Working Paper Series
No. 16001

Patent Claims and Patent Scope

Alan C. Marco
U.S. Patent and Trademark Office

Joshua D. Sarnoff
DePaul University

Charles A. deGrazia
U.S. Patent and Trademark Office (ADDX Corp.) and Royal Holloway, University of London

August 18, 2016

Hoover Institution Working Group on
Intellectual Property, Innovation, and Prosperity
Stanford University

www.hooverip2.org

U.S. Patent and Trademark Office
OFFICE OF CHIEF ECONOMIST
Economic Working Paper Series

Patent Claims and Patent Scope

Alan C. Marco
U.S. Patent and Trademark Office

Joshua D. Sarnoff
DePaul University

Charles A. deGrazia
U.S. Patent and Trademark Office (ADDX Corp.) and Royal Holloway, University of London

Original Version: January 2016
This Version: August 2016

USPTO Economic Working Paper No. 2016-04
Also available as Hoover IP[2] Working Paper

The views expressed are those of the individual authors and do not necessarily reflect official positions of the Office of Chief Economist or the U. S. Patent and Trademark Office. USPTO Economic Working Papers are preliminary research being shared in a timely manner with the public in order to stimulate discussion, scholarly debate, and critical comment.

For more information about the USPTO's Office of Chief Economist, visit www.uspto.gov/economics.

The USPTO's Patent Claims Research Dataset will be made available at www.uspto.gov/economics.

Patent Claims and Patent Scope

Alan C. Marco
U.S. Patent and Trademark Office

Joshua D. Sarnoff
DePaul University

Charles A. deGrazia
U.S. Patent and Trademark Office (ADDX Corporation) and Royal Holloway, University of London

Original Version: January 2016
This Version: August 2016

Abstract
Patent scope is one of the important aspects in the debates over "patent quality." The purported decrease in patent quality over the past decade or two has supposedly led to granting patents of increased breadth (or "overly broad" patents), decreased clarity, and questionable validity. Such patents allegedly diminish the incentives for innovation due to increased licensing and litigation costs. However, these debates often occur without well-defined measurements of patent scope. This paper explores two very simple metrics for measuring patent scope based on claim language: independent claim length and independent claim count. We validate these measures by showing that they have explanatory power for several correlates of patent scope used in the literature: patent maintenance payments, forward citations, the breadth of patent classes, and novelty. Using these data, we provide the first large-scale analysis of patent scope changes during the examination process. Our results show that narrower claims at publication are associated with a higher probability of grant and a shorter examination process than broader claims. Further, we find that the examination process tends to narrow the scope of patent claims in terms of both claim length and claim count, and that the changes are more significant when the duration of examination is longer.

Keywords: Patents, patent scope, patent claims, patent examination, patent quality, USPTO
JEL Classification Numbers: O3, O31, O32, O34, O38

Acknowledgements: We would like to thank Robert Kimble, Jesse Frumkin, Jamie Kucab, and Joseph Bailey, as well as numerous personnel at the US Patent and Trademark Office. We would also like to thank seminar participants at Harvard Business School and the Hoover Institution's IP^2 program.

The USPTO's Patent Claims Research Dataset will soon be made available at www.uspto.gov/economics.

1 Introduction

For many years, debates over the effectiveness of the patent system have focused on the central issue of "patent quality." In 2002, then-former Assistant Secretary of Commerce and Commissioner of the U.S. Patent and Trademark Office (PTO) Gerald Mossinghoff noted a "real concern that with the dramatic increase in the number of patent applications filed and patents granted - and with the influx of new and unavoidably inexperienced examiners hired to handle the workload - compromises to patent quality may be inevitable."[1] This increasing number of patents of purportedly diminishing quality supposedly led to dramatic increases in assertion of patents to extract rents through licensing and litigation, particularly by non-practicing entities (NPEs).[2] In turn, the purported decrease in patent quality supposedly led to diminished innovation due to increased licensing and litigation costs as well as to reduced sequential innovation in various industries, particularly in regard to software patents.[3]

---

[1] Gerald J. Mossinghoff & Vivian S. Kuo, *Post-Grant Review of Patents: Enhancing the Quality of the Fuel of Interest,* 43 IDEA 83, 83 (2002).

[2] *See, e.g.,* Patent Quality Improvement: Hearing Before the Subcomm. on Courts, the Internet, and Intellectual Property of the H. Comm. on the Judiciary, 109th Cong. 18 (2005) (statement of Richard J. Lutton, Jr., Chief Patent Counsel, Apple) ("The current patent system has given rise to too many low quality patents being issued, and a growing pattern of assertions of weak patents that threaten to damage productive companies and stifle innovation."); David L. Schwartz & Jay P. Kesan, *Analyzing the Role of Non-Practicing Entities in the Patent System,* 99 CORNELL L. REV. 425, 429 (2014) ("Under this narrative, NPEs assert marginal patents and read patent claims unreasonably expansively. Under any reasonable view, the patents are likely invalid or not infringed by the NPEs' targets. NPEs, who themselves do not innovate or introduce any products into the marketplace, merely extract rents from the large, innovative companies that they sue. They create fear of holdup by selecting venues where injunctive relief is available such as the International Trade Commission. They seek and accept 'nuisance' settlement amounts, far below the cost of litigation, so that the NPEs' targets have no incentive to defend in costly litigation."). *But see* John R. Allison & Ronald J. Mann, *The Disputed Quality of Software Patents,* 85 WASH. U. L. REV. 297, 298 (2007) ("In general, we find that patents the computer technology firms obtain on software inventions have more prior art references, claims, and forward citations than the patents that the same firms obtain on nonsoftware inventions. We also find that the patents that 'pure' software firms (those producing only software) obtain on software inventions have more prior art references, claims, and forward citations than the software patents obtained by the firms that derive revenues from other product lines.").

[3] *See, e.g.,* Arti K. Rai, *Improving (Software) Patent Quality Through The Administrative Process,* 51 HOUS. L. REV. 503, 505 (2013) ("Low-quality software patents … generate the usual negative static effects, in the form of either unnecessary licensing fees or deadweight loss. They also generate deleterious dynamic effects, as firms in the information and communications technology industries must accumulate large defensive arsenals in order to avoid being sued."). *Cf.* Jay Pil Choi & Heiko Gerlach, *Patent Pools, Litigation, and Innovation,* 46 RAND. J. ECON. 499, 499 (2015) ("If patents are sufficiently weak, patent pools with complementary patents reduce social welfare as they charge higher licensing fees and chill subsequent innovation incentives.").

Patent claim scope and claim clarity have been identified as significant concerns for patent quality.[4] Even the basic approach to determining claim meaning has been called into question.[5] Software patents in particular have been criticized for having unduly broad and/or unclear claims employing functional claiming language.[6] Thus, in 2003, a Federal Trade Commission (FTC) Report summarized hearings where "[m]any panelists and participants expressed the view that software and Internet patents are impeding innovation."[7] And in 2004, the authors of the most comprehensive litigation study of patents to that date noted – after determining that litigated (and by hypothesis more valuable) individual patents experienced significantly longer and more complex prosecutions at the PTO – that this "could suggest that the much-maligned PTO is doing a better job than expected in evaluating the patents that really matter, or it could mean that patent examiners are buried in paper by those critical applications."[8]

---

[4] *See, e.g.,* Rai, *supra* note 3, at 512 (identifying as concerns raised by commentators "the grant of patents with scope that exceeds their level of disclosure; and the grant of patents with unclear claim language that fails to provide adequate notice"); Dargaye Churnet, *Patent Claims Revisited,* 11 NW. J. TECH. & INT. PROP. 501, 502 (2013) ("Reading an entire patent application and gaining a thorough understanding of the claims may take weeks. Patent examiners, however, are expected to do so in less than 24 hours. It is no wonder, then, that many have questioned the quality of patents the PTO has issued."); Lee Petherbridge, *On Addressing Patent Quality,* 158 U. PA. L. REV. PENNUMBRA 13 (2009) (Polk Wagner "thoroughly and dispassionately identifies and examines the incentives that patent applicants and the Patent Office have to draft and issue, respectively, large quantities of patents with opaque disclosures and indeterminate claims.") (citing R. Polk Wagner, *Understanding Patent Quality Mechanism,* 157 U. Pa. L. Rev. 2135 (2009)): *cf.* John P. Zimmer, *To Infinity and Beyond: The Problem of Open-Ended Claim Language in the Unpredictable Arts,* 59 S. Car. L. Rev. 865, 867-68 (2008) (arguing for a more stringent enablement examination approach to open-ended claims having uncertain scope at one end of a range of claimed values).

[5] *See, e.g.,* Dan L. Burk & Mark A. Lemley, *Fence Posts or Sign Posts? Rethinking Patent Claim Construction,* 157 U. PA. L. REV. 1743, 1745 (2009) ("Claim construction is sufficiently uncertain that many parties don't settle a case until after the court has construed the claims, because there is no baseline for agreement on what the patent might possibly cover. Even after claim construction, the meaning of the claims remains uncertain, not only because of the very real prospect of reversal on appeal but also because lawyers immediately begin fighting about the meaning of the words used to construe the words of the claims.").

