v Pavement Salvage Co*, 396 US 57, 60 (1969); *Great Atlantic & Pacific Tea Co v Supermarket Corp*, 340 US 147, 153 (1950).

to fail to consider evidence of secondary considerations,[70] and would allow such evidence to "tip the scales" in favor of nonobviousness only in close cases where the three-factor *Graham* inquiry "[did] not produce a firm conclusion."[71] Following these rules, the district court in *Stratoflex, Inc.* made findings regarding the secondary considerations, but did not include them in her analysis because she determined the invention as a whole to be obvious based upon the three-part *Graham* inquiry.[72] As the Court had stated in *Sakraida:* "Though doubtless a matter of great convenience, producing a desired result in a cheaper and faster way, and enjoying commercial success, . . . [t]hese desirable benefits 'without invention will not make patentability.'"[73]

On appeal, Chief Judge Markey refused to treat the secondary considerations as secondary, insisting that evidence of secondary considerations is often "the most probative and cogent evidence in the record"[74] and that it "must always when present be considered."[75] Following Chief Judge Markey's lead, the Federal Circuit has held that commercial success and the other secondary considerations, although not conclusive on the issue of nonobviousness,[76] are a cen-

[70] See, e.g., *Stevenson v Grentec, Inc,* 652 F2d 20, 23 (9th Cir 1970) (ruling that a "failure to consider secondary factors [in determining obviousness] [was] not reversible error").

[71] *Digitronics Corp. v New York Racing Ass'n,* 553 F2d 740, 748-49 (2d Cir 1977); see also *Sakraida v Ag Pro, Inc,* 425 US 273, 282-83 (1976) ("Though doubtless a matter of great convenience, producing a desired result in a cheaper and faster way, and enjoying commercial success, . . . [t]hese desirable benefits 'without invention will not make patentability.'") (quoting *Great A&P Tea Co v Supermarket Equip Corp,* 340 US 147, 153 (1950)); *Medical Lab Automation, Inc v Labcon, Inc,* 670 F2d 671, 675 (7th Cir 1981).

[72] See *Stratoflex, Inc v Aeroquip Corp.,* 713 F2d 1530, 1539 (Fed Cir 1983).

[73] *Sakraida,* 425 US at 282-83 (quoting *Great A&P Tea Co v Supermarket Equip Corp,* 340 US 147, 153 (1950)); see also *Medical Lab Automation, Inc v Labcon, Inc,* 670 F2d 671, 675 (7th Cir 1981).

[74] *Stratoflex, Inc* 713 F2d at 1538-39.

[75] *Stratoflex, Inc v Aeroquip Corp,* 713 F2d 1530, 1538 (Fed Cir 1983); see also *Simmons Fastener Corp. v Illinois Tool Works, Inc,* 739 F2d 1573 (Fed Cir 1984) (reversing finding of obviousness for failure to consider evidence of secondary considerations); *WL Gore & Assocs. v Garlock, Inc,* 721 F2d 1540, 1555 (Fed Cir 1983).

[76] See *Richardson-Vicks, Inc v Upjohn Co,* 122 F3d 1476, 1481-84 (Fed Cir 1997) (ruling that patented invention was obvious despite evidence of commercial success and other secondary considerations); *Motorola, Inc v Interdigital Technology Corp,* 121 F3d 1461, 1472 (Fed Cir 1997) ("In reaching an obviousness determination, a trial court may conclude that a patent claim is obvious, even in the light of strong objective evidence tending to show non-obviousness."); *BF Goodrich Co v Aircraft Braking Sys Corp,* 72 F3d 1577, 1583 (Fed Cir 1996) ("Considering the minor difference between the claimed invention and the [prior art], the secondary considerations were not sufficiently compelling" to preclude a conclusion of obviousness.); *Newell Cos. v Kenney Mfg Co,* 864 F2d 757, 768-69 (Fed Cir 1988) (noting that secondary considerations "must be considered, [but] they do not control the obviousness conclusion").

tral, rather than secondary, factor in the obviousness inquiry.[77] In keeping with this more central role, the Federal Circuit has renamed these considerations, preferring the label "objective evidence of nonobviousness" rather than *Graham*'s label of "secondary considerations."[78] The Federal Circuit has also: (1) employed a broader range of secondary considerations as proof of nonobviousness[79]; (2) relaxed the required showing that the commercial success was the result of the nonobvious nature of the claimed invention, rather than some other factor, such as marketing[80]; and (3) restricted attempts to use secondary considerations (or the lack thereof) to establish that a patent was obvious.[81]

Like the Federal Circuit's substitution of its own suggestion test for the Court's synergy approach, the Federal Circuit's increased reliance on secondary considerations tends to reduce directly the likelihood that a litigated patent will be found obvious. As Professor Edmund Kitch warned more than thirty years ago, an increased reliance on secondary considerations, such as commercial success, to resolve questions of patent validity almost necessarily leads to a rule "that all patents that are litigated should be held valid."[82] As Professor

[77] See, e.g., *Uniroyal, Inc v Rudkin-Wiley Corp*, 837 F2d 1044 (Fed Cir) (reversing district court's finding of obviousness for failing to give more weight to evidence of secondary considerations); *Alco Standard Corp. v Tennessee Valley Authority*, 808 F2d 1490, 1492, 1499-1501 (Fed Cir 1986) (holding patent nonobvious on basis of secondary considerations despite the fact that the three-factor inquiry strongly suggested that patent was obvious in light of prior art). Indeed, the Federal Circuit has taken to identifying secondary considerations as a fourth *Graham* factor. See, e.g., *Robotic Vision Systems, Inc v View Eng'g, Inc*, 189 F3d 1370, 1376 (Fed Cir 1999); *Modine Mfg Co v Allen Group, Inc*, 917 F2d 538, 541 (Fed Cir 1990); *Loctite Corp. v Ultraseal, Ltd*, 781 F2d 861, 872-73 (Fed Cir 1985); *Oscar Mayer Foods Corp. v Con-Agra, Inc*, 35 USPQ 2d 1278, (Fed Cir 1994) ("Obviousness is a question of law with four factual predicates").

[78] See, e.g., *Gillette Co v S.C. Johnson & Son, Inc*, 919 F2d 720, 725 (Fed Cir 1990); Modine Mfg., 917 F2d at 541.

[79] See Kevin Rhodes, Comment, *The Federal Circuit's Patent Nonobviousness Standards: Theoretical Perspectives on Recent Doctrinal Changes*, 85 Nw U L Rev1051, 1071-72 (1991).

[80] See, e.g., Merges, 76 Calif L Rev at 824-25 (cited in note 12).

[81] Compare *Graham*,383 US at 18 (noting that secondary considerations "may have relevancy" "as indicia of obviousness or nonobviousness"); with *Gentry Gallery, Inc v Berkline Corp*, 134 F3d 1473, 1478 (Fed Cir 1998) (noting that evidence of secondary factors "can only further support nonobviousness") and *Custom Accessories, Inc v Jeffrey-Allan Indus, Inc*, 807 F2d 955, 960 (Fed Cir 1986) (holding that absence of commercial development or other secondary considerations is not evidence of obviousness, but only "a neutral factor"). Compare *Concrete Appliances Co v Gomery*, 269 US 177, 185 (1925) (relying on near-simultaneous invention by others to support ruling of obviousness) *and Fred Whitaker Co v E. T. Barwick Indus, Inc*, 551 F2d 622, 628 (5th Cir 1977); with *Environmental Designs, Ltd v Union Oil Co*, 713 F2d 693, 698 (Fed Cir 1983) (ruling that evidence of near-simultaneous invention by others not evidence of obviousness).

[82] Edmund Kitch, *Graham v John Deere Co: New Standards for Patents*, 1966 S Ct Rev 293.

Kitch explained, "it is unlikely that patents that are not commercially successful will be brought to litigation."[83] As a result, to the extent that commercial success becomes an important factor in determining a patent's validity, the very fact that the patent is worth litigating should establish its validity.[84]

Despite the fact that the Federal Circuit has overlooked, rewritten, and in some cases, expressly rejected the Court's interpretation of the nonobviousness doctrine, the Court has so far refused all invitations to reexamine the Federal Circuit's new nonobviousness doctrine. We can attribute this in part to the Federal Circuit's practice of reserving its most outrageous overreaching for those cases where a certiorari petition is unlikely. In "reversing" *Sakraida*, for example, the *Stratoflex* panel held that even under its substantially diminished nonobviousness requirement, the patented invention was obvious. By holding the patented invention obvious, the panel ensured that neither the plaintiff nor the defendant had an incentive to petition the Court for review: the defendant because it won; the plaintiff because it could not persuasively argue that its patent would have been valid if only the panel had applied the Court's more stringent nonobviousness standards. Safely insulated from Court review, the *Stratoflex* panel decision remained on the books and became binding on later Federal Circuit panels.

More generally, even where the Federal Circuit has not expressly rejected the Court's rulings, decisions of the Court have proven far less binding on the Federal Circuit than they have on the other Circuits. Given its exclusive jurisdiction over patent appeals, the Federal Circuit need not worry that another Circuit will reveal and ridicule the Federal Circuit's "rewriting" of Court precedent.[85] If circuit conflicts are an important signal to the Court in deciding which appellate decisions to review, then the absence of such "competition" may have given the Federal Circuit correspondingly greater leeway to work around seemingly binding Court authority.[86]

In any event, whatever the reason for the Court's silence on the obviousness issue, the net result has been that nearly thirty years has

[83] Id.

[84] Id.

[85] The Court's decision in *Holmes* re-establishes the possibility of such competition. See *Holmes Group, Inc v Vornado Air Circulation Systems, Inc*, 535 US 826 (2002) (holding that the Federal Circuit has exclusive jurisdiction over patent appeals only to the extent that the action for patent infringement appears in the complaint under the well-pleaded complaint rule).

[86] See *Holmes Group, Inc*, 535 US at 839 (Stevens, J, concurring) ("An occasional conflict in decisions may be useful in identifying questions that merit this Court's attention. Moreover, occasional decisions by courts with broader jurisdiction will provide an antidote to the risk that the specialized court may develop an institutional bias.")

passed since the Court's last decision on nonobviousness, *Sakraida*, and nearly forty years has passed since the Court's last decision that retains any influence on the issue, *Graham*. Sooner or later the Court will undoubtedly take up the issue again, but to date, the Court has, through its silence, acquiesced to the changes the Federal Circuit has made to the nonobviousness doctrine. Even under the Federal Circuit, the nonobvious requirement retains some semblance of life, as it remains expressly present in section 103 of the Patent Act and has continued to serve as a basis for finding a patent invalid in some cases.[87] However, it appears that nonobviousness's once dominant vitality has been substantially diminished. Where before the Federal Circuit's advent only substantial advances would satisfy the requirement, under the Federal Circuit, even slight advances over the prior art will likely prove nonobvious. Both the statistics and the doctrinal changes thus reflect a shift toward more routinely valid patents. In the next section, we consider the doctrinal changes to the infringement inquiry that have effectively narrowed the scope of existing patents.

B. Doctrinal Changes to the Infringement Inquiry: The Federal Circuit Acts and the Court Responds

Of the three federal statutes generally considered part of intellectual property, only one—the Patent Act of 1952—relies on written claims to delineate the scope of the property protected. Under both the Trademark Act of 1946 and the Copyright Act of 1976, we resolve questions of infringement by comparing the allegedly infringing trademark or work to the protected trademark or work of authorship directly. Only in patent law do we resolve the infringement question by comparing the allegedly infringing product or process to a written claim that defines the patent holder's discovery. In patent law, where the allegedly infringing product or process contains each element written in the patent claim, we say that the patent "reads on," and is therefore literally infringed by, the allegedly infringing product or process. Where a patent holder cannot establish literal infringement, a patent holder may nevertheless prevail if she can establish infringement under the so-called doctrine of equivalents.

Formally recognized by the Court in *Winans v. Denmead* in its December 1853 term,[88] the doctrine of equivalents expands the scope

[87] See, e.g., *Georgia Pacific Corp v United States Gypsum Co*, 195 F3d 1322 (Fed Cir 1999); *Richardson-Vicks, Inc v Upjohn Co*, 122 F3d 1476, 1481-84 (Fed Cir 1997) (ruling that patented invention was obvious despite evidence of commercial success and other secondary considerations); *Motorola, Inc v Interdigital Tech Corp.*, 121 F3d 1461 (Fed Cir 1997); *Para-Ordinance Mfg v SGS Importers Intl*, 73 F3d 1085, (Fed Cir 1995).

[88] *Winans v Denmead*, 56 US (15 How) 330 (Dec term 1853).

of a patent beyond its literal terms. Under the doctrine, a patent will extend to those products or processes, that while not literally covered by the patent claims, have elements that perform the same function in the same way to achieve the same result as each element set forth in the patent claim. Although a firmly established part of patent law for one hundred fifty years, the doctrine of equivalents has been controversial from its inception. Setting the terms for a debate that continues to this day, Justice Curtis, writing for the five member majority in *Winans*, insisted that the "property of inventors would be valueless . . . if the public are at liberty to make substantial copies of [the invention, merely by] varying its form and proportions."[89] On the other side, Justice Campbell, writing for the four dissenters, countered that "[n]othing, in the administration of [patent] law, will be more mischievous, more productive of oppressive and costly litigation, of exorbitant and unjust pretensions and vexatious demands, more injurious to labor" than a failure to limit a patent to its literal terms.[90]

Today, the debate over the doctrine of equivalents continues. On the one side, the specter of the "unscrupulous copyist" who will duplicate an invention's substance while avoiding the patent's literal terms—no matter how artfully drafted—cautions against restricting a patent to its literal claims.[91] On the other, the uncertainty and ambiguity that the doctrine injects into the question of infringement seem to contradict directly the statutory language requiring patent applications "particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention."[92]

Although the debate's substance has changed little over the years, the outcome has subtly shifted in the last decade. While Justice Campbell's arguments have not yet fully prevailed, judicial decisions have increasingly recognized the persuasive force of Justice Campbell's concerns and have therefore moved to narrow the doctrine of equivalents. In this effort, the Federal Circuit, after a brief flirtation with the doctrine, has led the way. Initially, in some of its first decisions, the Federal Circuit embraced an expansive view of the doctrine of equivalents, consistent with its pro-patent holder reputation.[93] However, the Federal Circuit's enchantment with the doctrine soon faded,

[89] Id at 342-43.

[90] Id at 347 (Campbell, J, dissenting, with Chief Justice Taney, Justice Catron, and Justice Daniel).

[91] *Graver Tank & Mfg Co v Linde Air Prods Co*, 339 US 605, 607 (1950).

[92] 35 USC § 112 (2002).

[93] See, e.g., *Martin v Barber*, 755 F2d 1564 (Fed Cir 1985); *Carman Indus., Inc v Wahl*, 724 F2d 932 (Fed Cir 1983); *Hughes Aircraft Co v United States*, 717 F2d 1351 (Fed Cir 1983).

and by 1987, the balance of power on the Federal Circuit had shifted towards a narrower view of the doctrine. In that year, the Federal Circuit, sitting en banc in *Pennwalt Corp. v. Durand-Wayland, Inc.*,[94] required a patent holder to prove equivalency on an element-by-element basis, rather than for the invention as a whole. By doing so, the Federal Circuit made it somewhat easier for the unscrupulous copyist to duplicate the substance of a patented invention without infringing the patent's claims.

To illustrate, consider the facts in *Pennwalt*. Pennwalt held a patent on a high-speed fruit sorting machine. Fruit was loaded onto a conveyor belt, and as the fruit traveled along the conveyor belt, sensors determined the weight and color of each piece. At the end of the belt, the fruit was sorted according to either its weight or its weight and color into the appropriate bin. To obtain its patent, Pennwalt could not simply submit its fruit sorter, however. Rather, under the Patent Act, Pennwalt had to draft a patent application "conclud[ing] with one or more claims particularly pointing out and distinctly claiming" its invention.[95] There are obvious difficulties in trying to describe in words an invention such as a fruit sorter in a manner sufficiently general to encompass potentially competing substitutes, yet sufficiently precise to satisfy the statutory requirement. Nevertheless, given the statutory language, Pennwalt had little choice but to make the attempt. Following the usual practice, Pennwalt included in its patent a number of claims of varying specificity, with the hope that should litigate result, at least one would be found both valid and infringed.[96] Although these claims vary, claim 10 is representative:

> 10. An automatic sorting apparatus comprising electronic weighing means for generating a signal proportional to the weight of an item to be sorted, first reference signal means for providing a predetermined number of reference signals, the value of each signal being established according to a predetermined criteria, first comparison means for comparing the signal generated by said electronic weighing means to the reference signals provided by said first reference signals means, optical detection means for generating a signal proportional to the color of an item to be sorted, second reference signal means for providing a predetermined number of reference signals, the value of

[94] 833 F2d 931 (Fed Cir 1987) (en banc), *cert. denied*, 485 US 961 (1988).