[6] *See, e.g.,* U.S. Government Accountability Office (GAO), Report to Congressional Committees, Intellectual Property: Assessing Factors that Affect Patent Infringement Litigation Could Help Improve Patent Quality, GAO-13-465, at 28-30 (Aug. 2013), available at http://www.uspto.gov/sites/default/files/aia_implementation/GAO-12-465_Final_Report_on_Patent_Litigation.pdf (last visited Nov. 3, 2015) (citing views of various stakeholders). *See generally* Mark A. Lemley, *Software Patents and the Return of Functional Claiming,* 2013 WISC. L. REV. 905.

[7] Federal Trade Commission, To Promote Innovation: The Proper Balance of Competition and Patent Law and Policy ch. 3 at 56 (Oct. 2003), available at http://www.ftc.gov/os/2003/10/innovationrpt.pdf (last visited Nov. 3, 2015).

[8] John R. Allison, et al., *Valuable Patents,* 92 GEO. L.J. 435, 438-39 (2004).

Of course, "patent quality" may have varying meanings, which depend on the user and the context. There are at least five "dimensions" of patent quality on which analysts of the patent system tend to focus, with the first three focused on the patent instrument itself: "(1) a patent's probable validity; (2) clarity of the patent (to different audiences); (3) faithfulness of the patent to the scope of the invention; (4) social utility of the patented invention; and (5) commercial success of the patented invention."[9] It is commonly agreed that only valid patents can be quality patents, but it is frequently disputed as to whether other measures of quality should be considered.[10] Further, some of the measures used in the past by the PTO to assess the rate of granting patents have been criticized, given that the various forms of continuing application practices[11] – including requests for continued examination (RCEs) in the same application[12] – suggest lower grant rates for applications that may ultimately issue, whether with identical or with different claims.[13] Such continued application practices increase the overall demand for examination services, regardless of whether the overall supply of such services is sufficient or reflects "rational ignorance" – i.e., reasonable limits on examination time expenditures from what would otherwise result in improved administrative validity decisions, given the substantial costs of expanding examination resources to address many patents of low innovation utility or low commercial value that will never be licensed or litigated.[14] These (and related)

---

[9] Christi J. Guerrini, *Defining Patent Quality,* 82 FORDHAM L. REV. 3091, 3096 (2014). *See, e.g.,* James E. Malackowski & Jonathan A. Barney, *What is Patent Quality? A Merchant Banc's Perspective,* 43 LES NOUVELLES 123, 124-28 (2008) (distinguishing low quality resulting from examination – validity – errors from low quality resulting from low standards of patentability).
[10] *See* Guerrini, *supra* note 9, at 3098 & n.27, 3099 & nn.28-30 (citing numerous sources).
[11] *See* 35 U.S.C. §§ 120, 121 (2014); 37 C.F.R. § 1.78(a) (2014).
[12] *See* American Inventors Protection Act of 1999, § 4403 (Title IV of the Intellectual Property and Communications Omnibus Reform Act of 1999 (S. 1948)), as enacted by Pub. L. No. 106-113, § 1000(a)(9), Division B, 113 Stat. 1501 (1999); 37 C.F.R. § 1.114 (2014).
[13] *See, e.g.,* Bruce A. Kaser, *Patent Application Recycling: How Continuations Impact Patent Quality & What The USPTO Is Doing About It,* 88 J. PAT. & TRADEMARK OFF. SOC'Y 426, 427-35 (2006). *See generally* Cecil D. Quillen Jr. & Ogden H. Webster, *Continuing Patent Applications and Performance of the U.S. Patent and Trademark Office,* 11 FED. CIR. B.J. 1 (2001); Cecil D. Quillen Jr. & Ogden H. Webster, *Continuing Patent Applications and Performance of the U.S. Patent and Trademark Office - Extended,* 12 FED. CIR. B.J. 35 (2002); Cecil D. Quillen Jr. & Ogden H. Webster, *Continuing Patent Applications and Performance of the U.S. Patent and Trademark Office - Updated,* 15 FED. CIR. B.J. 635 (2005-2006); Cecil D. Quillen Jr. & Ogden H. Webster, *Continuing Patent Applications and Performance of the U.S. Patent and Trademark Office – One More Time,* 18 FED. CIR. B.J. 379, 387-94 (2008-2009); Christopher A. Cotropia, Cecil D. Quillen Jr. & Ogden H. Webster, *Patent Applications and the Performance of the U.S. Patent and Trademark Office,* 23 FED. CIR. B.J. 179 (2013).
[14] *See generally* Mark A. Lemley, *Rational Ignorance at the Patent Office,* 95 NW. U. L. REV. 1495 (2001).

concerns led the PTO in 2007 to adopt rules to restrict continuation practice,[15] which ultimately were withdrawn in 2009 following litigation.[16]

The PTO itself has expressed concerns about patent quality, and has adopted many other initiatives over the last decade to address patent quality concerns.[17] In early 2015, the PTO adopted an "Enhanced Patent Quality Initiative," supervised by a new Deputy Commissioner for Patent Quality, which focuses on three "pillars" of "excellence": (1) work quality; (2) measuring patent quality; and (3) customer service.[18] Some of the measures being considered by the PTO include metrics of quality that go far beyond assessing validity of final determinations, such as processing time, correctness of intermediate actions, and (particularly) assuring clarity of claims and of other aspects of the prosecution record.[19] As noted by the PTO in 2011, its "previous focus on the correctness of actions taken by an examiner in an individual application has been widened to better encompass the entirety of the patent application and examination process."[20] And as noted by the PTO in 2015:

---

[15] *See* USPTO, Changes to Practice for Continuing Applications, Requests for Continued Examination Practice, and Applications Containing Patentably Indistinct Claims, 71 Fed. Reg. 48, 58-61 (Jan. 3, 2006) (proposing restrictions by amending Rules 78 and 114); USPTO, Changes To Practice for Continued Examination Filings, Patent Applications Containing Patentably Indistinct Claims, and Examination of Claims in Patent Applications, 72 Fed. Reg. 46716, 46837-41 (Aug. 21, 2007) (adopting amendments to Rules 78 and 114).

[16] *See* USPTO, Changes to Practice for Continued Examination Filings, Patent Applications Containing Patentably Indistinct Claims, and Examination of Claims in Patent Applications, 74 Fed. Reg. 52686, 52689-91 (Oct. 14, 2009) (withdrawing amendments to rules 78 and 114); Tafas v. Kappos, 586 F.3d 1369, 1370 (Fed. Cir. 2009) (en banc) (dismissing appeal without vacating District Court judgment); Tafas v. Doll, 328 Fed.Appx. 658 (Fed. Cir. 2009) (en banc) (granting en banc rehearing and vacating panel opinion); Tafas v. Doll, 559 F.3d 1345, 1359-63 (Fed. Cir. 2009) (panel decision invalidating PTO rule restricting continuation filings, but upholding PTO rule restricting RCEs); Tafas v. Doll, 541 F. Supp. 2d 805, 814-16 (E.D. Va. 2008) (District Court judgment invalidating amendments restricting both continuations and RCEs).

[17] *See, e.g.,* USPTO, 2010-2015 Strategic Plan (2010), available at http:// www.uspto.gov/about/stratplan/USPTO_2010-2015_Strategic_Plan.pdf (last visited Nov. 3, 2015) (identifying quality improvement as a critical priority); USPTO, The 21st Century Strategic Plan 5 (2003) [hereinafter "21st Century Strategic Plan"], available at http:// www.uspto.gov/web/offices/com/strat21/stratplan_03feb2003.pdf (last visited Nov. 3, 2015) (identifying patent quality as the PTO's "highest priority").

[18] *See* USPTO, Director's Forum: A Blog from USPTO's Leadership (Feb. 4, 2015), at http://www.uspto.gov/blog/director/entry/uspto_launches_enhanced_patent_quality (last visited Nov. 3, 2015); USPTO, Enhanced Patent Quality Initiative, http://www.uspto.gov/patent/initiatives/enhanced-patent-quality-initiative (last visited Nov. 3, 2015).

[19] *See, e.g.,* USPTO, Request for Comments on Enhancing Patent Quality, 80 Fed. Reg. 6475, 6476-80 (Feb. 5, 2015) (hereinafter "PTO Request for Comments 2015"). *See also* Guerrini, *supra* note 9, at 3099 & nn.31-32 (citing various PTO documents).

[20] USPTO, Adoption of Metrics for the Enhancement of Patent Quality Fiscal Year 2011, at

> [t]he USPTO recognizes that, in order for the patent system to fulfill its critical role in promoting innovation, issued patents must not only fully comply with all statutory requirements, but also contain an Official record that is unambiguous and accurate. Such a complete record provides patent boundaries that are clearly defined to the benefit of the patent owner, the courts, third-parties, and the public at large, giving inventors and investors the confidence to take the necessary risks to launch products and start businesses, and the public the benefit of knowing the precise boundaries of an exclusionary right.[21]

Many of the recent debates over the effectiveness of the incentive mechanisms created by patent rights have focused on the central issue of patent quality, but have treated patent examination as a "black box" or have worked backward from the characteristics of issued and litigated patents (or of patents that underwent some form of post-grant administrative reevaluation).[22] More recent scholarship, however, seeks to address more directly patent examination processes relating to patent quality, by looking at examination characteristics in light of the greater availability of such data. For example, Frakes and Wasserman (2015) have used information on application outcomes (including abandonments) to test their hypothesis that under conditions of resource constraints the PTO is more likely to grant applications in technology areas of higher continuation application filings.[23] Earlier, Frakes and Wasserman (2013) found – using PTO data from before and after the PTO acquired fee-setting authority – that the PTO was more likely to grant claims on "technologies with high renewal rates and patents filed by large entities, as the PTO stands to earn the most revenue by granting additional patents of these types."[24] Others, such as Régibeau and

---

http://www.uspto.gov/sites/default/files/patents/init_events/qual_comp_metric.pdf (last visited Nov. 3, 2015).