[95] 35 USC § 112, ¶ 2 (2002).

[96] Validity and infringement are resolved on a claim-by-claim basis. Each claim is "presumed valid independently of the validity of other claims," 35 USC § 282 (2002), and a defendant need only infringe one claim of a patent to be guilty of patent infringement.

> each signal being established according to a predetermined criteria, second comparison means for comparing the signal generated by said optical detection means to the reference signals provided by said second reference signals means, and generating a signal therefrom, clock means for incrementally signaling changes in the position of the item to be sorted, first position indicating means responsive to a signal from said clock means and said signal from said second comparison means for continuously indicating the position of an item to be sorted while the item is in transit between said optical detection means and said electronic weighing means, second position indicating means responsive to the signal from said clock means, the signal from said first comparison means and said first position indicating means for generating a signal continuously indicative of the position of an item to be sorted after said item has been weighed, and discharge means responsive to the signal from said second position indicating means for discharging the item at a predetermined one of a plurality of sorting positions.

For the sake of simplicity, we can break this claim down and summarize it as a set of elements or limitations,[97] as follows:

> 10. An automatic sorting apparatus comprising: (a) electronic weighing means; (b) weight comparison means; (c) optical color determining means; (d) color comparison means; (e) location-indicating means; and (f) discharge means.

When Durand-Wayland began selling two different types of automatic fruit sorting machines, Pennwalt sued for patent infringement. Although both parties agreed "that the Durand-Wayland machines are substantially the same as the machine described in the patent-in-suit insofar as the results achieved,"[98] there were some differences. Specifically, because Pennwalt developed its invention in the mid-1970s, it relied on hard-wired electronics and analog voltage signals in its pat-

[97] Although we can summarize the elements in this way for the sake of simplicity, I should also note that the claim is drafted using means-plus-function language. Such an approach is expressly authorized in the Patent Act. 35 USC § 112, ¶ 6 (2002). Although such language seems facially quite broad, means-plus-function language does not encompass every possible means to accomplish the specified result. Rather, the Federal Circuit interprets such language "to cover the corresponding structure . . . described in the specification and equivalents thereof." *Intellicall, Inc v Phonometrics, Inc*, 952 F2d 1384 (Fed Cir 1992). For that reason, each of the elements in Pennwalt's patent should be read to incorporate the specific structure (i.e. the physical circuitry) that Pennwalt described in its patent for practicing the invention.

[98] *Pennwalt Corp. v Durand-Wayland, Inc*, 225 USPQ 558 (N D Ga 1984), *aff'd*, 833 F2d 931 (Fed Cir 1987) (en banc).

ented invention. Durand-Wayland's devices, on the other hand, were developed in the 1980s and therefore used a general purpose microprocessor programmed to operate the fruit sorting device. As a result, where Pennwalt's invention kept track of the location of each piece of fruit through an analog signal proportional to the fruit's physical distance along the conveyor, the Durand-Wayland devices kept track of the location of a piece of fruit by its order in the fruit queue.

If the Federal Circuit had continued to evaluate the issue of equivalency by comparing the patented invention and the allegedly infringing device as a whole, there seems little doubt that Pennwalt would have prevailed. Just four years before, the Federal Circuit had held that patent infringement was established under the doctrine of equivalents where a defendant "merely employed a modern day computer to do indirectly what [the patent] taught it to do directly."[99] Yet, four years later, applying the doctrine of equivalents on an element-by-element basis, the Federal Circuit affirmed the district court's finding that Durand-Wayland machines did not infringe Pennwalt's patent either literally or under the doctrine of equivalents.[100] Although the Durand-Wayland devices kept track of each piece of fruit's position on the conveyor, they did so using the fruit's place in line, rather than its physical location along the conveyor. The courts concluded that this was not the same "way" as Pennwalt's patented invention. Because Durand-Wayland's devices lacked an equivalent of element (e), the devices did not infringe Pennwalt's patent,[101] and Durand-Wayland was therefore free to continuing selling its fruit sorters in direct competition with Pennwalt.

Although Pennwalt filed a petition for certiorari, the Court refused at that time to review the Federal Circuit's decision.[102] However, in two more recent cases, the Court has affirmed the Federal Circuit's vision of a narrower doctrine of equivalents. In the first, *Warner-Jenkinson Co. v. Hilton Davis Chemical Co.*,[103] the Court expressly recognized that "the doctrine of equivalents, when applied broadly, conflicts with the definitional and public-notice functions of the statutory claiming requirement."[104] Concerned that "the doctrine of equivalents, as it has come to be applied since *Graver Tank*, has taken on a life of its own, unbounded by the patent claims,"[105] the Court:

[99] *Hughes Aircraft Co v United States*, 717 F2d 1351, 1364 (Fed Cir 1983).

[100] *Pennwalt Corp.*, 833 F2d at 933-39.

[101] Id at 938-39.

[102] *Pennwalt Corp. v Durand-Wayland, Inc*, 485 US 961 (1988) (denying petition for certiorari).

[103] 520 US 17 (1997).

[104] *Warner-Jenkinson Co*, 520 US at 29.

[105] Id at 28-29.

(i) embraced the Federal Circuit's element-by-element approach to equivalency; and (ii) expressly recognized and broadened the reach of the doctrine of prosecution history estoppel.[106]

Classically in the law, where a party has made a representation that another has relied upon, an estoppel will arise that will preclude the party who has made the representation from subsequently acting in a manner contrary to the representation. Courts justify such an estoppel on the grounds that it would be unjust to allow a party to retract a representation once another has reasonably relied upon it. In patent law, the doctrine of prosecution history estoppel provides a similar rule: Where a patent holder has narrowed her patent claims during the application process, she may be estopped from reclaiming the scope given up by her amendment. In essence, the fact that a patent applicant has amended and narrowed her patent claims is taken as a factual statement or representation that the applicant does not intend to claim those devices or processes that fall within the broader claim language, but only those devices or processes that fall within the narrowed claim language. Where another party has relied upon that representation, by manufacturing or using a device that falls just outside the narrower claim language, it would seem similarly unjust to allow the patent holder to withdraw her representation.

Although we can therefore fit prosecution history estoppel comfortably within the general rationale for estoppel, we should not take a simple "representation plus reliance equals estoppel" formula too far. After all, we could also say that every claim, whether or not amended during prosecution, is a representation or statement of the patent scope intended by the applicant. To the extent that another party relies upon that representation, by making or using a device that falls just outside the literal scope of the claims, one could make a similar argument that an estoppel should arise. Although it approaches the issue from a slightly different perspective, such a general "claim" estoppel argument merely restates the reasons Justice Campbell offered for confining a patent's scope to its literal claims. Having decided to retain the doctrine of equivalents, even if in a decidedly narrower form, the Court has necessarily rejected this broader form of the estoppel argument.

The question thus becomes why we apply prosecution history estoppel, yet refuse to apply a general claim estoppel. Two reasons likely justify our willingness to treat narrowing amendments as more binding than we treat claim language generally. First, the fact of an amendment almost necessarily reflects a patent holder's conscious

[106] Id at 28-34.

choice between the original and amended claim language.[107] Moreover, if the amendment was made during a patent's prosecution, there will be a publicly-maintained record of the amendment, and we can establish the nature of the choice made directly by comparing the original and amended language. In contrast, where a claim has not been expressly amended, it is less clear whether a deliberate choice was made between a narrower and broader version of the claim. And even if we believe that such choices are inevitably made by the patent attorney in drafting the initial claim, there is no publicly-available record of the nature of those choices. As a result, determining the nature of the choice made and the broader scope foregone would prove more difficult. In addition, because amendments are usually made in response to feedback from the patent examiner and because amendments are usually focused on a particular phrase or limitation within a much longer claim, it is also possible that patent attorneys are somewhat more careful in drafting claim amendments than they are in drafting the claims generally. To the extent that narrowing amendments reflect more deliberate, careful choices that publicly establish a broader scope foregone and a narrower scope claimed, that might justify treating a narrowing amendment, but not claim language generally, as a representation upon which others can reasonably rely.[108]

Second, limiting the doctrine of equivalents to instances where a narrowing amendment has been made may be better tailored than a general claim estoppel doctrine towards ensuring that a patent holder cannot reclaim through the doctrine of equivalents subject matter to which she is not legally entitled. For example, in *Warner-Jenkinson Co.*, the inventors had developed a filtration process for purification of dyes.[109] During the patent's prosecution, the Hilton Davis inventors added the limitation "at a pH from approximately 6.0 to 9.0" to distinguish their filtration process from a previous patent that disclosed a similar filtration process at a pH above 9.0.[110] Because this prior patent either anticipated or rendered obvious the Hilton Davis

[107] See, e.g., *Festo Corp v Shoketsu Kinzoku Kogyo Kabushiki Co*, 535 US 722, 734-35 (2002) ("[Where a narrowing amendment is made during prosecution,] the prosecution history has established that the inventor turned his attention to the subject matter in question, knew the words for both the broader and narrower claim, and affirmatively chose the latter."); *Exhibit Supply Co v Ace Patents Corp*, 315 US 126, 136-37 (1942) (noting that "by the amendment [the patent holder] recognized and emphasized the difference between the two phrases . . . and the difference which [the patent holder] thus disclaimed must be regarded as material").

[108] See *Festo Corp.*, 535 US at 740 ("A patent holder's decision to narrow his claim through amendment may be presumed to be a general disclaimer of the territory between the original claim and the amended claim.").

[109] *Warner-Jenkinson Co*, 520 US at 23.

[110] Id at 22.

filtration process for a pH above 9.0, the Hilton Davis inventors were not entitled to a patent on their process for a pH above 9.0. To achieve that result, the patent examiner required the Hilton Davis inventors to amend their claim language to exclude the prior art literally. However, for that limitation to prove effective, we must also ensure that Hilton Davis cannot recapture the prior art through the doctrine of equivalents. The doctrine of equivalents already contains a limitation prohibiting a patent holder from using the doctrine to expand a patent to reach products or processes previously found in the prior art,[111] so the doctrine of prosecution history estoppel is not essential to that task. Nevertheless, prosecution history estoppel provides an additional basis for an alleged infringer to argue noninfringement—one perhaps more readily understood by judges and juries than the technologically complex issue of whether any given equivalent was a part of the prior art. Prosecution history estoppel can therefore serve to reinforce the boundaries of patentability. To the extent that narrowing amendments implicate the boundaries of patentability more directly than claim language generally, that might justify creating an estoppel in response to such amendments, but not to claim language generally.

Neither argument seems a particularly persuasive basis for treating narrowing amendments differently from general claim language. While patent attorneys should act deliberately and carefully in amending claim language, presumably they should also act deliberately and carefully in drafting the initial claim language. Moreover, patent attorneys draft the initial claims, just as much as amendments, to ensure that the application satisfies the relevant legal rules. If we need an estoppel argument as a backstop to the inherent limitations on the doctrine of equivalents in order to safeguard the boundaries of patentability in cases of amendment, it would seem that we would need such a backstop just as much for unamended claim language. Nevertheless, courts have applied prosecution history estoppel only where the claims were narrowed during the prosecution of the patent. Indeed, as the Court noted in *Warner-Jenkinson Co.*, the Court had historically applied prosecution history estoppel only in cases where a narrowing amendment was "made to avoid the prior art, or otherwise to address a specific concern—such as obviousness—that arguably would have rendered the claimed subject matter unpatentable."[112] Under this approach, prosecution history estoppel would plainly have barred Hilton Davis from using the doctrine of equivalents to claim that its patent was infringed by another's use of the pro-

[111] See, e.g., *Pall Corp v Micron Separations, Inc*, 66 F3d 1211, 1219 (Fed Cir 1995) ("[A] patent holder is estopped from recovering through equivalency that which was deemed unpatentable in view of the prior art.").

[112] Id at 30.

cess at a pH above 9.0. The "9.0" pH limitation was added expressly to overcome a prior art rejection. But under the traditional approach, because the lower pH limitation was not added to avoid the prior art, prosecution history estoppel would not so clearly apply.

On this issue, the Federal Circuit had held that prosecution history estoppel did not preclude Hilton Davis from asserting that Warner-Jenkinson's use of the process at a pH of 5.0 constituted infringement under the doctrine of equivalents. Finding no evidence in the record as to the reason for the lower pH limit, the Federal Circuit implicitly held that Warner-Jenkinson had failed to establish that the amendment adding a lower pH limitation was made to avoid the prior art. In the Federal Circuit's view, prosecution history estoppel did not therefore apply, leaving Hilton Davis free to "assert[] equivalency to processes such as Warner-Jenkinson's operating sometimes at a pH below 6."[113]

On appeal, the Court reversed. The Court held that where the reason for a narrowing amendment was unclear, a rebuttable presumption arose that the amendment was made for "a substantial reason related to patentability."[114] As a result, unless the patent holder could demonstrate that the narrowing amendment was made for some other reason, prosecution history estoppel would apply. Because there was no evidence in the record regarding the reason Hilton Davis added the lower pH limitation to its claim, the Court reversed and remanded the case, presumably to offer Hilton Davis a chance to introduce evidence regarding the reason for the lower pH limitation.[115]

Five years later, in the second recent case, *Festo Corp. v. Shoketsu Kinzoku Kogyo Kabushiki Co.*,[116] the Court again granted certiorari to review a Federal Circuit's decision on the doctrine of equivalents. In *Festo Corp.*, the Court addressed: (i) the reasons for an amendment that would not give rise to prosecution history estoppel; and (ii) the nature of the estoppel that would arise under prosecution history estoppel. As discussed above, to obtain a patent, an inventor must satisfy both the substantive requirements of sections 101, 102, and 103 with respect to the nature of the invention (novelty, nonobvious-

[113] *Hilton Davis Chem. Co*, 62 F3d at 1525.

[114] *Warner-Jenkinson Co*, 520 US at 33-34.

[115] *Warner-Jenkinson Co*, 520 US at 33. Subsequently, in *Festo Corp. v Shoketsu Kinzoku Kogyo Kabushiki Co*, the Federal Circuit limited a patent holder to the reasons for an amendment set forth in the patent's file wrapper. See *Festo Corp.*, 234 F3d 558, 586 n.6 (Fed Cir 2000), *vacated on other grounds*, 535 US 722 (2002). As a result, if a reason for an amendment unrelated to patentability did not appear in the file wrapper—and until the *Festo* decision, when a claim was amended for style or some other reason unrelated to patentability, the patent attorney had no reason to include the reason for the amendment in the file wrapper—the *Warner-Jenkinson Co* presumption would apply.

[116] 535 US 722 (2002).

ness, originality, and utility) as well as the procedural requirements set forth in section 112 with respect to the form of the patent application (enablement, best mode disclosure, definiteness). In its *Warner-Jenkinson Co.* decision, the Court had expressly stated that prosecution history estoppel would arise where a narrowing amendment was made to satisfy a substantive requirement of patent protection, but did not expressly resolve whether prosecution history estoppel would also arise for narrowing amendments made to satisfy the procedural requirements of the Patent Act.[117] Sitting en banc, the Federal Circuit ruled that prosecution history estoppel would arise if an amendment were made for any reason related to patentability, including compliance with section 112. Moreover, the Federal Circuit further held that a patent holder was bound to the reasons given in the patent's public prosecution history for the amendment and could not offer extrinsic evidence on that issue.[118] On these issues, the Court affirmed.[119] Unless an amendment was merely cosmetic, and so did not narrow the patent's scope, prosecution history estoppel would apply.[120]

On the second issue—the nature of the estoppel created by a narrowing amendment, the Federal Circuit ruled that prosecution history estoppel created a complete bar, prohibiting the application of the doctrine of equivalents entirely to those claim elements that were narrowed during prosecution in order to secure the grant of a patent.[121] Because claims are routinely amended and narrowed during patent prosecution—that is in some sense the point of the dialogue between examiner and applicant that constitutes patent prosecution—the Federal Circuit's complete bar rule would likely have estopped many, if not most, patent holders from claiming infringement under the doctrine of equivalents. As a practical matter, the complete bar would have substantially eliminated the doctrine of equivalents.

The *Festo Corp.* Court was not willing to go quite so far. Instead, the Court held that prosecution history estoppel creates a flexible bar

[117] *Compare Warner-Jenkinson Co*, 520 US at 32-33 (emphasizing that the Court had applied the doctrine in cases where the amendment was "made to avoid the prior art"), *with id.* at 33 (suggesting that patent holder cannot avoid presumption by introducing evidence that amendment was not made for "a substantial reason related to patentability").

[118] *Festo Corp*, 234 F3d at 586 n.6.

[119] *Festo Corp v Shoketsu Kinzoku Kogyo Kabushiki Co*, 535 US 722, 736-37 (2002) ("We agree with the Court of Appeals that a narrowing amendment made to satisfy any requirement of the Patent Act may give rise to an estoppel.").

[120] *Festo Corp*, 535 US at 737.

[121] *Festo Corp*, 234 F3d 558, 569 (Fed Cir 2000) ("When a claim amendment creates prosecution history estoppel with regard to a claim element, there is no range of equivalents available for the amended claim element. Application of the doctrine of equivalents to the claim element is completely barred."), *vacated*, 535 US 722 (2002).

that leaves at least some room for patent holders to claim infringement under the doctrine of equivalents even for those elements narrowed during prosecution.[122] As examples, the Court suggested that prosecution history estoppel would not preclude a patent holder from establishing infringement under the doctrine of equivalents where:

> The equivalent may have been unforeseeable at the time of the application; the rationale underlying the amendment may bear no more than a tangential relation to the equivalent in question; or there may be some other reason suggesting that the patent holder could not reasonably be expected to have described the insubstantial substitute in question.[123]

Given these examples, *Festo*'s flexible bar may not prove so flexible, particularly given the central role that the Federal Circuit will play in overseeing implementation of *Festo*'s flexible bar mandate.

Taken together, the Court's doctrinal changes, by broadening the reach of prosecution history estoppel and narrowing the doctrine of equivalents, will likely reinforce the trend towards narrower patents already evident under the Federal Circuit. Yet, *Festo*'s more enduring legacy is likely to come, not from the Court's specific doctrinal holdings, but from the rationale it offered for the doctrine of equivalents generally. Stepping away from (or perhaps merely unpacking the assumptions behind) the doctrine of equivalent's traditional basis, the *Festo* Court suggested that it is our imperfect command of language that justifies the doctrine. As the Court explained: "[T]he nature of language makes it impossible to capture the essence of a thing in a patent application."[124] Although offered in an attempt to explain and justify the doctrine's continued existence, the Court's justification may, by that logic peculiar to lawyers and three-year olds, become a limitation on the doctrine's application. If the inability to foresee perfectly all the possible variations and minor substitutions through which a competitor might duplicate the substance of an invention without infringing its literal terms justifies the doctrine of equivalents, then presumably the doctrine should not apply when the patent holder foresaw, *or reasonably could have foreseen,* the asserted equivalent at issue—an approach the Court seemed to adopt directly in cases of prosecution history estoppel.[125] This foreseeability limitation is not likely, however, to remain limited to cases involving prosecution history estoppel. Because the Court offered the "language-is-

[122] *Festo Corp,* 535 US at 737-41.

[123] Id at 740-41.

[124] Id at 731.

[125] Id. at 741 ("The patentee must show that at the time of the amendment one skilled in the art could not reasonably be expected to have drafted a claim that would have literally encompassed the alleged equivalent.").

imperfect" justification for the doctrine of equivalents generally, the Court's logic invites a similar foreseeability limit on the doctrine of equivalents generally.[126]

Under almost any plausible interpretation, the doctrine of equivalents operates primarily[127] within a relatively narrow range. In practice, it enables a patent holder to assert, and in some cases establish, infringement where the allegedly infringing device falls between: (i) what the patent holder actually claimed; and (ii) what the patent holder could legally, and with perfect information and foresight would, have claimed.[128] Although the doctrine's potential reach is therefore inherently limited, the doctrine plays a critical role in ensuring a patent's value as an exclusionary right because it covers precisely that area in which competitive, but potentially noninfringing, substitutes are likely to develop. When it created the doctrine of equivalents in *Winans v. Denmead* in 1853, the Court presumed that the patent holder intended to claim essentially all of the area between what was and what could have been claimed, absent some express language to the contrary.[129] In *Warner-Jenkinson Co.* and *Festo Corp.*, the Court has moved sharply towards reversing this presumption and has placed the burden squarely on the patent holder to justify expanding a patent beyond its claim's literal terms. Although patent attorneys will undoubtedly respond to the Court's decisions by adding additional claims and drafting existing claims more carefully,[130] there will likely remain, given the limited time and limited

[126] The Federal Circuit has already suggested such an approach. See *Sage Prods, Inc v Devon Indus, Inc*, 126 F3d 1420, 1425 (Fed Cir 1997) (holding that a patent holder cannot use the doctrine of equivalents to expand claim language to encompass foreseeable variations); see also *Johnson & Johnson Assocs Inc v R.E. Serv Co*, 285 F3d 1046, 1054-55 (Fed Cir 2002) (en banc) (per curiam) (holding that subject matter disclosed but not claimed in a patent cannot infringe under the doctrine of equivalents).

[127] The one exception is when the doctrine of equivalents extends a patent's scope to encompass after-developed equivalents. Cf. *Pennwalt Corp v Durand-Wayland, Inc*, 833 F2d 931, 938 (Fed Cir 1987) (en banc) ("The facts here do not involve later-developed computer technology which should be deemed within the scope of the claims to avoid the pirating of an invention.").

[128] The hypothetical claim approach to the doctrine of equivalents reflects this view of the doctrine of equivalents. See *Wilson Sporting Goods Co v David Geoffrey & Assoc*, 904 F2d 677 (Fed Cir 1990).

[129] *Winans v Denmead*, 56 US (15 How) 330, 343 (Dec term 1853) ("And, therefore, the patent holder, having described his invention, and shown its principles, and claimed it in that form which most perfectly embodies it, is, in contemplation of law, deemed to claim every form in which his invention may be copied, unless he manifests an intention to disclaim some of those forms.").

[130] Given that the Federal Circuit began the trend towards narrowing patent claims in the late 1980s, patent attorneys have likely already modified their patent applications accordingly. The available empirical evidence reflects an increase in the average number of claims per issued patent over the last ten years, but whether that is a conscious response to Federal Circuit's infringement rulings is not clear.

information available during a patent's drafting and prosecution, a gap between what a patent actually claims and what the patent could legally have claimed. Moreover, narrowing the doctrine of equivalents gives would-be competitors a distinct advantage. In applying for a patent, an applicant must attempt to guess how the market will develop and then draft a patent that covers *all* of the forms competition may take—a hard exercise generally made more difficult by the patent attorney's inevitable focus on the precise form of the invention before her. In contrast, the would-be competitor knows how the market and consumer preferences have developed. With the patent and its prosecution history to study at her leisure, the would-be competitor need only identify *one* means to introduce a competing product that falls outside the patent's literal scope.[131] Narrowing the doctrine of equivalents will likely therefore narrow the effective scope of any given patent.

C. Where Do We Go From Here: The Normative Desirability of Narrower, But Routinely Valid Patents

Although not conclusive on their own, both the statistics and the doctrinal revisions paint a compelling portrait of a switch from rarely valid, but broadly enforced patents to routinely valid, but narrowly enforced patents since the Federal Circuit's advent in 1982. Whether this switch is likely "to promote the Progress . . . of the Useful Arts" is a more difficult question. Under the traditional economic approach, the costs of patent protection arise because the exclusivity of patent rights is fundamentally incompatible with the nonrivalrous consumption of information. Because of this tension between private rights and public goods, establishing private rights over information entails some societal loss (given the inevitable absence of perfect price discrimination). Measured in the light of this traditional approach, the switch generates two, potentially offsetting, effects. On the one hand, effectively eliminating the nonobviousness requirement may impose some social loss by granting patents to innovations that would have been discovered and disclosed even without the inducement of a patent. On the other hand, narrowing the scope of patent protection may yield societal benefits by reducing the scope of patent exclusivity and reducing thereby the tension between a patent's private rights and the patented invention's public good character.

[131] The analogy I sometimes use with my students is that relying on the literal claims language alone to exclude would-be competitors is like playing poker where one player (the "patent holder") has to show her hand to the other (the "would-be competitor"). Although the patent holder in such a poker game may on occasion win a hand through particular skill or luck, the game is anything but fair.

Whether, on balance, the social losses within this framework will outweigh the gains is unclear.

In an attempt to understand more clearly the economic consequences of the switch, we will now examine an alternative approach that focuses on the underlying economic structure of patent law.

IV. THE ECONOMICS OF PATENT PROTECTION: THE COSTS OF UNIFORMITY

In order to limit the scope of our discussion, we begin with two assumptions. First, even in the absence of exclusive rights to the intangible information component of an innovative product, the private market, operating against a background of property rights in tangible things, will generate some incentives for innovation.[132] Nevertheless, there may remain innovations that would be desirable, in that their social value exceeds their social cost, yet unprofitable based upon the rents available from tangible property rights alone.[133] Second, although patents or copyrights are not the only mechanism available to redress this gap between desirability and profitability,[134] in some circumstances, a regime of exclusive rights in innovation will represent the best available mechanism, given the information, agency, and transaction costs otherwise present.

Even where patents or copyrights represent the best available alternative for ensuring certain types of innovation, patents and copyrights are not an ideal solution. The intangible information component of a work of authorship or an invention is characterized by nonrivalrous consumption, or as Professor Paul Samuelson phrased it, "one man's consumption does not reduce some other man's consumption."[135] Because the consumption of information is nonrivalrous, there is an inherent tension between a private ownership regime for information, and the concomitant right to exclude, and the information's public good character. The sole purpose of private ownership in this context is to permit the owner to exclude, or to threaten to ex-

[132] I will present formal models of the rents available from a lead-time advantage or from a reputation for innovation in Part V. See text accompanying notes 167-171.

[133] Although this gap between desirability and profitability is assumed for now, the model presented in Part V will formally establish this gap. See text accompanying notes 167-171.

[134] Some of the alternatives include direct government financing through research and development contracts, government buy-outs of patents, and prizes for innovation. For a discussion of the relative advantages and disadvantages of these approaches, see, for example, Wright, 73 Am Econ Rev 691 (cited in note 9); Kremer, 113 Quarterly J Econ 1137 (cited in note 9); Scotchmer & Green, 21 Rand J Econ 131 (cited in note 9).

[135] Paul Samuelson, *The Pure Theory of Public Expenditure*, 36 Rev Econ Stat 387, 387 (1954) (labeling such goods "collective consumption goods").

clude, non-payors. The threat of exclusion allows the owner to separate high and low reservation value consumers and to charge each consumer some price for access to the information component of her product. Inevitably, however, imperfect information and imperfect contract enforceability will lead an owner to exclude some consumers from access.[136] Yet because consumption of information is non-rivalrous, exclusion is never Pareto optimal.[137] As Professor Kenneth Arrow has explained: "In a free enterprise economy, inventive activity is supported by using the invention to create property rights; precisely to the extent that it is successful, there is an underutilization of the information."[138]

There is therefore a tension between private ownership and the

[136] If individuals had perfect information regarding the preference structures of others and contracts were perfectly and costlessly enforceable, a significant part of the justification for patent and copyright protection would disappear. With perfect information and perfectly enforceable contracts, the creator of an innovative product could use the threat that she will withhold access until the consumer agreed to enter into a contract that would adequately safeguard the information component of her product. The only gap left would be the one created by the risk of simultaneous, independent invention. Even that risk could be addressed by relaxing the antitrust laws to permit such simultaneous inventors to enter into binding contracts, dividing profits from their joint market.

[137] Professor Wendy Gordon has suggested that where a market price has been set for a public good that effectively excludes consumers unwilling to pay the market price, that is the ordinary operation of the market and hence not a market failure. See Wendy J. Gordon, *Fair Use as Market Failure: A Structural and Economic Analysis of the Betamax Case and Its Predecessors*, 82 Colum L Rev 1600, 1611-15 (1982). Such an assertion fails to recognize the key difference between public and private goods. If we were dealing with a private good, i.e. one characterized by rivalrous consumption, such market price exclusion can be Pareto optimal. For example, if all markets are complete and perfectly competitive, and the price of an apple was $1, there may be any number of consumers who would like an apple, but who are unable to afford the market price. However, if (as seems likely) apples are characterized by rivalrous consumption, in order to provide an apple to one of these excluded consumers, we would have to take an apple from one of the consumers willing to pay the market price. Because the consumer from whom we take the redistributed apple would be worse off (under the usual assumptions of self-interested utility), such a reallocation would not prove Pareto optimal. (Moreover, because the consumer from whom we take the apple apparently values it more highly than the otherwise excluded consumer to whom we give the apple, the amount that the otherwise excluded consumer would pay for the apple is insufficient to compensate the consumer from whom we take the apple for her loss.) The market price exclusion associated with private goods can therefore prove Pareto optimal. However, when we are dealing with public goods, the lack of rivalry means that we can provide access to the public good to one of the consumers otherwise excluded by the market price of the good without taking that good from anyone else. As a result, market price exclusion is not Pareto optimal for public goods.

[138] Kenneth J. Arrow, *Economic Welfare and the Allocation of Resources for Invention*, in Bureau of National Economic Research, *The Rate and Direction of Inventive Activity* 609, 617 (Princeton U, 1962).

public good character of information. This tension may lead to welfare losses where the market price for the public good excludes some consumers from access. It may also lead to welfare losses where information asymmetries, transaction costs, the limits of practical contract enforceability, or the potentially bilateral monopoly character of the negotiations preclude an innovator from entering into an otherwise desirable contract with another who wishes to build upon or reuse the information component of a patented or copyrighted product. This tension usually leads economists to suggest that determining the optimal scope of patent or copyright requires balancing the benefits derived from the *ex ante* incentive effects of protection against the costs that arise *ex post* from exclusion.[139] For example, Landes and Posner assert that: "For copyright law to promote economic efficiency, its principal legal doctrines must, at least approximately, maximize the benefits from creating additional works minus both the losses from limiting access and the costs of administering copyright protection."[140] Others, following Arrow's formulation, characterize the issue in terms of a trade-off between dynamic and static efficiency.

Because of the tension between public goods and private rights, even where patent protection represents the best available mechanism for encouraging certain types of innovation, we should presumably provide protection only if, and to the extent, necessary to ensure each innovation's expected profitability. Yet, historically, both patent and copyright have not attempted to tailor protection to such an individually optimal level. Instead, both patent and copyright have provided more-or-less uniform protection to the full range of innovative products that satisfy their more-or-less uniform sets of prerequisites. Uniformity means that each innovative product that satisfies the prerequisites receives roughly the same protection.[141] It also means that

[139] See Merges & Nelson, 90 Colum L Rev at 868 (cited in note 7) ("In most analyses of the different aspects of the patent system, concern has centered on a simple tradeoff. The analysis has concentrated on how changing patent coverage affects the balance between incentives to the inventor and underuse of the invention due to patent monopolies.").

[140] William M. Landes & Richard A. Posner, *An Economic Analysis of Copyright Law*, 18 J Leg Stud 325, 326 (1989). Later in their article, Landes and Posner argue that the "various doctrines of copyright law . . . can be understood as attempts to promote economic efficiency by balancing the effect of greater copyright protection—in encouraging the creation of new works by reducing copying—against the effect of less protection—in encouraging the creation of new works by reducing the cost of creating them." Id at 333.

[141] Within patent and copyright, courts often use the discretion inherent in statutory language to provide somewhat more protection for certain types of creative products. For example, copyright protection for factual works, such as encyclopedias, or

if we expand patent or copyright protection, we expand it uniformly for all protected products. Because of this uniformity, when we expand patent or copyright protection, we expand it both to the innovative products newly created in response to the expansion *and* to those innovative products that would have been created even without the expansion ("preexisting products").[142] In terms of social welfare, the additional products created in response to the expansion represent a simple gain (at least, to the extent that a regime of exclusive rights represents the best available alternative for ensuring the additional innovation). Admittedly, given the tension between private property and public goods, these additional products will be somewhat less valuable to society than they would have been if they could have been produced with no or less protection.[143] However, so long as these additional products would not have been forthcoming but for the expansion in protection, society is still better off with these additional products than without them (given the assumption that all other markets are complete and perfectly competitive). On the other hand, expanding protection to the preexisting products represents a pure welfare loss. Because the creation of these products could have been ensured with no or less protection, we cannot properly attribute any

---

useful works, such as computer programs, is generally less extensive than copyright protection for fictional or entertaining works. In copyright law, this variation generates the perverse result of providing the least protection and hence the least incentive for the works we most need, precisely because we most need them. See, e.g., Glynn S. Lunney, Jr., *Reexamining Copyright's Incentives-Access Paradigm*, 49 Vand L Rev 483 (1996). By focusing exclusively on the *ex ante* costs of limiting access to a work, such an approach is unlikely to ensure that copyright enables each author precisely to recover her reservation cost. For the doctrinal elements courts use to incorporate variability into patent law, see text accompanying notes 175-184.

[142] Please note the precise definition of "preexisting." Whether a product is preexisting for our purposes here is not a question of the timing of the innovation relative to the timing of the expansion in patent or copyright protection. Rather, it refers to an innovative product that would have been created even in the absence of the expansion, whether the expansion occurs before or after the creation of the innovative product at issue. By defining "preexisting" in this manner, I intend to leave open the possibility, alluded to by Ted Olson during oral argument in *Eldred v Ashcroft*, that an individual devotes her resources to the creation of an innovative product in the expectation of a future expansion in patent or copyright. While I do not find this possibility factually plausible for the Copyright Term Extension Act of 1998, I am using a definition of preexisting that assumes that we can identify without error which works would have been created without the expansion (or the expectation thereof) and which works would not have been.