[21] PTO Request for Comments 2015, *supra* note 19, at 6479.

[22] *See, e.g.,* Allison, et al., *supra* note 8; Shawn P. Miller, *What's the Connection Between Repeat Patent Litigation and Patent Quality? A (Partial) Defense of the Most Litigated Patents,* 16 STAN. TECH. L. REV. 313 (2013); John R. Allison, Mark A. Lemley & Joshua Walker, *Patent Quality and Settlement Among Repeat Patent Litigants,* 99 GEO. L.J. 677 (2011); John R. Allison, Mark A. Lemley & Joshua Walker, *Extreme Value or Trolls on Top? The Characteristics of the Most-Litigated Patents,* 158 U. PA. L. REV. 1 (2009); Stuart J.H. Graham, et al., *Post-Issue Patent "Quality Control": A Comparative Study of US Patent Re-Examinations and European Patent Oppositions,* NBER Working Paper 8807, at 1-5 (2002), available at http://www.nber.org/papers/w8807 (last visited Nov. 3, 2015).

[23] *See* Michael D. Frakes & Melissa F. Wasserman, *Does the U.S. Patent and Trademark Office Grant Too Many Bad Patents?: Evidence From A Quasi-Experiment,* 67 STAN. L. REV. 613 (2015).

[24] Michael D. Frakes & Melissa F. Wasserman, *Does Agency Funding Affect Decisionmaking?: An Empirical Assessment of the PTO's Granting Patterns,* 66 VAND. L. REV. 67, 70 (2013).

Rockett (2010), have looked at the relationship between technology development and invention importance compared to time in examination, using historical data specific to patents on genetically modified crops.[25] Yet others have focused on theoretical modeling of the examination process or of application filing behaviors,[26] and seek to draw empirical support from cross-country comparisons[27] or inferences for patent examination policy.[28]

As noted by Lemley and Sampat (2010), empirical analysis of actual examination practices was first made feasible around 2001 "when the PTO began publishing data on pending applications, and when the Patent Application Information Retrieval ('PAIR') system allowed the public to track the fate of those applications in real time."[29] The availability of patent examination data permitted analysis of grant rates, continuation practices, appeals, and other prosecution events.[30] But the PAIR data do not identify (without further hand-coding) the substantive grounds for the actions or the nature of the changes made to any claims during prosecution.[31]

As noted above, applicant claiming and PTO examination practices have been criticized, focusing on how purportedly "low quality" issued patents are treated in litigation.[32] However, there has been precious little empirical analysis of initial application claiming practices and changes to claims during the examination process. In contrast, numerous studies have looked at judicial

---

[25] *See* Pierre Régibeau and Katharine Rockett, *Innovation Cycles and Learning at the Patent Office: Does the Early Patent Get the Delay?,* 58 J. INDUS. ECON. 222, 222-24 (2010).
[26] *See, e.g.,* Stefano Comino & Clara Graziano, *How Many Patents Does It Take to Signal Innovation Quality?,* 43 INT'L. J. INDUS. ORG. 66, 66-69 (2015) (positing that "true innovators" are forced to patent more intensively in the presence of "bad patents").
[27] *See, e.g.,* Florian Schuett, *Patent Quality and Incentives at the Patent Office,* 44 RAND J. Econ. 313 (2013).
[28] *See, e.g.,* Bernard Caillaud & Anne Duchêne, *Patent Office in Innovation Policy: Nobody's Perfect,* 29 INT'L. J. INDUS. ORG. 242 (2011).
[29] Mark A. Lemley & Bhaven Sampat, *Examining Patent Examination,* 2010 STAN. TECH. L. REV. 2, 2 (2010).
[30] *See, e.g.,* Mark A. Lemley & Bhaven Sampat, *Is the Patent Office a Rubber Stamp?* 58 EMORY L.J. 181, 182-83 (2008).
[31] Note that the coding of actions in PAIR does not itself distinguish claim amendments from other application amendments (such as changes in the specification), although it is possible to read the associated, scanned documents to distinguish these types of amendments.
[32] *See, e.g.,* Petherbridge, *supra* note 4, at 16-18 (discussing various critiques focused on litigation); Alan Marco et al., U.S. Patent and Trademark Office, Patent Litigation and USPTO Trials: Implications for Patent Examination Quality 7-9 (January 2015) (discussing various studies of the relationship of patent quality to litigation and post-grant administrative reviews).

construction of claims at the district court and appellate court level, suggesting changes to how judges perform claim constructions and review thereof.[33]

This paper explores two claim-related metrics for patent scope. Specifically, for each published application and patent in our dataset we calculate:

1. the number of words used in the shortest independent claim (which we call independent claim length, or "ICL");[34] and,
2. the total number of independent claims (which we call "ICC").

We are able to observe the claim language for applications at the date of their pre-grant publication ("PGPub") and for granted patents at the date of issuance. We also calculate changes in ICL and ICC from publication to grant for each published application resulting in a grant ("publication-grant pair"). Moreover, we make the underlying claims data available for public use in order to stimulate more research in the area.[35] We validate ICL and ICC as measures of scope by testing the explanatory power with respect to several patent scope correlates from the previous literature: patent maintenance fee payments, forward citations, the number of technology classes to which the patent was assigned, and patent novelty as defined by Fleming (2001) and Strumsky et al (2012).

This paper presents the first large-scale analysis of patent application and granted patent scope changes during the examination process. Our results reveal several interesting features about the patent examination process. First, we find that applications with narrower claims (in terms of ICL) are more likely to be

---

[33] *See, e.g.,* Shawn P. Miller, *"Fuzzy" Software Patent Boundaries and High Claim Construction Reversal Rates,* 17 STAN. TECH. L. REV. 809 (2015); J. Jonas Anderson & Peter S. Menell, Informal Deference: A Historical, Empirical, and Normative Analysis of Patent Claim Construction, 108 NW. U. L. REV. 1 (2014); Thomas W. Krause & Heather F. Auyang, *What Reversals and Close Cases Reveal About Claim Construction: The Sequel,* 13 J. MARSHALL REV. INTELL. PROP. L. 525 (2014); Thomas W. Krause & Heather F. Auyang, *What Close Cases and Reversals Reveal About Claim Construction,* 12 J. MARSHALL REV. INTELL. PROP. L. 583 (2013); R. Polk Wagner & Lee Petherbridge, *Did* Phillips *Change Anything? Analysis of the Federal Circuit's Claim Construction Jurisprudence,* in INTELLECTUAL PROPERTY AND THE COMMON LAW (S. Balganesh ed. Cambridge U. Press 2011); David L. Schwartz, *Practice Makes Perfect?: An Empirical Study of Claim Construction Reversal Rates in Patent Cases,* 107 MICH. L. REV. 223 (2008); Kimberley A. Moore, Markman *Eight Years Later: Is Claim Construction More Predictable?,* 9 LEWIS & CLARK L. REV. 231 (2005); Joseph Scott Miller & James A. Hilsenteger, *The Proven Key: Roles and Rules for Dictionaries and the Patent Office and the Courts,* 54 AM. U. L. REV. 829 (2005).

[34] We also considered alternative measures for ICL including the average independent claim length and the length of the first independent claim. The results are largely insensitive to the definition of ICL.

[35] The USPTO's Patent Claims Research Dataset will soon be made available at www.uspto.gov/economics.

granted than those with broader claims. Second, the examination process itself tends to narrow the scope of patents. Patent prosecution tends to add 45 words, on average, to the shortest independent claim and tends to reduce the number of independent claims by 0.4 claims. Third, we find that broader applications (in terms of ICL and ICC) tend to have longer pendency times, both for abandoned applications and granted patents. Further, longer pendency periods tend to generate more significant narrowing of the patent between application and grant (in terms of both ICL and ICC). We also find significant variation over time in the breadth of patent applications and patent grants, contrary to conclusions drawn from some earlier analyses that suggested a high level of stability in claim lengths of issued patents over longer periods of time.[36]

This paper is organized as follows. In Section 2, we discuss the relationship between patent scope and our measures of ICL and ICC as proxies for the scope (breadth) of patent applications and granted patents. Section 3 describes the claims data and our resulting sample. Section 4 provides our descriptive analysis. We examine differences in the statistical distributions of ICL and ICC between abandoned applications and applications that are later granted, as well as the evolution of claims during prosecution. We also look at time trends for ICL and ICC, and cross-sectional differences for many types of application characteristics. Lastly, we consider the relationship between patent breadth and pendency. Section 5 provides several validations for using ICL and ICC as a measure of patent scope. Section 6 briefly concludes. We include appendices that provide a detailed description of the public use data sets, and the computer code that generated them.