[143] As Arrow has explained: "It is necessary to distinguish between the realized social benefit and the potential social benefit, which in this case, means the sale of the product at postinvention cost, c'. Clearly, the social benefit always exceeds the realized social benefit." Arrow, *Economic Welfare and the Allocation of Resources for Invention* at 622 (cited in note 138).

part of the value of these products to the expansion. Moreover, expanding protection for the information component of the preexisting products will limit the ability of others to take advantage of the lack of rivalry that characterizes their consumption and will thereby decrease directly their social value. The key to evaluating the desirability of expanded protection thus becomes a balancing of the value gained from the additional creative output that broader protection may ensure against the value lost from the reduced ability to exploit the nonrivalrous character of the preexisting products.

We can model the costs that arise from uniform protection formally. We begin by defining the total surplus for innovative product i as the sum of consumer and producer surplus for a particular level of patent or copyright protection, λ:

$$S_i(\lambda) = S_i^c(\lambda) + S_i^p(\lambda)$$

An individual will devote her resources to the creation of innovative product i so long as the expected producer surplus associated with the product is greater than or equal to the individual's reservation price, which is assumed to be zero. Following the traditional economic analysis of private rights and public goods, we will assume that an increase in λ will (over the range of potentially optimal λ)[144] increase the producer surplus, and decrease the consumer surplus, associated with each product i. Thus, $\partial(S_i^c(\lambda))/\partial\lambda < 0$ and $\partial(S_i^p(\lambda))/\partial\lambda > 0$. Moreover, in the absence of perfect price discrimination, there will be some slippage in the patent or copyright owner's ability to use the increased protection to convert consumer surplus into producer surplus. For that reason, the patent or copyright owner will be unable to capture completely the consumer surplus lost as a result of the increased protection. Increased protection will therefore lead to a loss in the total surplus associated with any given innovative product, thus $\partial(S_i(\lambda))/\partial\lambda < 0$. We will finally assume N innovative products are available, ranked in decreasing producer surplus order (for any given λ)

[144] As Professors Landes and Posner have explained:

> N_z measures the response of the number of works created to an increase in copyright protection. As we saw earlier, it can be either positive or negative. However, when z [the level of copyright protection] is set optimally, N_z will be positive. For suppose that N_z were negative at z^*. Since the same level of N could be attained at a lower z (because N increases initially and then falls as z rises), a lower z would yield a higher level of W [social welfare]. Not only would [the total cost of creating works] $E(N, z)$ be lower (since it is a positive function just of z when N is unchanged, and z would now be lower, but w (consumer and producer surplus per work before deducting the cost of expression) would be higher at a lower z for reasons explained in the previous section.

Landes & Posner, 18 J Leg Stud at 342 (cited in note 140).

from 1 to N. Given these assumptions, we can now model: (i) the optimal scheme of individually tailored patent protection; (ii) the optimal scheme of uniform patent protection; and (iii) the regulator's endogenous choice between these two regimes and the degree of individual tailoring to include in the patent system.

Optimal Individually Tailored Protection. With perfect information and costlessly enforceable legal rules, we could individually tailor patent and copyright to each individual innovative product.[145] With such individually tailoring, the regulator would set λ_i to maximize the social welfare from each innovative product. For each innovative product, the regulator would therefore solve:

$$\underset{\lambda}{\text{Max}}\ S_i^t(\lambda_i) = S_i^c(\lambda_i) + S_i^p(\lambda_i) \tag{1}$$

such that $S_i^p \geq 0$ (the innovator's individual rationality constraint).

To minimize the welfare losses that arise from the tension between private property and public goods, the regulator will set the optimal level of individual protection, λ_i^*, such that $S_i^p = 0$. Substituting $S_i^p = 0$ into (1), the regulator will set λ_i^* to maximize both total and consumer surplus (which in this case are the same) for each innovative product. By necessity given the regulator's objective function, $S_i^t(\lambda_i^*) \geq S_i^t(\lambda_u^*)$ with equality only for those products for which $\lambda_i^* = \lambda_u^*$, i.e., for the last product n produced under the uniform scheme of protection. In addition, with individualized protection, the regulator will set λ_i^* to ensure the existence of each innovative product for which $S_i^t(\lambda_i^*) \geq 0$. In other words, if the innovative product represents the most valuable use of society's resources, then the regulator will set λ_i^* to ensure its existence. As a result, so long as $S_N^t(\lambda_N^*) \geq 0$, the regulator will set individualized protection levels λ_i^* to ensure the existence of the full range of innovative products, from 1 to N.

With individualized protection, social welfare thus becomes:

$$S^t(\lambda_i^*) = \sum_1^N S_i(\lambda_i^*) \tag{2}$$

Optimal Uniform Protection In contrast, with uniform protection, λ, applied to all innovative products, there will be some product n for which the expected producer surplus exactly equals zero. For a given

[145] Because some research paths prove ultimately unsuccessful, the rents available from successful research paths must cover not only the costs of the successful path, but also a share of the costs of the unsuccessful paths. A full consideration of this issue is beyond the scope of this article. Instead, I have simply assumed that the government actor has accounted for the risk involved in research in selecting the optimal level of individualized protection.

λ, products 1 to n will generate a nonnegative expected producer surplus and will be created. For a given λ, products $n + 1$ to N will generate a negative expected producer surplus and will not be created.

Given this framework, social welfare at some initial level of protection, λ, becomes the sum of the consumer and producer surplus for the resulting innovative products, 1 to n:

$$S^t(\lambda) = \sum_{1}^{n} S_i(\lambda) \tag{3}$$

Increasing protection from this initial level will reduce the total surplus, but increase the producer surplus, associated with each innovative product. By ensuring a nonnegative expected producer surplus for additional innovative products, increased protection will lead to the production of additional products. If we consider the smallest increase in protection, from λ to λ', that shifts the break-even point from product n to product $n + 1$, then at this new level of protection, λ', social welfare becomes:

$$S^t(\lambda') = \sum_{1}^{n+1} S_i(\lambda') \tag{4}$$

In order to determine the marginal effect of increased protection, we subtract (3) from (4). Simplifying, rearranging terms, and solving for the optimal level of uniform protection by setting $\Delta S^t = 0$, we obtain:

$$S_{n+1}(\lambda') = -\sum_{1}^{n} [S_i(\lambda') - S_i(\lambda)] \tag{5}$$

With uniform protection, the optimal level of uniform protection, λ_u^*, is thus defined implicitly by (5) as the level of protection at which: (i) the social welfare gained from the additional creative products broader protection ensures; exactly equals (ii) the social welfare lost from protecting preexisting creative products more broadly than necessary.

We can therefore define the costs of uniformity, C_u, as the difference between social welfare with the optimal scheme of individualized protection and social welfare with the optimal scheme of uniform protection:

$$C_u = \sum_{1}^{N} S_i(\lambda_1^*) - \sum_{1}^{n} S_i(\lambda_u^*) \tag{6}$$

We can rearrange (6) in order to separate the costs of uniformity into two components:

$$C_u = \sum_{n+1}^{N} S_i(\lambda_i^*) + \sum_{1}^{n} [\Delta S_i(\lambda_i^*, \lambda_u^*)] \tag{7}$$

As (7) reveals, a uniform scheme of patent or copyright protection imposes two types of costs. First, because uniformity requires the regulator to consider the trade-off between encouraging additional innovative products and maximizing the social welfare associated with the preexisting products, the regulator will set λ_u^* at a level too low to ensure the expected profitability and hence the existence of innovative goods, $n+1$ to N. Even if goods $n+1$ to N represent the most valuable use of the resources at issue, expanding a system of uniform protection to ensure these additional innovative goods' existence is too costly given the restrictions such an expansion would place on exploitation of the preexisting innovative goods. With uniformity, society therefore loses the surplus associated with these additional innovative goods, as reflected in the first summation on the right-hand side of (7). Second, a uniform protection scheme also over-protects innovations 1 to $n-1$, enabling the patent or copyright owner to capture producer surplus strictly in excess of her reservation price. Again, in the absence of perfect price discrimination, total surplus for a given innovation, subject to the individual rationality constraint, is maximized only when $S_i^p = 0$. The total surplus associated with each of the innovations 1 to $n-1$ will therefore be lower under uniform protection than under individualized protection. With uniformity, society therefore also loses some part of the surplus associated with innovations 1 to $n-1$,[146] as reflected in the second summation on the right-hand side of (7).

The Endogenous Choice Between Uniform and Individually Tailored Protection. Given the costs of uniformity, the question arises why we continue to provide more-or-less uniform patent and copyright protection. The answer likely lies in the information and administrative costs that a system of individualized protection would entail. In order to tailor protection to ensure that an innovator received patent rights only if, and precisely to the extent, necessary to ensure that the innovator received her reservation cost for the innovation would require considerable information. If we tried to obtain the necessary information directly from the would-be patent holder, a moral hazard would arise.[147] The would-be patent holder would be tempted to overstate both the social value of her innovation and her

[146] Because $\lambda_i^* = \lambda_u^*$ for innovation n, $\Delta S_n(\lambda_n^*, \lambda_u^*) = 0$. It is therefore irrelevant whether the second summation on the right-hand side of (7) runs from 1 to $n-1$ or from 1 to n.

[147] The moral hazard parallels the tendency to underestimate costs for government defense projects in order to secure the contract to supply the project. See, e.g., A. W. Marshall & W. H. Meckling, *Predictability of the Costs, Time, and Success of Development*, in National Bureau of Economic Research, *The Rate and Direction of Inventive Activity* 461, 470-75 (Princeton U, 1962).

reservation price in order to secure a broad right of exclusion and the associated rents. Moreover, because broader patent rights likely correlate with increased rents, a menu of patent protection will likely fail to separate effectively high- and low-reservation cost innovators.

Existing attempts to tailor patent protection to particular technologies tend therefore to reflect efforts to achieve uniformity, rather than variation, in patent protection. For example, section 156 of the Patent Act varies patent law's otherwise uniform twenty-year duration for a prescribed set of patents. Although section 156 might appear initially to detract from patent law's uniformity, section 156 limits the availability of these extensions to those products that require premarketing regulatory approval.[148] Where an extension is available, section 156 further ties the length of the patent term extension to the length of the regulatory review process.[149] By allowing inventors to "recoup" the patent time during which regulatory approval was pending, and hence no commercial exploitation of the invention was possible, section 156 attempts to ensure that new drugs (and other products subject to FDA review) receive the same effective patent term as products not subject to a similar regulatory approval process.[150]

Moreover, even if we could overcome the information and administrative costs, a system of individualized protection would also entail potentially substantial agency costs. The early history of patent abuses by the English crown has cast a long shadow over patent law. As the Court noted in *Graham v. John Deere Co.*,[151] the Patent and Copyright Clause of the United States Constitution "was written against the backdrop of the practices—eventually curtailed by the Statute of Monopolies—of the Crown in granting monopolies to court favorites in goods or businesses which had long before been enjoyed by the public."[152] Although the English abuses are hundreds of years in the past, they remain a stark reminder of the excessive agency costs that a system of individualized protection may entail. Even if the sort of readily visible excesses of the English crown are unlikely in an open democracy, the uniformity of the patent and copyright systems likely remains an important bulwark against more

[148] 35 USC §§ 156(a)(4), (f) (2002).

[149] 35 USC § 156(c) (2002).

[150] As discussed, an optimal system of individually tailored intellectual property rights would leave each innovator with surplus exactly equal to her reservation price. We should therefore expect innovators to oppose strenuously any attempt to institute such an optimal system. On the other hand, innovators will happily support variations to the existing uniform scheme that expand protections for innovations in their field, whether or not such expansion is justified in terms of increased social welfare.

[151] 383 US 1 (1966).

[152] *Graham*, 383 US at 5.

subtle agency-cost driven biases, particularly where Congress or an administrative agency, rather than courts,[153] would be responsible for individually tailoring protection.[154]

Given these information and agency costs, a regulator seeking to maximize social welfare has four alternatives. First, the regulator can refuse to provide a system of exclusive rights for the innovative product at issue. With such selection, welfare becomes:

$$W = S_p(n_m) \tag{8}$$

where W equals social welfare, S_p is the social value of the innovative products at issue, and n_m is the number of innovative products that the market, operating against a background of private rights in tangible things, will ensure.

Second, the regulator could provide a uniform system of exclusive rights for the innovative products at issue and make no effort to exclude those innovative products that would be devised and disclosed in the absence of such protection. Here, the regulator solves (6) to identify λ_u^*. With such a uniform scheme of protection, social welfare becomes:

$$W_u = S_p(n(\lambda_u^*), \Delta\lambda_u) - E_u \tag{9}$$

where $\Delta\lambda_u$ is the extent to which innovative products, 1 to $n-1$, are overprotected and E_u is the information, agency, and transaction costs such a uniform scheme of protection entails.

Third, the regulator could provide a uniform system of exclusive rights for the innovative products at issue, but attempt to exclude from protection entirely those innovative products that would have been devised and disclosed in the absence of such protection. Here,

[153] Because federal judges are appointed for life, two of the principal means by which bribes and other illegitimate forms of influence can be concealed, campaign contributions and promises of future employment, are not as readily available for influencing federal judges. In addition, because the work of judges inevitably entails a written record of decision open to the public, the shadows in which attempts at improper influence can flourish are not as deep with judicial decision-making. As a result, agency costs likely are lower when judges, rather than other government agents, are responsible for individually tailoring patent rights.

[154] Even with uniform protection, these agency-cost biases can still influence the shape of protection. Thus, copyright producers, because of their transaction cost and collective action advantages over copyright consumers, have successfully persuaded Congress to expand both the duration and scope of copyright protection far beyond what can be rationally justified. However, if we took computer programs, factual works, and other useful works of authorship out of the system and isolated entertaining and fictional works within their own scheme of protection, there is every reason to believe that the protection accorded entertaining and fictional works would be even broader than it is today.

the regulator again solves (6) to identify λ^*_{ue}, but by limiting the class of preexisting products to which protection applies, the optimal level of uniform protection will be higher. Thus, $\lambda^*_{ue} > \lambda^*_u$ *and* $n(\lambda^*_{ue}) > n(\lambda^*_u)$. Moreover, if we assume that innovative products 1 to k would be devised and disclosed even without exclusive rights in their intangible information components, then such a scheme of protection will not protect, and hence will not overprotect, goods 1 to k. However, attempting to identify and exclude the preexisting innovative products from protection will entail increased information, agency, and transaction costs; thus, $E_{ue} > E_u$. With such an exclusionary, but otherwise uniform scheme of protection, social welfare becomes:

$$W_{ue} = S_p(n(\lambda^*_{ue}), \Delta\lambda_{ue}) - E_{ue} \tag{10}$$

Fourth and finally, the regulator could attempt to tailor protection to the individually optimal level for each innovative product eligible for protection. If we assume that the regulator can select any expenditure or cost level, e, in attempting to identify the precise reservation cost for each innovation and then to define and enforce a system of individualized system of patent protection, we can generalize the regulator's problem as follows:

$$\underset{e}{\text{Max}}\ W_i = S_p(n(\lambda_i), \Delta\lambda_i) - e \tag{11}$$

where W_i equals social welfare with such individually tailored protection, S_p is the social value of all patentable innovations, given the number of innovations created, n, and the difference for each innovation between the actual level of patent protection received and the individually optimal protection level, $\Delta\lambda_i$. As the regulator increases e, the regulator has better information and is better able to tailor protection for each innovation to its individual optimal. As e increases, $\Delta\lambda_i$ falls (because protection is more closely tailored to the individually optimal) and n increases (because protection more closely tailored to the individual optimal allows the regulator to set a higher maximum level of protection).

Taking the first order condition for (11), the regulator therefore chooses to expend effort e^* to solve:

$$\frac{dW_i}{de} = 0 \Rightarrow \frac{dS_p}{dn}\frac{dn}{de} + \frac{dS_p}{d(\Delta\lambda_i)}\frac{d(\Delta\lambda_i)}{de} = 1 \tag{12}$$

In other words, the regulator expends effort e^* to the point where the additional information, agency, and transaction costs entailed in more precisely tailoring the scope of protection to the individually optimal levels exactly equals the marginal social benefit of that additional effort in terms of: (1) the social value of additional innova-

tions better tailored protection ensures; and (2) the increased social value of preexisting innovations from more closely tailoring their protection to their respective individually optimal levels.

After determining the socially optimal effort entailed in individually tailoring rights, the regulator implements the resulting scheme of individually tailored rights and social welfare becomes:

$$W_i^* = S_p(n(\lambda_i^*), \Delta\lambda_i) - e^* \qquad (13)$$

Having identified the welfare consequences of each of the four alternative schemes of protection available, the question for the regulator becomes which scheme maximizes social welfare for the particular class of innovative products at issue.

Consequences of Uniformity. Before using this analysis to evaluate the switch to routinely valid, but narrowly enforced patents, the costs of uniformity identifies several principles that, from an efficiency perspective, should guide the structure of patent and copyright if we assume that the information and agency costs entailed in individualized protection are prohibitive.[155] First, as discussed above, if patent and copyright provide more or less uniform protection, patent and copyright should not provide protection sufficient to ensure the expected profitability of the full range of innovative products eligible for protection. Even where an innovative product represents the most valuable use of available resources, we must balance the value of an additional innovative product against the loss that arises from broader protection of the preexisting products. For that reason, an optimal uniform scheme of protection will provide protection that will

[155] We will revisit this assumption in Part VI. See text accompanying notes 172-183. One alternative that may deserve further exploration would be an auction system for patents. An auction system might provide an alternative to a system of government-tailored individualized intellectual property rights. Specifically, the government could "auction" a patent on a particular innovation to a party that agrees to develop the innovation within a particular time period. Would-be inventors could then bid for the patent by stating the minimum patent term that each would accept in order to develop the specified innovation. For example, in analyzing whether second-generation innovations should receive their own patent, Suzanne Scotchmer has argued that *ex ante* contracting between the holder of a patent on the first generation of a product and the would-be developers of the second-generation innovations should usually provide an appropriate incentive for the second-generation product. See Suzanne Scotchmer, *Protecting Early Innovators: Should Second-Generation Products be Patentable?*, 27 RAND J Econ 322 (1996). The difficulty with such an approach, both for second-generation licensing and for an auction scheme more generally, lies in the uncertainty associated with the inventive process. See, e.g., Marshall & Meckling, *Predictability of the Costs, Time, and Success of Development* (cited in note 147). If we cannot predict *ex ante* the timing or costs of innovation, or even its nature, then both *ex ante* contracting and an auction scheme are unlikely to function effectively.

leave some desirable innovative products unprofitable.[156] This insight may help explain the traditional exclusion from either patent or copyright protection of a variety of innovative products, including such disparate products as business methods and clothing designs. Even if providing patent or copyright protection would encourage the production of additional desirable innovations in these fields, excluding these fields from protection remains desirable so long as: (i) the information and other costs entailed in identifying and excluding the preexisting innovative products in the field from protection would prove prohibitive; and (ii) the resulting social cost from extending protection to the preexisting products would exceed the social value of the additional innovations that protection would generate.

Second, although we must be aware of potential information and agency costs, we should remain alert for opportunities to limit uniformity costs by incorporating variation into an otherwise uniform system of protection.[157] For example, if Congress were to expand protection by extending the duration of copyright or patent, Congress could minimize the costs such a uniform expansion would impose by limiting the term extension to those innovative products that would not have been devised or disclosed but for the additional term. However, Congress is unlikely to have sufficient information to distinguish these innovative products from those that would have been created even without the expansion. Nevertheless, even in the absence of the information required to achieve such precise tailoring, Congress could extend the term, but require the rights holder to satisfy certain periodic formalities. The Patent Act embraces such an approach, providing a uniform term of twenty years from the date the patent application was filed, but requiring the payment of maintenance fees at three points to maintain the patent in force for the full term.[158] While such an approach does not strictly limit the duration of protection to that necessary to ensure a given innovation (and is not therefore an incentive compatible mechanism), at the very least,

[156] This provides another reason why we should accord little weight to social rate of return studies. By identifying innovative products just outside the margins of existing protection with positive social rates of return, they do not establish a basis for expanding patent or copyright protection. Rather, they are simply identifying an inevitable consequence of an optimal scheme of uniform protection.

[157] For an exploration of how the nonobviousness doctrine, claims language, and the doctrine of equivalents can introduce variation into otherwise uniform patent protection, see text accompanying notes 175-183.

[158] Sections 41(b) and 154(a)(2) of the Patent Act incorporate such a requirement. In order to receive the full twenty-year term, a patent holder must pay maintenance fees on or before (or within a six month grace period) 3 years and 6 months after grant of the patent, 7 years and 6 months after grant, and 11 years and 6 months after grant. 35 USC §§ 41(b), 154(a)(2) (2003).

it ensures that protection will end once the expected rents from an additional term of protection for a particular innovation exceed the costs of applying for the extension. For a given patent term,[159] such an approach should tend to reduce (somewhat)[160] the costs of an otherwise uniform patent term, without undermining significantly the incentive to innovate.[161]

Third, and most importantly, to minimize the costs of uniformity, we should limit application of a uniform system of IPRs to "similar" innovative products. Innovative products are "similar" to the extent that: (1) a given set of uniform prerequisites defines when a significant gap will arise between the desirability of an innovative product (relative to alternate uses of the resources) and its expected profitability; and (2) a given set of uniform exclusive rights approximates the protection precisely necessary to close that gap. As (7) reflects, the costs of uniformity increase to the extent that the optimal level of uniform protection, λ_u^*, differs from the optimal level of individual protection, λ_i^*, for each of the innovations, $i = 1, N$, eligible for protection under a particular system of intellectual property rights ("IPRs"). On the other hand, the costs of uniformity tend to decrease as λ_u^* approaches λ_i^* for each of the eligible innovative goods. In the exceptional case, where $\lambda_i^* = \lambda_u^*$ for all of the innovations eligible for protection, the costs of uniformity would be zero.

While both patent and copyright cover a sufficiently wide range of innovative products that λ_u^* is unlikely to equal λ_i^* for the full range of innovative products eligible for protection, patent and copy-

[159] This assumption is important. If the use of such formalities becomes a justification for longer terms of protection, the question becomes more complex. The formalities would still entail some cost savings by removing those products for which the rights holder sees little or no commercial value from intellectual property protection. However, some innovative products that would have been created even with the shorter term may nonetheless remain commercially value for the longer term. As a result, the rights holder would comply with the formalities to preserve protection for such an innovative product, tying down access to the product's information component for a longer time than necessary to ensure the innovative product's existence. In short, because formalities leave the decision to extend protection to the rights holder, they are not well-tailored to minimizing the social costs of a longer term of protection. While such formalities are desirable for a given term of protection, they should not become a justification for a longer term of protection.

[160] Because patent holders will allow only a patent to expire only where the costs of renewal exceed the expected value of the patent rents, we should expect failures to renew only for the least valuable patents. To the extent that a patented innovation has such trivial economic value, the social costs of extending protection are also likely to prove trivial. (Recall that we have postulated the social costs of protection are some fraction of the unrestricted social value.)

[161] See Francesca Cornelli & Mark Schankerman, *Patent Renewals and R&D Incentives*, 30 RAND J Econ 197 (1999); Suzanne Scotchmer, *On the Optimality of the Patent Renewal System*, 30 RAND J Econ 181 (1999).

right historically were each tailored to redress a specific desirability-profitability gap. For example, copyright, at least for its first hundred years, focused on the desirability-profitability gap that would otherwise arise for innovative products subject to rapid and inexpensive mechanical copying. By prohibiting mechanical or near-mechanical duplication, copyright sought to narrow the desirability-profitability gap for such works. At the same time, however, copyright limited the costs that would arise from such protection in two ways. First, copyright strictly limited eligible subject matter to the types of innovative products particularly susceptible to such copying. The Copyright Act of 1790, for example, limited copyright protection to "maps, charts, and books"—the types of innovative products particularly susceptible to rapid and inexpensive mechanical duplication given the technology available (i.e. the printing press) at that time.[162] Second, nineteenth century copyright also strictly limited the scope of protection available, granting an author the exclusive right "to multiply copies," but leaving others free to build upon, translate, and otherwise reuse the work.[163] By narrowly defining the innovative products eligible for copyright protection and by limiting the scope of protection provided, copyright historically tended to remedy the desirability-profitability gap mechanical duplication would otherwise create without imposing undue costs. Unfortunately, over the course of the twentieth century, that sensible focus, and copyright's consequential efficiency in encouraging the creation of additional innovative products without unduly restricting the use of preexisting products, has been lost.

From the outset, patent law has taken a slightly different approach. Unlike copyright law, patent law defines the subject matter eligible for its protection broadly, encompassing any "machine, manufacture, process, or composition of matter."[164] Patent also defines its scope of protection broadly, prohibiting another from "making, using, selling, or offering to sell" the patented invention. Unlike copyright law, patent law not only prohibits another's copying, it also prohibits another who has developed the same innovative product entirely independently from making, using, or selling it.[165] Rather than rely on a narrow

[162] See Act of May 31, 1790, ch. 15, § 1, 1 Stat. 124.

[163] See, e.g., *Perris v Hexamer*, 99 US 674, 675-76 (1878) (defining copyright as "the exclusive right of multiplying copies of what he has written or printed"); accord *Greene v Bishop*, 10 F Cas 1128, 1133-34 (D Mass 1858) (No 5,762).

[164] 35 USC § 101 (2003).

[165] Although copyright today has become far longer in duration than patent, patent and copyright initially had similar durations. Compare Act of Apr. 10, 1790, ch.7, § 1, 1 Stat 109 (authorizing patents with term of fourteen years), with Act of May 31, 1790, ch 15, § 1, 1 Stat 124, 124 (authorizing copyrights with initial term of fourteen years followed by a renewal term of an additional fourteen years if the author was still living and satisfied the renewal formalities).

subject matter or a narrow scope of protection to limit its uniformity costs,[166] patent law has instead relied on a strict nonobviousness requirement. Where a copyright requires only originality—essentially a requirement that the innovative product be the author's own work—a patent requires nonobviousness. To satisfy this requirement, a patent applicant has to demonstrate not only that the invention at issue was his own, but also represented a significant advance over the work of others. Unlike copyright law, patent law thus attempts to limit the costs its protection would otherwise impose by limiting patents to those innovations that reflect a substantial technical advance.

Figure 4 summarizes the respective combinations of eligible subject matter, level of creativity required for protection, and scope of protection historically associated with copyright and patent protection.

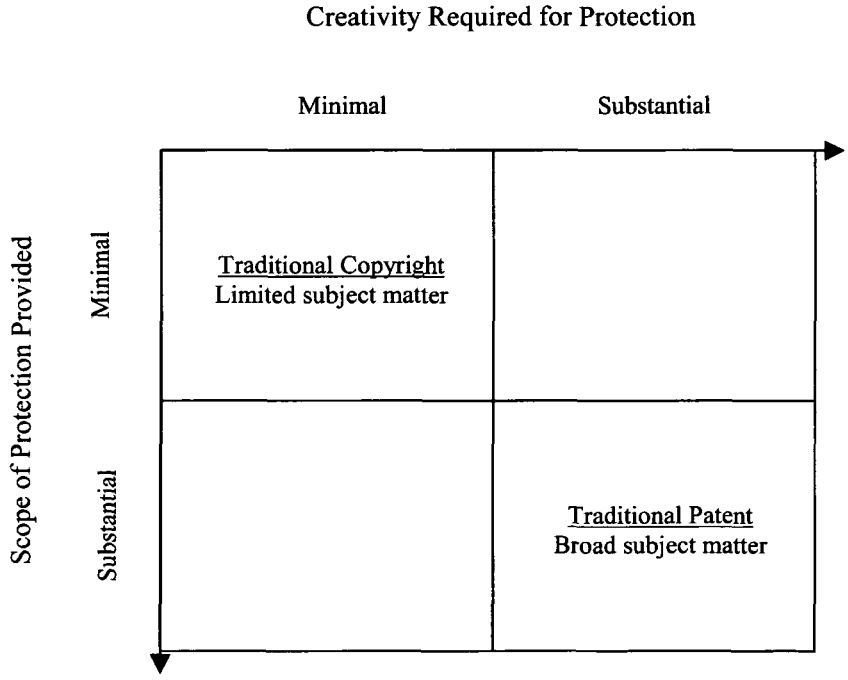

Figure 4. Traditional Balancing of Creativity Required and Scope of Protected Subject Matter Against Scope of Protection Provided

[166] See, e.g., *Diamond v Chakrabarty*, 447 US 303, 309 (1980). ("The Committee Reports accompanying the 1952 Act inform us that Congress intended statutory subject matter to 'include anything under the sun that is made by man.' S. Rep No 1979, 82d Cong, 2d Sess 5 (1952); HR Rep No 1923, 82d Cong, 2d Sess 6 (1952).").

Through its modifications to patent law, the Federal Circuit, with sometimes the acquiescence of the Court and sometimes its support, seems determined to push patent protection from the lower right quadrant towards the upper left quadrant—the traditional purview of copyright law. Although both the traditional and Federal Circuit approaches implicitly recognize the costs of uniformity, they do so from diametrically opposed perspectives. Because the traditional approach provides broad protection, limiting the costs of such broad protection (uniformly provided) requires strict limits on the availability of such protection—hence an interpretation of the nonobviousness doctrine that limits protection to the most creative technical advances. In contrast, the Federal Circuit's approach offers protection even to minor technical advances. Because protection is widely available, limiting the costs of such widely available protection (uniformly provided) requires strictly limiting the scope or extent of protection provided—hence the narrow interpretation of the doctrine of equivalents. Focusing on the costs of uniformity, within the structure of patent law, thus suggest that we should not evaluate the changes to the nonobviousness doctrine and the doctrine of equivalents independently, but as a package. Changing the ease with which protection may be obtained dictates a corresponding need to change the scope of protection and *vice versa*. Because it recognizes this trade-off, a switch to routinely valid, but narrowly enforced patents is almost certainly preferable to a switch to routinely valid patents, broadly enforced or to rarely valid patents, narrowly enforced.

As between the upper left quadrant and the lower right quadrant, the question becomes which quadrant better corresponds to the desirability-profitability gap likely to arise for innovations eligible for patent protection. Through the nature of protection provided, each approach suggests implicitly the desirability-profitability gap that it is attempting to address. The traditional approach suggests that there is a desirability-profitability gap for unusually creative products, which gap requires a fairly broad scope of protection to redress. The Federal Circuit's approach suggests that there is a desirability-profitability gap for routinely creative products, which gap requires a relatively narrow scope of protection to redress. In order to explore which of these approaches has correctly identified the relevant gap, the next section presents formal models of the rents available to encourage innovation in the absence of patent protection. Depending upon where these models suggest the relevant gap between desirability and expected profitability is likely to arise, we can determine whether (i) the Federal Circuit's approach of providing narrow patents for all advances or (ii) the traditional approach of providing broader

patents, but only for exceptional advances, is better tailored "to promote the Progress of the useful Arts."

V. IDENTIFYING THE RELEVANT DESIRABILITY-PROFITABILITY GAP

In the absence of patent protection, private rights in tangible property alone will provide two types of incentive for innovation. First, in the almost certain absence of perfect competition, an individual who introduces a discrete innovation into the market will have some time where she faces competition only from the range of preexisting products available. Precisely to the extent that consumers consider her innovation a desirable improvement over the preexisting products, an innovator will enjoy some market power and a corresponding ability to price the tangible goods embodying her innovation above marginal cost. Sooner or later, however, this lead-time period will end. It may end when competitors notice the innovator's rents and copy the innovation in an effort to obtain a share of the rents for themselves. It may also end if another, working entirely independently, happens to re-create the same innovation and bring it to the market shortly after the innovator. These new entrants will usually offer products that consumers consider closer substitutes for the original than the preexisting range of products, and will therefore usually reduce the extent to which an innovator can price her product above marginal cost in the post-entry period. Nevertheless, particularly in an oligopolistic market, prices for the innovative product are unlikely to fall to marginal cost. Product differentiation, tacit collusion, and a variety of other strategies will tend to ensure that the innovator continues to earn some rents even after competitive entry has occurred.

The second type of incentive arises not from the introduction of a single innovation, but from an innovator's repeated introduction of innovations over time. In many industries, consumers will lack perfect information regarding product quality. If consumers desire a product at the industry's creative edge, but lack perfect information both as to where that creative edge lies and as to which products fall along it, a company, by maintaining a consistently innovative product line, may establish a reputation as an innovator. A consumer would not therefore need to know the current state of running shoe technology, clothing design, or the latest mathematical advances in portfolio diversification theory to find innovative products. Instead, she could simply turn to a company with a reputation for innovation in those fields. Once developed, a reputation for innovation would

become both a source of rents for, and an informal guarantee of, future innovation.[167]

The following sections model formally each of these sources for innovation rents.

A. Innovation Rents in the Absence of Patent Protection: Lead-Time Rents

The first source of innovation rents available in the absence of patent protection arises from the lead-time advantage usually available to innovators. If we assume that consumers can identify an innovative product or service when it is introduced and consumers find a particular innovation desirable, an innovator will be able to charge somewhat more for her product until such time as a competitor either copies or independently re-creates a similarly innovative product. As such competitors enter the market, both the price our innovator can charge for the original and her share of the innovation's market will steadily fall.

As a result of declining price and market share, we should expect the lead-time rents available to decrease steadily (more or less) over time, allowing us to model the lead-time rents using geometric decay beginning from some initial price, p:

$$R_i = \int_{t_i}^{\infty} \delta^t p e^{-\varphi t} dt \qquad (14)$$

In equation (14), t_i represents the time at which the innovation is first introduced, δ is a discount factor, and φ is an increasing function of the number of independent re-creators, N_i, and the number of copiers, N_c, that enter the market. Both N_i and N_c are determined endogenously. We will assume that additional independent re-creators and additional copiers will continue to enter the market until the expected rents available to each precisely equal their respective costs of entering the market.