---

[36] *See, e.g.,* Kristen J. Osenga, *The Shape of Things to Come: What We Can Learn from Patent Claim Length,* 28 SANTA CLARA COMPUTER & HIGH TECH. L.J. 617, 619-37 (2012). *Cf.* Johannes Koenen & Martin Peitz, *Firm Reputation and Incentives to "Milk" Pending Patents,* 43 INT'L J. INDUS. ORG. 18, 18-23 (2015) (discussing equilibrium effects of reputation to seek only meritorious grants and benefits from extending beyond that and from examination errors); Stephen Yelderman, *Improving Patent Quality with Applicant Incentives,* 28 HARV. J.L. & TECH. 77, 78-81 (2014) (arguing that various measures, such as fees, could be used to affect applicant willingness to file overbroad claims).

2 Patent Claims and Patent Prosecution

Our analysis proceeds from the theoretical presumption that the word length of an independent claim and the scope of the claim (equivalently, claim breadth) tend to be *negatively* correlated: adding more words to a claim should generally decrease its scope of potential application.[37] A patent application contains two distinct parts: the specification and the claims.[38] The specification encompasses a written description and background, along with drawings or figures. The claims represent the legal metes and bounds of the invention. Importantly, the specification may not be significantly altered after filing, whereas it is common to amend claims during prosecution.

Typically, applicants have an incentive to file an application with the broadest claims to which they think they are entitled. There is no incentive for the applicant to excessively narrow the claims, ex ante, before the examiner has done her search; that would be the legal equivalent of leaving money on the table. Broader claims translate to a larger set of technologies that the owner can exclude others from using, and making it more difficult for competitors to invent around. During examination, a search may reveal prior art that renders the applicant's claim(s) unpatentable under novelty or obviousness standards. In that case, the examiner rejects the application and the applicant typically amends the claim(s) or abandons the application. In order to circumvent the prior art, claims must be narrowed so that they are not so broad as to overlap with the prior art. Consequently, amendments almost always involve narrowing. Further, this process almost always involves adding words to the claim: modifiers, qualifiers, or other details. The patent prosecution process itself provides it's own support: applicants have no incentive to narrow claims, except to respond to examiners' rejections. Yet, as we show below, the vast majority of independent claims grow longer during prosecution, in response to rejections. Thus, there is at least a

---

[37] *See, e.g.,* Benedikt Szmrecsányi, On Operationalizing Syntactic Complexity, in JOURNÉES INTERNATIONALS D'ANALYSE STATISTIQUE DES DONNÉES TEXTUALLES 1037- 38 (2004) ("determining length in words — to assess syntactic complexity is by all means one that is nearly as accurate as the more sophisticated and cognitively, conceptually, or even psychologically 'more real' methods"). *Cf.* Nicholas van Zeebroeck, Bruno van Pottelsberghe de la Potterie & Dominique Guellec, *Claiming more: the Increased Voluminosity of Patent Applications and its Determinants,* Centre Emile Berneim working Paper No. 06/018, Université Libre de Bruxelles – Solvay Business School, text at n. 21-22 (Mar. 2007) ("As technology becomes more complex, more words may be required to describe and claim it."). *See generally* Thomas Wasow, *Remarks on Grammatical Weight,* in 9 LANGUAGE VARIATION AND CHANGE 81-105 (Cambridge U. Press 1997).

[38] Technically the specification as defined by 35 USC 112 contains the written description and the claims. However, it is common in the industry to refer to the "spec" as distinct from the claims.

correlation between the narrowness of claims and the length of claims, *ceteris paribus*.

Patent prosecution generally involves several rounds of rejection and amendment. A common practice is for applicants to include a very broad independent claim, along with narrower dependent claims. The examiner may reject the independent claim while indicating approval for a dependent claim, which by law must have narrower scope. In that instance, the applicant may "roll up" at least one dependent claim limitation into the original independent claim to form a new, longer and narrower independent claim. For example, claim 1 of U.S. patent application 10/495,059 was modified by the applicant to include most of the language of claim 1 as originally filed, as well as the additional limitations of dependent claims 5 and 6 as originally filed. [39] This additional language narrowed the original independent claim 1 such that, as modified, issued independent claim 1 was allowable over the prior art of record. By legal construction, a dependent claim incorporates the independent claim language and adds a limitation, which requires adding words (e.g., "A device as described in claim 1, such that…"). Thus, by definition a dependent claim roll up will be longer and narrower than the original independent claim.

Where claim language is ambiguous or vague, the examiner may reject the claim under section 112.[40] Clarification by adding words normally narrows the claim scope because it excludes a set of potential embodiments, whether by restricting the meaning of the ambiguous or vague language or by specifying a narrower conception of the things (or relevant properties of things) that the meaning denotes. Note that the approach of treating additional words as narrowing does not necessarily mean that comparing two different claims from different patents on unrelated inventions will permit a general inference that the longer claim implies the narrower scope. Rather, it only indicates that adding words in a particular application tends (all else being equal) to add limitations that reduce or otherwise restrict claim scope. However, comparing word lengths within narrow technology groups may be appropriate. Further, comparing word lengths across patents may enable us to observe general trends over time.

---

[39] See Appendix A for the full text claim language of application 10/495,059 as reflected in U.S. pre-grant publication 20050065799 (published March 24, 2005) and U.S. patent 7,769,690 (issued August 3, 2010).
[40] 35 USC 112.

In particular instances it may be possible to add words to a claim without narrowing its scope, even without regard to specialized claim formats. This is the case when the claim contains a list of possible embodiments, each separated by the word "or." Adding another possible embodiment to the list would add words and potentially add scope.[41] We analyzed our presumptions for robustness against two particular claim formats for which the addition of words may be more likely to expand than to narrow claim scope: claims using the connecting word "or"; and Markush claims that use the words "selected from."

Our observations below focus on claim length and on changes to claim length in particular patent applications during prosecution. Claims as published in the PGPub are a good indication of the claims at filing, because only 8.1 percent of patents have any claim amendments between the date of filing and the date of publication. Further, the change in independent claim length acts as a good proxy for assessing changes to patent scope *during prosecution.* In contrast, in an effort to assess changes to claiming practices over decades, Osenga (2012) looked at average independent and dependent claim length *at grant alone,* using small samples of randomly selected patents. She found that claim length practices had remained surprisingly stable over five decades, notwithstanding significant doctrinal and technological changes.[42] In contrast, we find significant variations in claim length from 1976 to 2014 for granted patents and 2001 to 2014 for published applications.

With respect to the number of independent claims we presume—based on principles of patent prosecution—that more independent claims implies a broader patent scope. That is, adding an independent claim should tend to increase a particular patent's scope,[43] and should never decrease the patent's scope. Claims are subject to the interpretive principle of claim differentiation,[44] and

---

[41] Similarly, a Markush claim provides alternatives as being "selected from the group consisting of A, B, and C" (MPEP 803.02). Adding more elements to the group would add words and increase the scope.
[42] *See* Osenga, *supra* note 36, at 619-22, 632-37.
[43] In comparison, many scholars have used a count of total claims. For example, Allison and Lemley (2000) performed analyses on the total number of claims at grant, based on the assumption that comparative increases *across* unrelated patents in the total number of claims should reflect either increased complexity or increased value of the technology sought to be protected, given that additional claims will normally cost patent applicants additional filing fees and drafting and prosecution costs. With respect to scope, however, the number of independent claims is more accurate, because dependent claims may not be broader than their independent claims.
[44] *See, e.g.,* World Class Tech. Corp. v. Ormco Corp., 769 F.3d 1120, 1125 (Fed. Cir. 2014) ("'The doctrine of claim differentiation creates a presumption that distinct claims, particularly an independent claim and its

consequently a new independent claim should not be entirely subsumed in the scope of another independent claim. Thus, the number of independent claims should be an indicator of the patent's scope, and the change in the number of independent claims during prosecution should reflect the narrowing or broadening of scope due to claim amendments.

Accordingly, we provide our analyses under the presumptions that *ceteris paribus,* a patent's scope is correlated with (1) fewer words in its shortest independent claim (broader claims), and (2) a greater number of independent claims. Therefore, as the length of the shortest independent claim increases, and as the number of independent claims decreases, the scope of patent should narrow. We validate these presumptions in Section 5.

3 Data

We build our claims data sets[45] from publicly available full-text information on pre-grant publications and patent grants. Machine readable claims information is readily available on published patent documents, including the patent grant itself (since 1976), as well as the pre-grant publication (since 2001). Unfortunately, the individual claim amendments during prosecution are only available as image files in the electronic file record of the application (the "image file wrapper"). Further, the bulk data files incorporate the entire text of the patent, not just the claims, and the claims themselves are not individually parsed.

To develop the datasets, we first cleaned and identified the claims section of each bulk file for published applications and patents. Second, we applied an algorithm to the parsed files to identify individual claims as well as the dependency relationships between claims. From the parsed claims text, we measured the length of each claim based on word count. We created data sets at the claim level and summary statistics at the document level.[46]

---

dependent claim, have different scopes.'") (quoting Kraft Foods, Inc. v. Int'l Trading Co., 203 F.3d 1362, 1368 (Fed.Cir.2000)). *See generally* Joshua D. Sarnoff & Edward D. Manzo, *An Introduction to, Premises of, and Problems with Patent Claim Construction,* in CLAIM CONSTRUCTION IN THE FEDERAL CIRCUIT § 0:4(2) (2014 on-line ed. Thompson West). Again, this does not necessarily mean that comparing two unrelated patents with different numbers of claims will indicate that the patent with the larger number of claims has the greater scope.

[45] The USPTO's Patent Claims Research Dataset will soon be made available at www.uspto.gov/economics.

[46] The data sets are provided at www.uspto.gov/economics. More information about the methodology and the structure of the data sets can be found in the appendices.

We provide claim-level data for each pre-grant publication of a U.S. patent application (publication, or PGPub) filed after November 29, 2000 and published before January 1, 2015. Similarly, we provide claim-level data for each patent granted between January 1, 1976 and before January 1, 2015. We also create publication-patent pairs for analysis. Unfortunately, the available sources do not contain parsable claims at the time applications are abandoned, and thus we cannot directly compare claims at the time of abandonment to claims at the time of issue.

For the purposes of this paper we examine document-level (patent-level or PGPub-level) claims statistics. We calculate two document-level statistics of primary interest:

- ICL (independent claim length): the word count of the shortest independent claim in the document. This is often, but not always, the first claim.
- ICC (independent claim count): the number of independent claims in the document.

From the full data set, we constructed publication-patent pairs, for those applications for which we can identify both a PGPub and a granted patent. These pairs enable the observation of changes in an application's claims between publication and grant. For these pairs we define ΔICL and ΔICC, which represent the value of ICL or ICC at grant less the corresponding values at publication. Note that the shortest independent claim at grant may be a different claim number than the shortest independent claim at publication. First, claims may be renumbered at various times during prosecution and the particular forms on which claim amendments are made are not machine readable. Second, amendments may cause the shortest independent claim on the PGPub to grow longer than another independent claim.

Table 1 summarizes our final sample, which represents 3.9 million PGPubs, 4.9 million granted patents, and 2.1 million publication-patent pairs. For publications, the table shows that abandoned applications tend to have broader claims relative to applications that are later granted; abandoned applications have 15 fewer words in their shortest independent claims, at the median. Further, granted patents are narrower at grant than at publication; they tend to gain 45 words in their shortest independent claims, and lose 0.4 independent claim between publication and issuance. We discuss these differences in greater detail in Section 4, below.

4 Analysis

4.1 Comparing pre-grant publications and granted patents

We generated frequency distributions of ICL and ICC (Figures 1 and 2) for all publications and grants, for application years 2001-2014. PGPubs are separated based on whether they resulted in abandonments or grants (pending applications are not shown). These distributions demonstrate how ICL and ICC vary across all applications during prosecution and by disposal type.

From the ICL distributions, it is notable that: (1) applications with narrower claims at the time of publication are more likely to be granted, and (2) granted patents are narrower at the time of grant than at the time of publication. These facts suggest that the prosecution process leads to narrower claims and narrower patents. This is consistent with the practice to roll-up dependent claims into independent claims; the practice would lead to longer independent claims.

We find that the different distributional characteristics between the ICL for PGPubs and for patents indicate that examination not only increased ICL from publication to grant, but also disproportionately decreased the concentration of very short independent claims at grant; i.e., prosecution shifted the distribution to the right (Figure 1). In other words, the prosecution and examination process on average narrows the scope of applications by increasing ICL from 106 words at publication to 156 words at grant for application years 2001 to 2014. Unfortunately, we cannot observe the distribution of abandoned applications at the time of disposal, which would provide insight into separating the relative effects of examination from the initial filing choices by applicants. Nevertheless, the overall distribution of ICL for applications that are later abandoned has the same general shape as for those that are later granted, except that applications that go abandoned have a larger mass of shorter claims for PGPubs. This confirms that allowances are less frequent for applications that have claims of greater scope.

On the other hand, abandoned applications tend to have fewer claims at publication than those that are allowed. With regard to claim counts, the distribution of ICC is right-skewed for PGPubs and for patents at grant. These distributions differ significantly between PGPubs that are later granted (PGPub-grants), PGPubs that are later abandoned (PGPub-abandonments), and granted patents as shown in Figure 2. PGPub-grants have the highest concentration at three independent claims, whereas PGPub-abandonments and patent grants have the highest concentration at one independent claim. The mean of ICC is slightly lower for PGPub-abandonments (2.99 claims) than for PGPub-grants (3.07 claims).

The different distributions for ICC suggest (contrary to what would be expected) that abandoned applications have *narrower* scope compared to applications that result in grants. However, two alternative potential explanations are possible. First, applications with a single independent claim may be more likely to go abandoned when that independent claim is rejected; applications that have more than one independent claim are more likely to continue prosecution if one independent claim is rejected. Second, it is possible that applications that include very broad claims may be more likely to include fewer of such claims or fewer categories of such claims (e.g., product and process claims). The mean number of independent claims for PGPubs, 2.97, is consistent with the maximum number of allowable independent claims per patent application that avoid incurring excess claim fees.

4.2 Trends

Figures 3 to 10 show the trends over time in claims for PGPubs and for grants, which provide some insights into applicant filing behavior as well as potential changes in examination practice. For patents, we can observe claims information from 1976 to 2014. For published applications, we can observe claims data from 2001 to 2014. The figures graph annual arithmetic means for three different cohort aggregations.

First, we define cohorts based on the year of their *final disposition*, whether that was an abandonment or grant, and compare it to information on granted patents using the year of their issue – which we refer to as "cohort comparisons." Second, we compare PGPubs and patent grants based on publication date and the patent issue date, respectively. These "contemporaneous comparisons" provide an indication of how both application and patent claims are changing in a particular year, rather than by looking at the year in which a published application was disposed. Third, we examine publication-patent pairs, which we refer to as "paired comparisons." This permits us to measure trends over time in the *change* in claim language during prosecution, by computing annual means of ΔICL and ΔICC within applications that are granted.[47] Because paired comparisons provide an indication only of the changes to ICL or ICC of a given application, this measure is the least likely to suffer from problems of cross-patent scope comparisons.

A few stylized facts emerge from examining the trends in Figures 3-10:

- There have been significant trends in ICL and ICC over time. Figures 3 and 4 show the trend in ICL and ICC, respectively for granted patents. There is a notable shift towards broader patents from 1984-2004, after which there is a shift towards narrower patents (2004-2014). The trend holds for both ICL (Figure 3) and ICC (Figure 4).

- PGPub-grants tend to be narrower than PGPub-abandonments based on ICL, and granted patents are narrower still (Figures 5 and 7). This holds across the observable range (2001-2014) whether measured by the cohort comparison (Figure 5) or the contemporaneous comparison (Figure 7). It confirms what we observed for the full distribution in Figure 1.

- PGPub-grants and PGPub-abandonments are virtually identical at the means based on ICC, whether measured by the cohort comparison (Figure 6) or the contemporaneous comparison (Figure 8).

- Granted patents are narrower than PGPubs as measured by ICC. However, that difference has been getting smaller over the last decade as measured by the cohort comparison (Figure 6), and virtually disappearing at the mean as measured by the contemporaneous comparison (Figure 8).

- There has been an upward trend in the number of words added to ICL between publication and grant as shown in Figure 9. At the same time, the number of independent claims removed from applications has gone from -0.7 in 2001 to -0.2 in 2014 (Figure 10). These facts are consistent with the cohort comparisons in Figures 5 and 6.

---

[47] The paired comparisons are aggregated based on the date of disposal (issuance). As with cohort comparisons, it is feasible to aggregate by date of publication, which may better highlight applicant filing rather than examination behaviors.

Overall, the trends reinforce the conclusion that the examination process reduces the scope of patent applications. On average, examination adds words to the shortest independent claim and reduces the number of independent claims. Further, the trend over the last decade has been towards narrower patents. ICL for patents has grown significantly since 2004 even while the claim length of published applications has remained more flat (Figure 5). This indicates that the examination process may have become more stringent. The contemporaneous comparisons show that applicant's may have responded to this in recent years by filing narrower claims. However, one should note that the publication time series are censored, because most applications published in 2014 were still pending by the end of 2014. So, the values for recent years include only those applications that received a fast allowance or that abandoned early. (We show below that pendency is correlated with the scope of incoming applications.)

The change in trends beginning around 2004 may correspond to various Patent Office examination quality initiatives adopted following the PTO's 2003 21st Century Strategic Plan and July 2003 legislative hearings on patent quality, including expanded reviews of primary examiners' work, "second-pair-of-eyes" reviews, and quality assurance reviews.[48] USPTO quality initiatives adopted around 2004 and later may have influenced examiner and subsequently applicant behaviors, thus leading to narrower patents over the last decade.

4.3 Relationships between patent scope and examination pendency

The differences in scope between allowed and abandoned applications suggests that there may be differences in patent prosecution based on the scope of the incoming applications, aside from the difference in allowance itself. To investigate this we focus on examination pendency, which is an issue central to applicants, the PTO, and to Congress (Mitra-Kahn, et al, 2013).