If we assume that all of the independent innovators are equally efficient at innovating, then the lead-time rents available to each are identical as are the costs of innovating. Additional independent innovators will therefore continue to enter the market until the lead-time rents available (as defined in (14)) exactly equal the cost of innovation, or until $R_i = C_i$.

[167] If we define a product's innovativeness as the relevant quality that consumers are seeking, this result is suggested by Klein and Leffler's analysis. See Benjamin Klein and Keith B. Leffler, *The Role of Market Forces in Assuring Contractual Performance*, 89 J Pol Econ 615 (1981).

Copiers must wait until after an innovation is introduced before beginning the imitation process. The rents (or quasi-rents since they are going to earn zero profit) available therefore begin at some time t_c, rather than at time t_i, where $t_c > t_i$. If we assume further that consumers consider the products offered by innovators and imitators to be perfect substitutes for each other, then the rents available for imitation will equal:

$$R_c = \int_{t_c}^{\infty} \delta^t p e^{-\varphi t} dt \tag{15}$$

Under these assumptions, the rents available from copying are strictly less than the rents available for innovation. An individual will therefore rationally choose to engage in copying only if the costs of copying are strictly less than the costs of innovating. Under the zero profit condition, additional imitators will continue to enter the market until $R_c = C_c$.

Even in the absence of patent protection, the lead-time rents available will prove sufficient for those innovative products with an expected cost, C_i, less than or equal to R_i. We can define this threshold cost as $\overline{C}$. Because the number of independent innovators and the number of copiers are endogenously determined as functions of the price consumers are initially willing to pay for the innovation, the costs of innovation, the costs of imitation, and the time required to imitate, we can define $\overline{C} = f(p, C_c, t_c)$.

Comparative Statics. Before comparing R_i to the ideal, taking the partial of R_i with respect to φ establishes that R_i is a decreasing function of φ and also therefore a decreasing function of N_i and N_c. From equation (15), it is readily apparent that, all else constant, R_c and hence N_c decrease as the time required to imitate an innovation increases. Similarly, holding R_c constant, N_c will decrease as C_c increases. R_i therefore increases as the time and cost of imitation increase.

Comparison to the Ideal. If we assume that all other markets are complete and perfectly competitive, so that a partial equilibrium analysis is appropriate, then the social value of the innovation in our model is defined by:

$$W_i = \int_{t_i}^{\infty} \delta^t p dt \tag{16}$$

From a social perspective, investing in an innovation is desirable so long as:

$$W_i \geq C_i \tag{17}$$

However, the lead-time rents available in the absence of patent protection will generally fall short of W_i. The only instance where R_i will equal W_i will be where the innovator is a monopolist so that $N_i = N_c = 0$ and therefore $\varphi(N_i, N_c) = 0$. In the usual case, where competitive entry can occur, the lead-time rents available will leave a potential gap between profitability and desirability when:

$$W_i \geq C_I \geq R_i \tag{18}$$

B. Innovation Rents in the Absence of Patent Protection: Reputation Rents

The second source of innovation rents available in the absence of patent protection arises when consumers desire an innovative product or service, but lack perfect information regarding which products or services incorporate the latest innovations. In the presence of imperfect information, developing a reputation for innovation can provide a source of rents (or quasi rents) to fund innovation. Consider a case where companies engage in a continual process of product improvement and where all consumers purchase the product either once (or not at all) on a periodic basis. To simplify the calculations involved, we will use the following three-period model, with each period having a uniform duration of t:

In period 0, a company chooses whether or not to invest in research and development to ensure that its product or service remains at the cutting edge of innovation. If the company chooses to invest, then it will have an innovative product to offer at the start of period 1. Alternatively, the company may opt to copy the innovation of others. If a company chooses to copy, the company makes no R&D investment in period 0. Instead, so long as at least one of its competitors engages in R&D and introduces an innovative product at the start of period 1,[168] the company copies that innovation during period 1. Copying the product requires some time, however. As a result, a copier will not introduce its version of the innovative product at the start of period 1, but after some fraction, $\varphi \in (0,1)$, of period 1 has elapsed.

A given number of homogenous consumers, represented by the range $[0,1]$, are each willing to pay a premium, p, over the price for the non-innovative version of the product in order to receive the innovative version of the product. For the sake of convenience and to focus on the incentive for innovation, we will assume that the price for the non-innovative version of a product is zero. Each consumer purchases the product once during time periods 1 and 2, with consumer pur-

[168] If no competitor innovates, then copying is not an option. In such a case, the only choice is between innovating or not innovating.

chases distributed uniformly over each time period. In order to ensure equilibrium, we will assume that consumers can determine the total market share of the companies that will innovate in period 1, but they cannot determine before their purchase whether a particular company's product is the innovative or non-innovative version. After they have purchased, they will learn whether the product they purchased was innovative. If they receive a non-innovative product in period 1, then they will assume that that company's products will remain non-innovative in period 2. Consumers who are disappointed with a company's products in period 1 will not only refuse to pay p for that company's products in period 2, but will switch and purchase the product from an innovative company in period 2.[169]

Let μ_i be the company's market share at the start of period 1, μ be the total market share of innovative companies, and the remainder, $1-\mu$ be held by competitors who imitate. In period 1, the company will be able to charge μp to all its customers, whether the company innovates or not. In period 2, however, if the company was innovative (i.e. invested in R&D in period 0), then the company will be able to charge p to its existing market share. The company will also capture a proportionate share of the consumers who purchased a non-innovative product from a copying competitor. With a discount rate per period of δ, the company's discounted revenue from being innovative is:

$$R_i = \delta\mu_i p\left\{\mu + \delta\left[1 + \varphi\left(\frac{1-\mu}{\mu}\right)\right]\right\} \tag{19}$$

If the company chooses to imitate, rather than innovate, then it will charge μp to all of its customers in period 1. However, because of the time required for imitation, a fraction, φ, of its consumers will actually receive a non-innovative version of the product. As a result, in period 2, a copying competitor will lose a corresponding fraction of its market share and be able to charge p only to its remaining customers. A copying competitor will therefore earn discounted revenue of:

$$R_c = \delta\mu_i p[\mu + \delta(1-\varphi)] \tag{20}$$

A reputation for innovation therefore generates an incentive to innovative precisely equal to the difference between the discounted innovation revenue and the discounted imitation revenue:

$$I_R = R_i - R_c = \delta^2\mu_i p\varphi\left(\frac{1}{\mu}\right) \tag{21}$$

[169] The working assumption behind this formulation of the model is that consumers, if they purchased a non-innovative version of the product in period 1, will poll a sufficiently large sample of their peers and purchase from one of the companies unanimously reported to be innovative.

If innovating entails a fixed cost C_i in period 0, and imitating entails a fixed cost C_C in period 1, then a company will invest in the R&D necessary to keep its product at the innovative forefront so long as:

$$I_R + \delta C_C \geq C_i \tag{22}$$

Comparative Statics Analysis. Before comparing I_R to the ideal, taking partial derivatives of I_R with respect to a company's market share, the time required for imitation, and the proportion of the company's competitors engaged in copying rather than innovation establishes the following:

1) The innovation rents available from reputation are a nondecreasing function of a firm's market share.

Specifically, if we assume that μ_c remains constant as μ varies, then the partial of I_R with respect to μ is:

$$\frac{\partial I_R}{\partial \mu} = \delta^2 p \varphi \left(\frac{1}{\mu}\right) > 0$$

On the other hand, if the company at issue is the only innovative firm in the market, so that $\mu = \mu_i$, then (21) becomes:

$$I_R = \delta^2 p \varphi \tag{21'}$$

In this case, the partial of I_R with respect to μ exactly equals zero.

2) The innovation rents available from reputation increase with the time required for others to copy an innovation.

If we assume that μ_i and μ remain constant as φ varies, then the partial of I_R with respect to φ is:

$$\frac{\partial I_R}{\partial \varphi} = \delta^2 \mu_i p \left(\frac{1}{\mu}\right) > 0$$

3) The innovation rents available from reputation increase with the proportion of competing firms that imitate, rather than innovate.

To see this, we will define the proportion of the company's competitors engaged in copying, rather than innovation as $\lambda \in [0,1]$. Formally, let $\lambda = 1 - \mu/1 - \mu_i$ or $\mu = 1 - \lambda(1 - \mu_i)$. Substituting for μ, we can rewrite (21) as:

$$I_R = \delta^2 \mu_i p \varphi \left[\frac{1}{1 - \lambda(1 - \mu_i)}\right] \tag{21''}$$

Given (21''), the partial of I_R with respect to λ becomes:

$$\frac{\partial I_R}{\partial \lambda} = \delta^2 \mu_i p \varphi \frac{1 - \mu_i}{[1 - \lambda(1 - \mu_i)]^2} \geq 0$$

with equality when $\mu = 1$.

Comparison to the Ideal. If we assume that all other markets are complete and perfectly competitive, so that a partial equilibrium analysis is appropriate, then the social value of the innovation in our model is defined by:

$$W_i = \delta(1 + \delta)p \tag{23}$$

In the absence of patent protection, the reputation rents otherwise available will generally fall short of the ideal unless the company is a monopolist.[170] If the company is a monopolist, such that $\mu_i = 1$, then the discounted revenue from innovation in (19) will exactly equal the optimal level of incentive, W_i, as set forth in (23). Moreover, if the company is a monopolist, there is no other from whom the company can copy, and if no one innovates, then consumers will not (by assumption) pay p for any product. As a result, the discounted revenue

[170] A second exception may arise if an innovator can license her innovation to her competitors. See, e.g., Arrow, *Economic Welfare and the Allocation of Resources for Invention* at 619-22 (cited in note 138). If the innovator agrees to license her innovation to all of her competitors, enabling the competitors to introduce the innovative product either at the start of period 1 or in any event before they could introduce the innovative product through copying, then competitors would be willing to pay a license fee for such earlier access to the innovation. The maximum license fee that a competitor would be willing to pay would depend upon the fixed cost of imitating and the additional revenue earned as a result of the time saved by licensing. If there are a sufficient number of competitors willing to license, and the cost and time required to imitate are sufficiently large, the license fees available could equal (or by eliminating the deadweight loss of competitors' reinventing the wheel) exceed the optimal innovation incentive. While a full consideration of this possibility is beyond the scope of this article, such a licensing arrangement would present severe coordination problems. For a firm to invest in period 0 relying on potential licensing revenue at the start of period 1, the firm would have to know which competitors would be willing to license in period 1, rather than innovate themselves. Moreover, firms intending to license at the start of period 1 would have to know that some other firm would innovate and hence have the innovative version of the product available for licensing at the start of period 1. In addition, if such innovation and subsequent licensing appears an attractive option for one firm and the firms face similar innovation and copying costs, the other firms would also presumably prefer to innovate and collect licensing fees, leaving no firms interested in licensing. Patents might prove useful as a means of facilitating (or forcing) such licensing agreements, but because of the two- to three-year delay typically involved in obtaining a patent, it is not clear that patents are well-tailored to facilitate licensing in the model's context of continual, ongoing product improvement over relatively short time periods. For an approach to the issue of patent scope that focuses on the role of licensing for second generation products in the absence of uncertainty, see Scotchmer, 27 RAND J Econ (cited in note 155).

from copying in (20) will exactly equal zero. Under these assumptions, the reputation rents available to a monopolist will exactly equal the optimal level of incentive.[171]

In the more general case, where $\mu_i < 1$, I_R will usually fall short of W_i, creating the potential for a gap between profitability and desirability. From a social perspective, investing in the innovation is desirable so long as:

$$W_i \geq C_i \tag{24}$$

However, a reputation for innovation will ensure innovation only to the point where $I_R + \delta C_C \geq C_i$. A reputation for innovation therefore leaves a potential gap between profitability and desirability when:

$$W_i \geq C_i > I_R + \delta C_C \tag{25}$$

As was the case for lead-time rents, the availability of reputation rents will tend therefore to ensure some level of innovation. Precisely how much will depend on the time and expense required to imitate, rather than innovate, and upon the market share of the innovator. If we take the practical difficulty of imitating, the proportion of competitors who will innovate, and the innovator's market share as given, and assume that a reputation for innovation will develop, then there will be some threshold innovation cost at which the costs of innovation will precisely equal the reputation rents, such that $C_i = I_R + \delta C_C$. If we define that threshold cost as $\overline{C} = f(\mu_i, \lambda, \varphi, C_c)$, then patent protection is necessary to close the gap between profitability and desirability left by reputation rents only for those innovations whose cost exceeds $\overline{C}$.

C. The Relevant Gap: Unusually Creative Innovations

This analysis suggests that in the absence of patent protection the rents available from the lead-time advantage or the reputation obtained through innovation leave a gap between profitability and de-

[171] Please note that if consumer demand for the innovative product is not a step function, then in the absence of perfect price discrimination, the reputation rents available even to a monopolist will fall short of the optimal level of incentive. Because a step function demand is unlikely to represent actual market conditions, reputation rents will almost invariably fall below the optimal level of incentive. However, for purposes of identifying where the relevant profitability-desirability gap arises, this fact is not relevant for two reasons. First, it does not change the qualitative conclusion that there is some incentive available for routine innovation even in the absence of patent protection. Second, we must be careful not to justify providing patent protection using the desirability-profitability gap that arises from an inability to price discriminate perfectly if the innovator will remain equally unable to price discriminate perfectly with a patent.

sirability for innovations whose cost exceeds some threshold level, $\overline{C}$. Even in the absence of patent protection, the availability of lead-time rents or reputation rents will ensure an adequate incentive for the creation of less costly (or "routine") innovations. Particularly given the availability of copyright protection to guard against simple mechanical duplication, the relevant profitability-desirability gap that patent law should address appears to be the gap that arises for those more costly innovations that otherwise lack sufficient incentive from the marketplace.

Having identified the underlying economic considerations and the relevant profitability-desirability gap, we now move to an evaluation of the normative desirability of the quiet revolution that the Federal Circuit has brought to patents measured against the costs of uniformity and the information costs of individually tailored (or variable) protection.

VI. UNIFORMITY, VARIABILITY, AND INFORMATION COSTS

Although the switch to narrower, but routinely valid patents appears to miss the relevant desirability-profitability gap, the switch may nonetheless enhance social welfare if the information and other transaction costs entailed in tailoring patent protection more closely to the relevant gap are sufficiently high. As discussed, uniformity in patent protection is a potentially rational response to high information and other transaction costs. With perfect information and costlessly enforceable legal rules, a government regulator should individually tailor protection to each innovation, ensuring that each innovation receives precisely that protection, but no more, necessary to ensure its development and disclosure. Gathering the information and then enforcing the legal rules necessary to achieve such individually tailoring is costly, however. To explore this trade-off, and its implications for the normative desirability of the switch from rarely valid, but broadly enforced, to routinely valid, but narrowly enforced patents, we begin with a numerical example.

A. An Illustration of the Uniformity-Variability Trade-Offs

Assume that there are four innovations potentially eligible for patent protection. Each innovation has an unrestricted social value of 35, but the innovations require varying levels of patent protection to ensure their expected profitability, ranging from no protection for innovation A to level 3 protection for innovation D. As the degree of patent protection increases, the associated restrictions on the use of the

Table 1. Distribution of Potentially Patentable Innovations

Innovation	Protection Level Required	Unrestricted Social Value
A	0	35
B	1	35
C	2	35
D	3	35

information aspect of the protected innovation steadily increase, and the associated social value of the protected innovation steadily decreases. For purposes of this example, we will assume that the social value of an innovation falls by 10 for each level of patent protection provided. Table 1 summarizes these assumptions:

Case #1: No Specific Information. In the first case, the government regulator is aware of the distribution of the innovations, their unrestricted social values, and the reduction of social value associated with increasing levels of patent protection. While the regulator can also distinguish innovations from prior art, the government regulator has no information regarding the level of patent protection required to ensure the expected profitability (and hence, the existence) of any given innovation. Because the regulator cannot distinguish those innovations that would not have been devised or disclosed but for the inducement of a patent, the regulator cannot enforce a rule that would exclude innovation *A* from patentability. Similarly, lacking the necessary information to differentiate the various innovations, the regulator cannot tie the protection provided any given innovation to the level required to ensure its expected profitability. As a result, the regulator has no choice but to provide a uniform level of protection for all innovations. Given the information available, and the resulting choice the regulator faces, the optimal decision is to provide level 1 to all innovations.

Uniformly providing level 1 protection increases social welfare compared to providing no protection. By uniformly providing level 1 protection, the regulator ensures the expected profitability of innovation *B*, generating an increase in social welfare of 25—the level 1 restricted social value of innovation *B*. On the other hand, providing uniform level 1 protection also reduces the social value of preexisting innovation *A* by 10. Moving from level 0 to level 1 uniform protection thus generates a net social gain of 15 and is therefore desirable.