We measure pendency by total pendency: the time from filing to final disposal. We use this measure of pendency to determine how ICL or ICC at publication are associated with the time in prosecution. Figures 11 and 12 represent scatter plots of ICL and ICC against total pendency. As one might

---

[48] *See, e.g.,* 21st Century Strategic Plan, *supra* note 17, at 8-9 (discussing measures to improve examiner competency and to enhance quality assurance techniques); National Academy of Public Administration, Report for the U.S. Congress and the USPTO, U.S. Patent and Trademark Office: Transforming to Meet the Challenges of the 21st Century 66-67 (Aug. 2005).

expect (or hope), we find a positive correlation between application scope and total pendency. That is, broad applications tend to have higher pendency. This is shown in Figure 11 by a negative correlation between ICL and total pendency (i.e., fewer words indicates a longer pendency). Correspondingly, Figure 12 shows a positive correlation between the number of independent claims and pendency.

These results are intuitive, particularly for ICC, as examiners should require more time to evaluate each additional independent claim (which by hypothesis should require independent evaluation). The results are identical if we restrict the definition of pendency to examination pendency only (post-first-action pendency).[49]

If broad patents have a longer pendency, a natural question is whether the longer pendency has any impact on the resulting claims at the time of final disposal. With our data we cannot observe claims at the time of abandonment. However, we can investigate the relationship between pendency and the claims at disposal for granted patents. More precisely, we are interested in the relationship between pendency and the *change* in claims for granted patents.

For our publication-patent pairs, we calculate the change in ICL and the change in ICC between publication and grant (as defined in Figures 9 and 10). Our interest is in whether these differences are correlated with pendency. In both cases, we find that greater pendency is associated with more narrowing of the claims during prosecution. Figure 13 shows that there is a positive correlation between pendency and ΔICL (more time is correlated with more words added to the claim). Correspondingly, Figure 14 shows that greater pendency is associated with more independent claims being removed during prosecution.

In short we find that broader applications are subject to longer pendency, and longer pendency is associated with more significant narrowing of claims, both in the length of claims, and the number of claims. This is confirmed in Figures 15 and 16, which show the change in ICL and ICC during prosecution, against the values of ICL and ICC at publication, respectively. The scatter plot (Figure 15) shows a negative correlation between ICL and ΔICL, indicating that broader

---

[49] Post-first-action pendency measures the time from the first action by an examiner to the time of final disposal. This definition of pendency reflects the time under examination at the office, which is impacted by both examiner and applicant behavior.

applications (a low value of ICL at publication) experience greater narrowing (a larger value of ΔICL). Figure 16 also shows a negative correlation between ICC and ΔICC, again indicating that broader applications (a high value of ICC at publication) experience greater narrowing (a more negative value for ΔICC). However, about 25% of applications do not have a change in ICL between publication and grant, and over 50% do not have a change in the number of independent claims.

4.4 Application characteristics

We find that different characteristics of applications can lead to statistically significant differences in measures of scope. However, the general patterns about scope discussed above hold for all groupings: narrower applications tend to be granted, and the prosecution process tends to narrow applications.

Table 2 provides the ICL and ICC for PGPubs grouped by entity size,[50] examination unit (technology center),[51] technology category,[52] and parent application type.[53] The technology center analysis was generally similar to that for the NBER technology categories; thus, we restrict our discussion to the technology centers. For each case, the ICL is higher for PGPub-grants relative to PGPub-abandonments. The number of claims is not substantially different

---

[50] Entity status is based on fee payments at the time of filing. Small and micro entities are combined as a single category relative to large entities.

[51] There are eight technology centers (TCs) used during our period of study, including Biotechnology and Organic Chemistry (1600), Chemical and Materials Engineering (1700), Computer Architecture, Software, and Information Security (2100), Computer Networks, Multiplex Communication, Video Distribution, and the Security (2400), Communications (2600), Semiconductors, Electrical, and Optical Systems and Components (2800), Transportation, Construction, Electronic Commerce, Agriculture, National Security and License & Review (3600), and Mechanical Engineering, Manufacturing and Medical Devices/Processes (3700).

[52] NBER technology categories, as defined by Hall, Jaffe, and Trajtenberg (2001) and Marco, et al (2015) are: Chemical (1), Computers and Communications (2), Drugs and Medical (3), Electrical and Electronics (4), Mechanical (5), and Other (6).

[53] Parent application type or application status relative to the parent. If there was no parent (a first time filing), we identified the application as having "no parent" (not applicable, or USNA). For applications having a parent application, we identified the type of such application. These were divided into applications having a parent that was: a foreign application (Foreign, or FOR); a Patent Cooperation Treaty (PCT) application (which was further subdivided by the designated office of the parent – either PCT-foreign or PCT-US); a prior US non-provisional application (and if so, the relationship to that parent application as discussed below), or a US provisional application (US-provisional, or US-PRO). If the application had a prior US non-provisional application as its parent, we denoted the application's relationship to the parent as a continuation (CON), a divisional (DIV), or a continuation-in-part (CIP) application to a US application.

between PGPub-grants and PGPUB-abandonments across application characteristics, which is consistent with the aggregate results in Figures 6 and 8.

There are some notable characteristics that differ from the means. With regard to technology, applications in *biotechnology* (TC 1600) have the largest difference in ICL between granted applications and abandoned applications: approximately 28 words. This is driven by the very low values for PGPub-abandonments, which are about 12-15 words below the PGPub-abandonment mean of 94 (from Table 1). However, these applications tend to have the most independent claims at filing. Further, biotech is the only examination unit for which PGPub-abandonments have more claims, on average, than PGPub-grants.

TC 3600 (including transportation, construction, e-commerce, and agriculture) tends to have the longest claims (125 words for PGPub-grants and 107 words for PGPub abandonments, relative to the means of 111 and 94, respectively). These applications also tend to have the fewest independent claims. Surprisingly, small and large entities look almost identical at the mean for ICL and ICC. Applications with foreign parents tend to be narrower than average at filing, having higher ICL and lower ICC. The broadest patents at filing tend to be those with US provisional parents.

Table 3 provides the ICL and ICC for publication-patent pairs, by groups based on application characteristics. By comparing claims at publication to claims at grant, we can identify the average change in claims during patent prosecution. The publication values in Table 3 match those found in Table 2 for granted applications. There are several interesting facts that emerge from Table 3. Most notably, for each group applications are narrowed between publication and grant, in terms of ICL and ICC. We also see interesting differences between application types.

Small and large entity applications tend to be similar at filing, but small entities experience greater narrowing during prosecution, leading to 5 more words and 0.25 fewer claims at issue relative to large entities. Biotech applications again stand out relative to other examination groups: they are not significantly narrowed with respect to ICL (only 11 words), but they lose an average of 1.5 independent claims during prosecution. This is likely based on nature of the invention and the terminological (nomenclature) conventions for how certain types of inventions (particularly chemical products) are claimed. Computer-related patents—on the other hand—are more subject to increases in ICL than to decreases in ICC.

Parent types reveal some interesting facets about application sources. Both foreign and PCT-foreign sourced applications are filed with the longest independent claims (123 and 120 words, respectively), yet they are among the highest with respect to changes in ICL during prosecution (44 and 48 words, respectively). This means that the ICL of the resulting patents has an ICL of more than 165 words—more than 15 words higher than the next highest parent type. This is perhaps surprising because foreign applications may already have been through an examination process in the home jurisdiction, and thus may be "pre-narrowed." Further, the other application types with significant narrowing during prosecution are those with no parent and those with provisional parents (adding 51 and 48 words, respectively). Yet, those applications tend to be filed with the broadest claims (99 words at the mean). One might expect that "new" applications, with no previous non-provisional filings, would be filed with broad claims. Thus, it is surprising that foreign applications and new applications are narrowed by similar amounts.

Continuations and divisionals of regular US applications had the largest ICL at publication and had the smallest ΔICL among U.S. applications (+29.6 and +31.5 words, respectively). It is likely that continuations tend to be narrower when filed and require fewer changes from application to grant than other applications, because continuations have already gone through at least one round of US prosecution before the continuation was filed.

5 Validation

To validate our ICL and ICC measures of patent scope, we employ several statistical tests to compare these measures with post-grant outcomes and other variables traditionally correlated with patent scope, as shown in Tables 4a and 4b. The tests extend the previous literature and examine the impact of patent scope—based on ICL and ICC—on (1) forward citations, (2) the number of Cooperative Patent Classification (CPCs)[54] subclasses to which the patent was assigned, (3) patent maintenance, and (4) a novelty measure based on whether the granted patent was issued in a "new" US patent classification subclass. We use a variant of the validation method from Lerner (1994), which analyzes the relationship between a proxy for patent scope—the number of 4-digit International Patent

[54] The CPC classification system was jointly developed by the USPTO and European Patent Office (EPO) and is a descendent of the IPC classification. For more information on t the CPC classification system, please visit http://www.cooperativepatentclassification.org/.

Classifications (IPCs) a patent was assigned—to the number of forward citations assigned to a given patent and to the incidence of litigation. We present evidence that our measures of patent scope explain traditional scope proxies in a way consistent with Lerner (1994). We also discuss how our measures relate to the results from the USPTO's *Patent Litigation and USPTO Trials: Implications for Patent Examination Quality*, which examined the relationship between the incidence of litigation and ICL and ICC at grant. Our results show that the relationship between our measures of patent scope and the outcome variables above is consistent with other validation tests of patent scope in the literature.