In contrast, moving from level 1 to level 2 is undesirable, if the increased protection is made uniformly available to all innovations. Moving to a uniform level 2 ensures the expected profitability of innovation *C* and thereby generates social value of 15—the social value

of innovation *C* when use of innovation *C* is restricted by level 2 protection. However, if the protection is uniformly provided, moving to level 2 protection reduces the social value associated with preexisting innovations *A* and *B* by 10 a piece. Moving to level 2 therefore generates a net social loss of –5.

Case #2: Regulator Can Identify Those Innovations That Will Be Devised and Disclosed Without a Patent, But Cannot Distinguish Between Innovations B, C, and D. Again, given the limited information available, the regulator will choose uniform protection for those innovations that receive a patent. However, because the regulator can distinguish those innovations that would be devised and disclosed without a patent, *i.e.* innovation *A*, the regulator will adopt and enforce a rule excluding innovation *A* from patent protection. Facing a choice of providing some level of uniform protection to the remaining innovations, the regulator will then choose to offer level 2 protection uniformly to innovations *B*, *C*, and *D*.

By moving from level 1 protection to level 2, the regulator ensures the existence of innovation *C* and thereby generates social value of 15. Because the regulator has excluded innovation *A* from patent protection, moving to level 2 protection reduces the social value of only one preexisting innovation, innovation *B*. Although moving from level 1 to level 2 protection reduces the social value of innovation *B* by 10, in this case, the switch to level 2 generates a net social gain of 5 and is therefore desirable. By excluding innovation *A* from patentability, the regulator reduces the social cost of expanding uniform patent protection and can therefore offer a more expansive level of uniform protection to the remaining innovations. The regulator can therefore offer broader protection than was optimal in Case #1 and can thereby ensure the existence of the more expensive innovation *C*.

However, even after excluding innovation *A* from protection, switching to level 3 protection remains undesirable. Providing level 3 protection would ensure the expected profitability of innovation *D* and thereby generate a social gain of 5—the level 3 restricted social value of innovation *D*. Yet, if the same level 3 protection were uniformly provided to innovations *B* and *C* as well, providing uniform level 3 protection would reduce the social value of these preexisting innovations by 10 a piece. With uniform protection for innovations *B*, *C*, and *D*, moving from level 2 to level 3 protection would therefore generate a net social loss of –15, even if innovation *A* were excluded from protection. As a result, even though there would likely be a positive social rate of return associated with innovation *D*, the costs of uniformity dictate that we should not extend patent protection to ensure innovation *D*'s expected profitability.

Case #3: Regulator Has Perfect Information. If the regulator has

perfect information and can costlessly enforce an individually-tailored (or variable) system of patent protection, then the regulator would exclude innovation *A* from patent protection, would provide level 1 protection to innovation *B*, level 2 protection to innovation *C*, and level 3 protection to innovation *D*.

Choosing the Optimal Case: While these cases are laid out as if the information constraints were externally imposed on the regulator, we can also use this example to illustrate the process by which a regulator would endogenously choose whether to expend the resources necessary: (1) to distinguish (factually and legally) innovation *A* from innovations *B*, *C*, and *D*; and (2) to distinguish between innovations *B*, *C*, and *D*. As established in Part IV above, the relevant rule is straightforward: If the information and other transaction costs entailed in moving the legal system from Case #1 to Case #2, or from Case #2 to Case #3, are less than the marginal increase in the social value of the innovations under the respective optimal schemes, then the regulator should choose to spend the resources necessary to move from one case to the next. For example, given the assumed distribution of innovations and social values, moving from Case #1 to Case #2 generates a net social gain of 15.[172] If gathering the necessary information to distinguish innovation *A* from the others and then legally enforcing the exclusion of innovation *A* from patent protection costs less than 15, then the regulator should choose to expend the necessary resources to move the legal regime from Case #1 to Case #2. Similarly, moving from Case #2 to Case #3 also generates a net social gain of 15.[173] Again, if gathering the necessary information and enforcing the necessary legal rules to move from Case #2 to Case #3 costs less than 15, then the regulator should choose to expend the information necessary to move the legal regime to Case #3.

B. Evaluating the Normative Desirability of Narrower, but Routinely Valid Patents

In this light, the normative desirability of the switch to narrower, but routinely valid patents depends entirely on the relative magnitudes of the marginal gains from more closely tailoring patent protection to the relevant desirability-profitability gap we have identified in Part V

[172] Comparing the respective optimal schemes, such a move increases the social value associated with innovation *A* by 10 by excluding it from patent protection, reduces the social value associated with innovation *B* by 10 by increasing its patent protection from level 1 to level 2, and adds the social value of innovation *C*, restricted by level 2 protection, which is 15.

[173] This gain comes from reducing the level of patent protection provided to innovation *B* from level 2 to level 1 and from ensuring the creation of innovation *D*.

against the marginal information and other transaction costs such tailoring would entail. Although the precise magnitude of these respective gains and losses is an empirical matter, this analysis gives us a clearer picture of the costs the switch implicates and the circumstances under which the switch might prove desirable.

First, to the extent that the Federal Circuit's evisceration of the nonobviousness requirement effectively extends patent protection to those technical advances that would have occurred in any event, the Federal Circuit's doctrinal changes to the nonobviousness requirement limit the patent system's ability to encourage more costly innovation. In our example, the switch to narrower but routinely valid patents would essentially move us from Case #2 to Case #1. Under the traditional economic analysis, this is undesirable because it extends protection to innovation *A*, even though such protection is unnecessary to ensure innovation *A*'s development and disclosure.[174] Yet, this is not the only cost of the switch. After recognizing the costs of uniformity, we can see that granting patents to routine technical advances also limits our ability to use patent protection to ensure the expected profitability of high cost innovations. As we extend patent protection over a wider range of preexisting innovations, the costs of any given expansion in uniform protection increase. Granting patents to innovations, such as innovation *A*, that would have occurred even without a patent drags down the optimal level of uniform protection. Granting patents to routine technical advances will therefore limit our ability to provide patent protection sufficient to ensure the expected profitability of desirable, but higher cost innovations.

Second, in addition to limiting the range of innovation we can encourage through a uniform patent system, the switch towards narrower, but routinely valid patents also removes two of the three key doctrines through which courts could introduce desirable variability into the level of patent protection provided individual innovations.[175]

[174] In addition, under the traditional analysis, if we focus solely on those innovations that would still exist, the social loss from providing patent protection to innovation *A* would be offset by the social gain from providing only level 1 protection to innovation *B*.

[175] Professors Burk and Lemley have suggested that the use of the "person having ordinary skill in the art" or "PHOSITA" in the nonobviousness requirement implicitly incorporates such variability across arts. See Dan L. Burk & Mark A. Lemley, *Is Patent Law Technology Specific?*, 17 Berkeley Tech L J 1155 (2002). As they explain, for those arts where the PHOSITA has more skill, that will make the nonobviousness requirement tougher to satisfy because the more highly skilled PHOSITA will more readily see any given difference between a claimed invention and the prior art as obvious. Although they did not extend their argument to the infringement inquiry, PHOSITA also plays a similar role in the doctrine of equivalents. One of the factors

Even after we interpret the nonobviousness requirement to exclude from patentability those technical advances that would be disclosed or devised without a patent, there will remain a wide range of innovative products that patent protection could potentially ensure. There will be some innovative products whose expected cost only slightly exceeds $\overline{C}$. For these innovative products, some minimal level of patent protection will suffice to ensure expected profitability. There will also be some innovative products whose expected costs far exceed $\overline{C}$. For these innovative products, we will need to promise far more substantial patent protection to ensure their expected profitability.

Traditionally, courts used the nonobviousness doctrine,[176] the literal language of the patent claims, and the doctrine of equivalents to introduce the desired variation into patent law's otherwise uniform protection. For example, courts have traditionally granted "pioneering inventions"[177] a broader range of equivalents and a correspond-

---

courts have identified as relevant to determining whether a given element is an equivalent to a claimed element is whether a PHOSITA would recognize the given element as a substitute for the claimed element. Again, in those arts where the PHOSITA has a higher skill level, presumably the PHOSITA will more readily recognize such substitutability and hence a broader range of equivalents will apply. While varying the skill level of the PHOSITA associated with different arts can therefore vary the requirements for and scope of patent protection for different arts, such an approach does not vary the requirements for or scope of protection within an art. Once the skill level of the PHOSITA for a given art has been set, presumably that same skill level will apply to all patents in the relevant art and hence reinforce uniformity of patent protection for each given art. My working assumption is that, *within each art*, there is a range of innovations that patent protection could potentially ensure. As a result, manipulation of the PHOSITA standards across different arts cannot introduce the requisite variation within an art.

[176] Although the nonobviousness requirement is a bright line in a legal sense—obvious technical advances are not patentable; nonobvious technical advances are—application of the nonobviousness requirement to particular cases is a necessarily human and therefore imprecise exercise. If we place technical advances on a spectrum from the least significant advances to the most significant, at one end will be those technical advances that virtually all judges in all cases will find obvious; at the other end will be those technical advances that virtually all judges will find nonobvious. As we move from one end of the spectrum to the other, the chance that a technical advance, given the judge drawn and the attorneys and parties involved, will survive an obviousness challenge steadily increases. If patented inventions otherwise receive identical protection, increasing or decreasing the chance that the invention will be found obvious can introduce variation into the effective protection provided any given patented invention.

[177] The Court has defined a pioneering invention as one that is "a distinct step in the progress of the art, distinguished from a mere improvement or perfection of what had gone before." *Westinghouse v Boyden Power Brake Co*, 170 US 537, 562 (1898); see also *Texas Instr, Inc v US Intl Trade Comm'n*, 846 F2d 1369, 1370 (Fed Cir 1988) (adopting the definition of pioneering invention from *Westinghouse*, 170 US at 562).

ingly broader scope under the doctrine of equivalents.[178] By narrowly interpreting both the nonobviousness requirement and the doctrine of equivalents, the switch to routinely valid, but narrowly enforced patents sharply reduces the room for judicially tailoring patent protection to the individually optimal level. After the switch, courts can rely only on the more limited discretion left in interpreting the words of the patent claim to introduce the desired variation in a patent's scope. While interpreting patent claims offers some room for individually tailoring protection,[179] the discretion available through claim interpretation is not as well suited as the nonobviousness requirement and the doctrine of equivalents to matching the economic rents generated to the reservation cost for a particular innovation.

The key issue in determining a patent's effective scope is its effectiveness at excluding competitors from a distinct product market, either by increasing the cost of introducing, or delaying the introduction of, would-be competitors' products. For a number of reasons, the doctrine of equivalents addresses this issue far more directly than the process of claims interpretation. A claim is written before the market has developed, when the precise nature of consumer preferences and the various forms that competitors' products can take is as yet uncertain. The general rule of claim construction—that "the construing court interprets words in a claim as one of skill in the art *at*

[178] The Federal Circuit initially embraced this aspect of the doctrine of equivalents, see *Perkin-Elmer Corp v Westinghouse Elec Corp*, 822 F2d 1528, 1532 (Fed Cir 1987); *Thomas & Betts Corp v Litton Sys, Inc*, 720 F2d 1572, 1579 (Fed Cir 1983), but has subsequently questioned it. See *Augustine Medical, Inc v Gaymar Indus., Inc*, 181 F3d 1291, 1301 (Fed Cir 1999). Despite this questioning, this application of the doctrine of equivalents, along with the doctrine more generally, survives, at least for now. See, e.g., *Molten Metal Equipment Innovations, Inc v Metaullics Systems Co*, 2003 US App. LEXIS 1821, at *11, 14 (Fed Cir 2003) (reinstating jury verdict finding infringement based upon the following instructions: "In the event an invention achieves a major or extraordinary advance over the prior art, and as such may properly be characterized as a pioneering invention, the claims are entitled to a broad or liberal range of equivalents. On the other hand, if the advances over the prior art are narrow or minor, the range of equivalents is correspondingly more restricted and the claims are entitled to only a narrow range of equivalents.").

[179] For example, in the case of pioneering inventions, a panel of the Federal Circuit has stated:

> Without extensive prior art to confine and cabin their claims, pioneers acquire broader claims than non-pioneers who must craft narrow claims to evade the strictures of a crowded art field. Thus, claim scope itself generally supplies broader exclusive entitlements to the pioneer.

Augustine Medical, Inc, 181 F3d at 1301. The absence of a well-established record of prior arts for new patentable subject matters, such as computer programs and business methods, has also left room for often overly broad claims in these newly patentable fields.

the time of invention would understand them"[180]—reinforces the *ex ante* nature of the discretion inherent in claim interpretation. In addition, while patent attorneys always attempt to balance obtaining broader claim language against the legal prerequisites of patentability, at the drafting stage, uncertainty over the form competition may eventually take,[181] as well as the certainty that only a fraction of issued patents will ever result in litigation, tend to tilt that balance towards satisfying the legal prerequisites for patentability.

In contrast, under the doctrine of equivalents, we resolve the question whether a substituted element is the equivalent of a claimed element as of the time the infringement occurs.[182] We resolve the doctrine of equivalents issue only for those patents the value of which the fact of litigation has proven. And traditionally, the doctrine of equivalents, rather than focus on the necessarily imprecise words used to claim the invention, focused more directly on the extent to which the allegedly infringing product or process is likely to serve as a competitive substitute for the patented invention. However artful a patent attorney may be, focusing on whether the allegedly infringing device or process "performs substantially the same function in substantially the same way to achieve substantially the same result" is likely to address far more directly the competitive substitutability of defendant's device or process than attempting to match the defendant's device or process to the precise words of a patent claim. The discretion available under the doctrine of equivalents is therefore likely to prove far more effective at identifying competitive substitutes and, for that reason, far more useful in tailoring individually optimal patent protection.[183]

[180] *Eastman Kodak Co v Goodyear Tire & Rubber Co*, 114 F3d 1547, 1555 (Fed Cir) (emphasis added), *modified on other grounds on rehearing*, 114 F3d 1564 (Fed Cir 1997).

[181] Professor Mark Lemley has made a similar argument with respect to the Patent and Trademark Office's often cursory examination of patent applications. See Mark A. Lemley, *Rational Ignorance at the Patent Office*, 95 Nw U L Rev 1495 (2001).

[182] See, e.g., *Warner-Jenkinson Co v Hilton Davis Chem Co*, 520 US 17, 37 (1995) ("Insofar as the question under the doctrine of equivalents is whether an accused element is equivalent to a claimed element, the proper time for evaluating equivalency—and thus knowledge of interchangeability between elements—is at the time of infringement, not at the time the patent was issued.").

[183] In criticizing the traditional rule that pioneering innovations receive a broader range of equivalents, the *Augustine Medical* panel offered the following argument:

> At the outset, this court notes that no objective legal test separates pioneers from non-pioneers. Furthermore, it is impossible for this court or the PTO to predict the future of any given technology and thereby determine the likelihood that an invention will open vast new vistas of innovation. The peripheral claiming system itself, however, makes the best distinction between pioneers and

A vibrant nonobviousness doctrine offers similar advantages. Relying on claim language alone presupposes that there is some definable relationship between the scope of a patent and the rents available to the patent owner. Yet, one of the great mysteries in economics is how prices will behave in cases other than perfect competition or monopoly. If protection results in an effective monopoly over the innovative product's market, then economic theory provides a reasonably definitive guide to the likely rents available to the patent holder. If the monopoly rents available happen to equal precisely the reservation cost for the innovation, then matching patent scope to reservation cost is relatively straightforward. If, however, the monopoly rents available would exceed the innovation's reservation cost, the question becomes more difficult. One possibility would be to narrow the scope of the patent, so that the patent holder no longer possesses an effective monopoly. Others could then enter the market, and we can hope that the resulting imperfect competition may generate rents that approximate the innovation's reservation costs.

The difficulty with such an approach is that once we allow others to enter the market, economic theory no longer provides a single answer to the rents the patent holder will likely collect. Even a move from one firm in a market to two firms in a market renders uncertain the rents a firm will collect. Under one economic model—the Cournot model—prices in a market with two competitors offering identical goods fall somewhat from monopolistic levels, but remain well above competitive levels. In contrast, under another model—the Bertrand model—prices in the same duopoly market fall to marginal cost.

As a result of these uncertainties, if we know that the reservation cost for a given innovation is exactly half the monopoly rents available, there is no clear answer as to how to achieve that return if we rely solely on narrowing the patent's scope. In contrast, we could ensure the appropriate incentive by granting a patent that would provide an effective monopoly over the market, if valid, but impose a

---

non-pioneers. Pioneers enjoy the benefits of their contribution to the art in the form of broader claims.