Lerner (1994) found that a proxy for patent scope, the number of 4-digit IPCs, was positively and significantly related to the number of forward citations a patent receives. An increase in the number of 4-digit IPCs assigned to a patent reflects an increasing number of distinct technologies incorporated into the invention, which can be interpreted as increasing broadness of a given patent. Lerner (1994) used a simple Poisson regression to examine the relationship between the dependent variable, a count of forward citations for a given patent, and the independent variable, the number of IPCs. He also controlled for the time since grant, to account for varying exposure time among patents in his sample of biotechnology firms. The results show that as the number of IPCs increases, the number of forward citations in a given patent increases as well.

We extend Lerner's analysis to include maintenance rates and forward citations (following van Zeebroeck, 2011) and a novelty indicator based on Fleming (2001) and Strumsky et al (2012). Further, we include the number of CPCs as a dependent variable. More precisely, our dependent variables include two count variables and two binary indicators are defined as follows:

- *Forward citations*. A count of the number of citations received by the patent within three years of the issue date.
- *Number of subclasses*. A count of the number of unique CPC subclasses (4-digit) assigned to the patent.
- *Fully maintained*. A binary indicator of whether the patent was maintained to its maximum statutory term (paying the requisite fees at 3.5, 7.5, and 11.5 years after grant).
- *New subclass*. A binary indicator of whether the patent was classified in a "new" subclass, according the US Patent Classification system. "New" is defined as being within 12 months of the first use of the subclass (see Fleming, 2001 and Strumsky et al, 2012).

We expect each indicator to be positively correlated with patent scope, along the lines of Lerner's argument and the findings in van Zeebroeck (2011). Forward citations have long been used by economists as a correlate of patent value and scope (van Zeebroeck, 2011). Patent maintenance is closely related to the concepts of patent value and patent scope. According to Bessen (2008), "[t]he implicit value of a patent is revealed when its owner pays a renewal fee, implying that the patent is worth more than the fee required to keep it in force." Broader patents, *ceteris paribus*, have wider applicability than a narrower patent representing similar underlying technologies, and should therefore be more valuable.

First-movers in a technology space have the opportunity to patent fundamental inventions. These seminal patents can be expected to have broader scope than the incremental inventions that follow (Strumsky et al, 2012). Until the conversion to CPC in 2015, the USPTO regularly re-evaluated the US Patent Classification system. New classes or subclasses were created retrospectively based on whether a significant volume of the "new" inventions had been filed, so that a new subclass would make routing and search easier. For instance, class 977 (nanotechnology) was created in August, 2004.[55] The classification effort had the purpose to "[f]acilitate the searching of prior art related to Nanotechnology," and to "[f]unction as a collection of issued U.S. patents and published pre-grant patent applications relating to Nanotechnology across the technology centers." As such, it added the cross-reference classification to already issued U.S. patents. The earliest patent in class 977 is US patent 4,107,288, issued in 1978,[56] a full 26 years before the creation of the class. These early patents represent the first patents identified by the USPTO that are relevant for prior art search in the technology.

We expect ICL and ICC to be negatively and positively correlated, respectively, to our patent scope indicators. To confirm this hypothesis, we run Poisson regressions with *forward citations* and the *number of subclasses* as dependent variables, and a linear probability model (ordinary least squares) for the *fully maintained* and *new subclass* indicators. We also include year fixed effects and US Patent Classification fixed effects, to control for differences in

---

[55] See USPTO memo dated August 25, 2004 at http://www.uspto.gov/web/offices/pac/dapp/opla/documents/nanotechdig.pdf (accessed August 9, 2016).
[56]

claim length and citation behavior by applicants between classes and across years. Tables 4a and 4b present the results of the regressions.

For each of the four dependent variables we estimate three models based on the explanatory variables: ICL, ICC, and ICL and ICC together. Each model includes year fixed effects and US Patent Class fixed effects. Our expectation is that ICL will have a negative coefficient and ICC will have a positive coefficient, both of which correspond to a positive correlation between our scope measures and the dependent variables of value and scope.

For ICC, all coefficients are positive and statistically significant at the 1 percent significance level for all specifications. For ICL all coefficients are negative and statistically significant at the 0.1 percent significance level with three exceptions. The coefficient is positive for *forward citations* when combined with ICC (Model 6), and it is negative but not statistically significant for the *new subclass* specifications. The robustness of the results across specifications implies that ICL and ICC are useful measures of patent scope. Because the models that include both measures tend to have the expected signs further imply that ICC and ICL represent different aspects of patent scope.

As further evidence that ICL and ICC represent patent scope, we rely on results from Marco et al (2015). There, the authors find that patent scope—as measured by average independent claim length and independent claim count—is correlated with the incidence of patent litigation. Lanjouw and Schankerman (2001) explain why patent breadth should be positively correlated with litigation. Thus, the result provides more evidence that ICL and ICC are indicators of patent scope.

6 Conclusion

This paper presents the first large-scale analysis of patent claim language as it applies to patent scope. We define two document-level measurements of scope that should be useful to researchers interested in patent value and patent quality: independent claim length (ICL) and independent claim count (ICC). Our hypotheses that ICL is negatively correlated with patent scope and ICC is positively correlated with patent scope are born out in several ways. First, we find that the narrowing process that occurs during examination tends to add words to the shortest independent claim and tends to remove independent claims, leading to greater ICL and lower ICC. Second, our formal validation exercise shows that ICL and ICC independently explain patent maintenance, forward citations, the breadth of patent classes, and—to a lesser extent—novelty.

As shown above, using very simple claim length and claim count metrics to model application and patent scope can provide useful information about patent prosecution. For instance, we show that narrower applications tend to have shorter examination times, and that longer examination times lead to more significant narrowing of the original application claims. As a measure of scope, we expect ICL and ICC to be the most meaningful for intra-application comparisons and intra-technology comparisons. However, we believe that the results presented here provide ample evidence that claim text can be usefully exploited by researchers to measure patent scope.

Our continuing research agenda includes more in-depth analysis into the examination process, as well as exploring how natural language processing techniques can be applied to claim text. By making these data widely available we hope to stimulate more research into the usefulness of analyzing claim text in order to understand patent scope and its relationship to examination quality and patent quality.

References

"Adoption of Metrics for the Enhancement of Patent Quality Fiscal Year 2011." *U.S. Patent and Trademark Office* (blog). Accessed November 3, 2015. http://www.uspto.gov/sites/default/files/patents/init_events/qual_comp_metric.pdf.

Allison, John R., and Ronald J. Mann. "The Disputed Quality of Software Patents." *Washington University Law Review* 85 (2007): 297.

Allison, John R., Mark A. Lemley, Kimberly A. Moore, and R. Derek Trunkey. "Valuable Patents." *Georgetown Law Review* 92 (2004): 435.

Allison, John R., J. H. Walker, and Mark A. Lemley. "Patent Quality and Settlement among Repeat Patent Litigants." *Georgetown Law Review* 99 (2011): 677.

Allison, John R., Mark A. Lemley, and Joshua Walker. "Extreme Value or Trolls on Top? The Characteristics of the Most-Litigated Patents." *University of Pennsylvania Law Review* 158 (2009): 1.

Allison, John R., and Mark A. Lemley. "Who's Patenting What? An Empirical Exploration of Patent Prosecution." *Vanderbilt Law Review* 53 (2000): 2107.

Anderson, J. Jonas and Peter S. Menell, “Informal Deference: A Historical, Empriical, and Normative Analysis of Patent Claim Construction,” *Nw. U. L. Rev.* 108 (2014): 1.

Burk, Dan L. and Mark A. Lemley, “Fence Posts or Sign Posts? Rethinking Patent Claim Construction,” *U. Pa. L. Rev. 157* (2009): 1743.

Caillaud, Bernard, and Anne Duchêne. "Patent Office in Innovation Policy: Nobody's Perfect." *International Journal of Industrial Organization* 29, no. 2 (2011): 242.

Choi, Jay Pil. "Patent Pools And Cross-Licensing In The Shadow Of Patent Litigation." *International Economic Review* 51, no. 2 (2010): 441.

Dargaye Churnet, “Patent Claims Revisited,” *Nw. J. Tech. & Int. Prop.* 11 (2013): 501.

Collins, Kevin E. "The Reach of Claim Scope Into After-Arising Technology: On Thing Construction and the Meaning of Meaning." *Connecticut Law Review* 41 (2008): 493.

Comino, Stefano, and Clara Graziano. "How Many Patents Does It Take to Signal Innovation Quality?" *International Journal of Industrial Organization* 43 (2015): 66.

Cotropia, Christopher A., Cecil D. Quillen, Jr., and Ogden H. Webster. "Patent Applications and the Performance of the U.S. Patent and Trademark Office." *The Federal Circuit Bar Journal* 23 (2013): 179.

Federal Trade Commission. *To Promote Innovation: The Proper Balance of Competition and Patent Law and Policy.* October 2013. Accessed November 3, 2015. http://www.ftc.gov/os/2003/10/innovationrpt.pdf.

Fleming, Lee. “Recombinant Uncertainty in Technological Search.” *Management Science* Volume 47, Issue 1, pp. 117-132 (2001). http://dx.doi.org/10.1287/mnsc.47.1.117.10671.