Augustine Medical, Inc v Gaymar Indus, Inc, 181 F3d 1291, 1301 (Fed Cir 1999). The *Augustine Medical* panel's argument is curious. By the time a patent issues and comes to litigation, there is often no need for the court or the PTO "to predict the future" to determine whether a patented invention has proven pioneering. The "new vistas of innovation" will either have developed or not. If they have not, certainly the court will be in a better position to evaluate the likelihood that they will develop at the time litigation ensues that either the PTO, the patent applicant, or the patent attorneys were at the time the application was filed and prosecuted.

fifty percent chance that the patent will be held invalid as obvious. If patent holders are risk neutral, as one would expect for corporations with large patent portfolios, then the expected rent from a fifty percent chance of having an effective monopoly would approximate the appropriate incentive.

By strictly limiting the discretion available under the nonobviousness doctrine and the doctrine of equivalents, the Federal Circuit's switch to routinely valid, but narrowly enforced patents limits our ability to introduce desirable variation into the otherwise uniform protection patents provide. Limiting the patent system's ability to tailor protection to a particular innovation's individually optimal level limits in turn the patent system's ability to provide protection sufficient to ensure the expected profitability for the full range of desirable innovations eligible for patent protection.

While a conclusive empirical resolution of the optimal level of variability in patent protection is beyond the scope of this paper, our analysis establishes that the ultimate question is whether the information and other costs entailed in crafting and enforcing a somewhat more discerning system are justified by the benefits from: (1) reducing the extent to which some innovations are overprotected in a more uniform system; and (2) increasing our ability to provide protection sufficient to ensure the expected profitability of more costly innovations. The switch to routinely valid patents is normatively desirable under our analysis if and only if we cannot reliably (or at a reasonable cost) separate those innovations that would have been devised and disclosed without a patent from those innovations that would not be devised and disclosed but for a patent. Similarly, the switch to narrower patents is normatively desirable if and only if we cannot reliably (or at a reasonable cost) differentiate those innovations that require somewhat more extensive patent protection to ensure their expected profitability from those innovations that require somewhat less. Only if these information costs are prohibitive should we close our eyes and embrace the Federal Circuit's "one size fits all" patent protection.

C. Brief Discussion of the Supposed Advantages of the Switch

Although the Federal Circuit has not offered a policy justification for its rewriting of the nonobviousness requirement, the Federal Circuit and the Court have offered two arguments to justify their decisions to narrow the doctrine of equivalents. First, the Federal Circuit has argued that a narrower doctrine of equivalents forces the patent holder to internalize the cost of defining a patent's scope. In *Sage*

Products, Inc. v. Devon Industries, Inc.,[184] the Federal Circuit held that the doctrine of equivalents could not be used to encompass a foreseeable variation of the patented invention.[185] In justifying that rule, the panel explained:

> [A]s between the patent holder who had a clear opportunity to negotiate broader claims but did not do so, and the public at large, it is the patent holder who must bear the cost of its failure to seek protection for this foreseeable alteration of its claimed structure. . . . Because the doctrine of equivalents blurs the line of demarcation between infringing and non-infringing activity, it creates a zone of uncertainty, into which competitors tread only at their peril. Given a choice of imposing the higher costs of careful prosecution on patent holders, or imposing the costs of foreclosed business activity on the public at large, this court believes the costs are properly imposed on the group best positioned to determine whether or not a particular invention warrants investment at a higher level, that is, the patent holders.[186]

Although the panel's decision to portray the doctrine of equivalents as creating undesirable externalities renders the panel's reasoning superficially attractive, an arbitrary choice of phrasing does not represent a reasoned analysis. While the public at large benefits from competition, it also benefits from innovation. While the patent holder could have chosen to spend more on patent prosecution, would-be competitors can similarly choose to spend more to ensure that they avoid a patent's reach.[187] A vibrant doctrine of equivalents may limit

[184] 126 F3d 1420 (Fed Cir 1997).

[185] *Sage Prods, Inc*, 126 F3d at 1425.

[186] Id (citations omitted).

[187] Ian Ayres and Paul Klemperer have pointed out, as prices move from a competitive level to a full monopoly level, social costs are highest, relative to the additional profit accruing to the patent holder, for the last increment, as the patent holder's price moves from just below the full monopoly price to the full monopoly price. See Ayres & Klemperer, 97 Mich L Rev 985 at 989-93 (cited in note 12). Although I am not certain of the practical significance of this point for optimal design of a patent system, its corollary is that the social costs of supracompetitive pricing are lowest, relative to the patent holder's additional profit, for the first increment, as the patent holder's price moves incrementally from a competitive level to just above a competitive level. If, as Ayres and Klemperer argue, this relationship between social cost and rents suggests that we should leave a patent holder somewhat uncertain of the precise scope of her patent in order to discourage her from charging the full monopoly price, see id at 993-1007, we may also want to leave would-be competitors uncertain as to the patent's precise scope. Uncertainty for would-be competitors would tend to ensure that the patent holder could charge a price, at least, slightly above a competitive level and could thereby capture the rents available from such pricing. Such an approach would tend to provide patent holders with rents at the lowest social cost.

competitors' ability to enter a market, but narrowing the doctrine of equivalents is likely to increase the costs of patenting[188] and to reduce the effective scope of patents.[189] To the extent that patents play an important role in innovation, narrowing the doctrine of equivalents is likely to reduce the incentives for, and hence the resources invested in, innovation. Whether society is better off with somewhat more innovation or somewhat less is a difficult question, but it is not one that we can answer by pretending that society's interests lie solely in leaving room for competition.

Second, both the Federal Circuit and the Court have reiterated Justice Campbell's concern that a vibrant doctrine of equivalents injects undue uncertainty into a patent's scope.[190] But it is a mistake to assume that claim language, simply because it is written down, is necessarily clear. The language of a claim is not like the language of a deed. With a deed, the language describes physical dimensions and boundaries of a parcel of land and can do so quite literally. A patent claim, on the other hand, is meant to describe an intangible boundary—the scope and limits of an invention—and can necessarily do so only approximately. There is no reason to believe, simply as a matter of logic, that would-be competitors can better predict how their device or process will fare against a formal and legalistic interpretation of the language of a patent claim, than they can predict how it will fare against a more pragmatic test focusing on the essence of the patented invention.

Empirically, if the shift towards a narrower doctrine of equivalents increased certainty and predictability, then we should find a significant decrease in success rate variability under the Federal Circuit.[191]

[188] Even the Federal Circuit has acknowledged this point. See *Sage Prods, Inc*, 126 F3d at 1425 ("This court recognizes that such reasoning places a premium on forethought in patent drafting. Indeed this premium may lead to higher costs of patent prosecution.").

[189] Perhaps the Federal Circuit hopes that patent holders will draft and pursue vigorously broader claims so that the effective scope of protection remains effectively the same, but with less uncertainty. Yet, the striking increase in the percentage of no infringement results reflected in Figure 3 suggest that so far the effect of the Federal Circuit's doctrinal changes has been to narrow the effective scope of patents. See text accompanying notes 29-31.

[190] Compare *Warner-Jenkinson Co v Hilton Davis Chem Co*, 520 US 17, 29 (1997) ("There can be no denying that the doctrine of equivalents, when broadly applied, conflicts with the definitional and public-notice functions of the statutory claiming requirement."), with *Festo Corp v Shoketsu Kinzoku Kogyo Kabushiki Co*, 535 US 722, 732-33 (2002) (recognizing the uncertainty the doctrine of equivalents may create, but noting that "[t]hese concerns . . . are not new").

[191] This is particularly true given that the Federal Circuit's evisceration of the nonobviousness requirement has eliminated one of the significant sources of such variability.

Yet, the data reveals exactly the opposite—variability in the success rate has increased under the Federal Circuit. If we take the success rate for each of the six pre-Federal Circuit time periods as discrete data points, then the success rate over those six periods had a standard deviation of only 3.25 percent. In contrast, if we take each of the nine post-Federal Circuit time periods as discrete data points, then the post-Federal Circuit success rate had a standard deviation of 9.54 percent. Even if we limit our sample to the six post-Federal Circuit time periods since the Federal Circuit began to move towards a narrower doctrine of equivalents in *Pennwalt Corp.*, the standard deviation of the success rate remains a strikingly high 10.2 percent. If we assume that the considerations that drive the process by which parties select cases for appellate resolution have remained roughly constant, the increased variability in success rates suggests that parties are less able to predict appellate litigation outcomes under the Federal Circuit.[192] This in turn suggests that the Federal Circuit and its doctrinal changes have brought less certainty and predictability to patent enforcement.[193]

Reversal rates on the infringement issue provide further support for the proposition that narrowing the doctrine of equivalents does

[192] To explain this increased variability, it is not enough to suggest that district courts are unprepared to apply correctly the legal rules of claim interpretation. Kimberly A. Moore, *Are District Judges Equipped to Resolve Patent Cases?*, 15 Harv J L & Tech 1 (2001) (arguing that reversal rates are high because district judges are not well-equipped to interpret patent claims). Even if district judges routinely make mistakes that are unlikely to survive the *de novo* standard of review for claim construction that the Federal Circuit has imposed, see *Cybor Corp v FAS Techs*, 138 F3d 1448, 1454-56 (Fed Cir 1998) (en banc), increased variance in success rates should arise only if the parties are unable to predict how the Federal Circuit will resolve the claim construction issue. *De novo* review should make that prediction easier, because it is no longer clouded by the possibility that the level of deference given to a district court's claim construction under a clearly erroneous standard might vary among Federal Circuit judges. My own sense, agreeing with a view recently expressed by Polk Wagner and Lee Petherbridge, is that there remains a contingent of Federal Circuit judges that continue to follow the traditional approach to patents on both validity and infringement issues. Lunney, 7 Mich Telecommun & Tech L Rev 363 at 393 (cited in note 16) ("Years of jurisprudence based upon the traditional perspective are unlikely to disappear without a trace, particularly where the Court has so far refused to repudiate its own longstanding jurisprudence reflecting that perspective. Even some members of the Federal Circuit may retain some continuing commitment to the traditional perspective."); R. Polk Wagner & Lee Petherbridge, *Is the Federal Circuit Succeeding? An Empirical Look at Claim Construction* (available at http://www.law.upenn.edu/fac/pwagner/research.html) (last visited April 29, 2003). The increased variance thus arises from: (i) differences among Federal Circuit judges in interpreting claims; (ii) the random selection of judges for particular panels; and (iii) the Federal Circuit's refusal to identify the panel of judges who will hear a case until the morning of the appellate argument.

[193] Compare Dreyfuss, 64 NYU L Rev at 9-10 (cited in note 12) (arguing that Federal Circuit had brought greater certainty and predictability to patent enforcement).

not ensure certainty.[194] In the last year of our sample, for example, of the ninety-eight cases in which the Federal Circuit addressed the issue of infringement, the Federal Circuit reversed or vacated the district court's judgment on the issue of infringement in thirty-eight of them.[195] If a district judge with no particular stake in the issue has such trouble resolving the issue correctly, even after hearing argument from presumably competent counsel on both sides, there is little reason to believe that a would-be competitor, with advice only from its own counsel and its own self-interest in entering the market at stake, will prove better able to determine the precise boundaries of a patent under the Federal Circuit's approach.

These data suggest that the supposed certainty and predictability of claim language is simply a myth. In the end, because words will necessarily prove imprecise in defining the boundary of an invention, there is little reason to believe that emphasizing claim language will establish the certainty and predictability in defining patent boundaries that the Federal Circuit and the Court hope for. To the contrary, the available empirical evidence suggests that parties were better able

[194] Both Christian Chu and Kimberly Moore have documented the high reversal rates involving claim construction. See Christian A. Chu, *Empirical Analysis of the Federal Circuit's Claim Construction Trends*, 16 Berkeley Tech LJ 1075, 1097-1100 (2001) (noting that despite the changes to the infringement inquiry, "the promises of pre-trial predictability and expedient patent litigation seem to remain a tantalizing dream"); Moore, 15 Harv J L & Tech at 14 (cited in note 192) ("This means that more than one in four appealed patent cases involving claim construction result in overturning the judgment reached by the district court solely for claim construction reasons.").

[195] These cases include: *Intermatic Inc v Lamson & Sessions Co*, 273 F2d 1355 (Fed Cir 2001) (reversing judgment of infringement for improper claim construction and holding that, properly construed, patent could not be infringed as a matter of law), *vacated for further consideration in light of Festo Corp*, 535 US 722 (2002); *Xerox Corp v 3Com Corp*, 267 F3d 1361 (Fed Cir 2001) (reversing summary judgment of noninfringement for improper claim construction); *Tapco Intl Corp v Van Mark Prods Corp*, 2001 US App LEXIS 18330 (Fed Cir 2001) (reversing summary judgment of infringement for improper claim construction); *Durel Corp v Osram Sylvania, Inc*, 256 F3d 1298 (Fed Cir 2001) (reversing judgment of infringement for improper claim construction of term "oxide coating"); *Unique Coupons, Inc v Northfield Corp*, 2001 US App LEXIS 12839 (Fed Cir 2001) (reversing judgment of infringement for improper claim construction); *Somfy, SA v Springs Window Fashions Div, Inc*, 2001 US App LEXIS 8482 (Fed Cir 2001) (reversing summary judgment of noninfringement and remanding for trial on infringement under the doctrine of equivalents); *Mentor H/S, Inc v Medical Device Alliance, Inc*, 244 F3d 1365 (Fed Cir 2001) (reversing judgment as matter of law of noninfringement and reinstating jury verdict of contributory infringement); *Optimal Rec Solutions, LLP v Leading Edge Techs*, 2001 US App LEXIS 5772 (Fed Cir 2001) (vacating holding of noninfringement for improper claim construction); *AFG Indus v Cardinal IG Co*, 239 F3d 1239 (Fed Cir 2001) (vacating summary judgment of noninfringement for improper claim construction); *Wenger Mfg, Inc v Coating Mach Sys, Inc*, 239 F3d 1225 (Fed Cir 2001) (reversing summary judgment of noninfringement for improper claim construction).

to predict infringement outcomes under the traditional doctrine of equivalents' more pragmatic function-way-result test than they are under the precise words of the patent claim, legally interpreted.

Moreover, even if we assumed—contrary to what the empirical evidence suggests—that a narrower doctrine of equivalents slightly increased certainty in identifying patent boundaries, that increased certainty would come at a steep price. By reducing the doctrine of equivalents' ability to serve as a source of desirable variability in patent protection, narrowing the doctrine of equivalents directly limits our ability to tailor protection to ensure profitability for a wider range of desirable innovations eligible for patents. The certainty that the Federal Circuit seeks thus comes at the expense of ensuring the expected profitability (and hence, the likely existence) of more expensive, but still socially desirable innovations.

VII. INNOVATION, INFORMATION, AND PROPERTY RIGHTS

Under the traditional economic analysis of patents, the costs of patent protection arise directly from the tension between private rights and public goods. However, if providing patent protection ensures the creation of a desirable information product and does so more efficiently than the plausible alternatives, such as direct government subsidies, the fact that the information product could have been more valuable still in the absence of the patent's protection has little practical significance. Rather than follow the traditional analysis, this article identifies the uniformity of patent protection as the principal source of a patent's social cost. To the extent that we provide the same protection to all information products that satisfy a uniform set of prerequisites, broader patent protection entails a trade-off between: (i) the social value of the additional information products broader protection ensures; and (ii) the reduced social value associated with the preexisting information products protected more broadly than necessary to secure their discovery and disclosure.

Measured against the costs of uniformity, the switch to routinely valid, but narrowly enforced patents has two principal consequences. First, while, in combination, the social losses from loosening the nonobviousness requirement may roughly approximate the social gains from narrowing the scope of patents, at least, for those routine innovations that will remain, extending patent protection to a wider range of innovative products that would have been forthcoming with no or less protection reduces the optimal level of uniform protection. Second, at the same time, the switch removes from the patent system two of the three doctrines by which courts could attempt to vary the

level of protection provided to particular innovations. By reducing the ability of courts to tailor protection to each individual innovation, the switch to routinely valid, but narrowly enforced patents pushes us towards a more uniform, "one size fits all" system of patent protection. Given the costs of uniformity, the switch to routinely valid, but narrowly enforced patents will limit the range of desirable innovations that the patent system can ensure.

Appendix I

				Reason for Failure		Non-Final	
Period	# of Cases	"Success"	"Failure" Total	Invalid/Unenforceable	Not Infringing	Pro-Patentee	Pro-Infringer
1944-1946	45	14	28	24	4	2	1
1954-1956	86	28	51	37	18	7	0
1964-1965	57	21	33	19	19	2	1
1966-1967	60	18	36	30	6	4	2
1975-1976	59	19	36	27	11	3	1
1981-1982	63	24	29	24	5	9	1
1984-1985	76	27	30	18	12	19	0
1986-1987	99	38	39	17	26	19	3
1988-1989	86	43	31	13	18	10	2
1990-1991	93	27	42	17	24	19	5
1992-1993	103	31	49	20	33	12	11
1994-1995	92	27	50	11	39	14	1
1996-1997	173	49	94	38	55	26	4
1998-1999	172	38	103	29	75	26	4
2000-2001	228	38	136	34	109	50	4