Frakes, Michael D., and Melissa F. Wasserman. "Does the U.S. Patent and Trademark Office Grant Too Many Bad Patents?: Evidence From A Quasi-Experiment." *Stanford Law Review* 67 (2015): 613.

Frakes, Michael, and Melissa F. Wasserman. "Does Agency Funding Affect Decisionmaking?: An Empirical Assessment of the PTO's Granting Patterns." *Vanderbilt Law Review* 66 (2013): 67.

Graham, Stuart J., Bronwyn Hall, Dietmar Harhoff, and David Mowery. "Post-Issue Patent "Quality Control": A Comparative Study of US Patent Re-examinations and European Patent Oppositions." *NBER Working Paper 8807*, 2002. Accessed November 3, 2015. http://www.nber.org/papers/w8807.

Guerrini, Christi J. "Defining Patent Quality." *Fordham Law Review* 82 (2014): 3091.

Hall, Bronwyn, Adam Jaffe, and Manuel Trajtenberg. "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." *NBER Working Paper 8498*, October 2001.

Kaser, Bruce A. "Patent Application Recycling: How Continuations Impact Patent Quality & What The USPTO Is Doing About It." *Journal of the Patent and Trademark Office Society* 88 (2006): 426.

Koenen, Johannes, and Martin Peitz. "Firm Reputation and Incentives to 'Milk' Pending Patents." *International Journal of Industrial Organization* 43 (2015): 18.

Krause, Thomas W. & Heather F. Auyang, "What Close Cases and Reversals Reveal About Claim Construction," *J. Marshall Rev. Intell. Prop. L.* 12 (2013): 583.

Krause, Thomas W. & Heather F. Auyang, "What Reversals and Close Cases Reveal About Claim Construction: The Sequel," *J. Marshall Rev. Intell. Prop. L.* 13 (2014): 525.

Lemley, Mark A. "Software Patents and the Return of Functional Claiming." *Wisconsin Law Review* 2013 (2013): 905.

Lemley, Mark A. "Rational Ignorance at the Patent Office." *Northwestern University Law Review* 95 (2001): 1495.

Lemley, Mark A., and Bhaven Sampat. "Examining Patent Examination." *Stanford Technology Law Review* 2010 (2010): 2.

Lemley, Mark A., and Bhaven Sampat. "Is the Patent Office a Rubber Stamp?"

Emory Law Journal 58 (2008): 181.

Lerner, Joshua, "The Importance of Patent Scope: An Empirical Analysis," *RAND J. Econ.* 25 (1994): 319.

Malackowski, James E., and Jonathan A. Barney. "What Is Patent Quality? A Merchant Banc's Perspective." *LES Nouvelles* 43 (2008): 123.

Marco, Alan C., Michael Carley, Steven Jackson, and Amanda F. Myers. "The USPTO Historical Patent Data Files: Two Centuries of Innovation." *U.S. Patent and Trademark Office Working Paper*, 2015.

Marco, Alan C., Richard D. Miller, Kathleen Kahler Fonda, Pinchus M. Laufer, Paul Dzierzynski, and Martin Rater, "U.S. Patent and Trademark Office, Patent Litigation and USPTO Trials: Implications for Patent Examination Quality" (January 2015). Accessed November 3, 2015. http://www.uspto.gov/sites/default/files/documents/Patent%20litigation%20and%20USPTO%20trials%2020150130.pdf

Miller, Shawn P. "What's the Connection Between Repeat Patent Litigaiton and Patent Quality? A (Partial) Defense of the Most Litigated Patents." *Stanford Technology Law Review* 16 (2013): 313.

Mitra-Kahn, Benjamin, Alan C. Marco, Michael Carley, Paul D'Agostino, Peter Evans, Carl Frey, and Nadiya Sultan, "Patent Backlogs, inventories, and pendency: an International Framework" (June 2013). Accessed August 16, 2016. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/311239/ipresearch-uspatlog-201306.pdf

Moore, Kimberley A., "*Markman* Eight Years Later: Is Claim Construction More Predictable?," *Lewis & Clark L. Rev.* 9 (2005): 231.

Mossinghoff, Gerald J., and Vivian S. Kuo. "Post-Grant Review of Patents: Enhancing the Quality of the Fuel of Interest." *IDEA: The Journal of Law and Technology* 43 (2002): 83.

National Academy of Public Administration, "Report for the U.S. Congress and the USPTO, U.S. Patent and Trademark Office: Transforming to Meet the Challenges of the 21st Century" (Aug. 2005).

Osenga, Kristen J. "The Shape of Things to Come: What We Can Learn from

Patent Claim Length." *Santa Clara High Technology Law Journal* 28 (2012): 617.

Petherbridge, Lee, "On Addressing Patent Quality," *U. Pa. L. Rev. PENNumbra* 158 (2009): 13.

Quillen, Cecil D., Jr., and Ogden H. Webster. "Continuing Patent Applications and Performance of the U.S. Patent and Trademark Office." *The Federal Circuit Bar Journal* 11 (2001): 1.

Quillen, Cecil D., Jr., and Ogden H. Webster. "Continuing Patent Applications and Performance of the U.S. Patent and Trademark Office - Extended." *The Federal Circuit Bar Journal* 12 (2002): 35.

Quillen, Cecil D., Jr., and Ogden H. Webster. "Continuing Patent Applications and Performance of the U.S. Patent and Trademark Office - Updated." *The Federal Circuit Bar Journal* 15 (2006): 635.

Quillen, Cecil D., Jr., and Ogden H. Webster. "Continuing Patent Applications and Performance of the U.S. Patent and Trademark Office – One More Time." *The Federal Circuit Bar Journal* 18 (2009): 379.

Rai, Arti K. "Improving (Software) Patent Quality Through The Administrative Process." *Houston Law Review* 51 (2013): 503.

Régibeau, Pierre, and Katharine Rockett. "Innovation Cycles And Learning At The Patent Office: Does The Early Patent Get The Delay?" *The Journal of Industrial Economics* 58 (2010): 222.

Sarnoff, Joshua D. and Edward D. Manzo, "An Introduction to, Premises of, and Problems with Patent Claim Construction," in *Claim Construction in the Federal Circuit* § 0:4(2) (2014 on-line ed. Thompson West).

Schuett, Florian. "Patent Quality and Incentives at the Patent Office." *The RAND Journal of Economics* 44, no. 2 (2013): 313-36.

Schwartz, David L., "Practice Makes Perfect?: An Empirical Study of Claim Construction Reversal Rates in Patent Cases," *Mich. L. Rev.* 107 (2008): 223

Schwartz, David L. and Jay P. Kesan, "Analyzing the Role of Non-Practicing Entities in the Patent System," *Cornell L. Rev.* 99 (2014): 425.

Strumsky, Deborah, José Lobo, and Sander Van der Leeuw, "Using patent technology codes to study technological change." *Economics of Innovation and New Technology* 21 (3) pp. 267-286 (2012). http://www.doi.org/10.1080/10438599.2011.578709.

Szmrecsányi, Benedikt. "On Operationalizing Syntactic Complexity." *JOURNÉES INTERNATIONALS D'ANALYSE STATISTIQUE DES DONNÉES TEXTUALLES*, 2004, 1037-038.

U.S. Government Accountability Office,. *Report to Congressional Committees, Intellectual Property: Assessing Factors That Affect Patent Infringement Litigation Could Help Improve Patent Quality, GAO-13-465,* August 2013. Accessed November 3, 2015. http://www.uspto.gov/sites/default/files/aia_implementation/GAO-12-465_Final_Report_on_Patent_Litigation.pdf.

van Zeebroeck, Nicolas, Bruno van Pottelsberghe de la Potterie, and Dominique Guellec. "Claiming More: The Increased Voluminosity of Patent Applications and Its Determinants." *Research Policy* 38, no. 6 (2009): 1006-020.

van Zeebroeck, Nicolas. "The puzzle of patent value indicators," *Economics of Innovation and New Technology*, Volume 20, Issue 1 (2011). http://dx.doi.org/10.1080/10438590903038256.

Wagner, R. Polk, "Understanding Patent Quality Mechanism," *U. Pa. L. Rev.* 157 (2009): 2135.

Wagner, R. Polk & Lee Petherbridge, "Did *Phillips* Change Anything? Analysis of the Federal Circuit's Claim Construction Jurisprudence," in *Intellectual Property and the Common Law* (S. Balganesh ed. Cambridge U. Press 2011).

Wasow, Thomas. "Remarks on Grammatical Weight," in *Language Variation and Change* (Cambridge U. Press 1997): 81.

Yelderman, Stephen. "Improving Patent Quality with Applicant Incentives." *Harvard Journal of Law & Technology* 28 (2014): 77.

Zimmer, John P., "To Infinity and Beyond: The Problem of Open-Ended Claim Language in the Unpredictable Arts," *S. Car. L. Rev.* 59 (2008): 865.

Tables and Figures

Table 1. Distributional statistics for pre-grant publications (2001-2014) and patent grants (1976-2014)

		ICL				ICC			
	Frequency	**Mean**	**P25**	**P50**	**P75**	**Mean**	**P25**	**P50**	**P75**
Publications (2001-2014)									
Later Abandoned	1089427	94.2	46	75	115	3.03	1	2	3
Later Granted	2113273	111.4	58	90	137	3.08	2	3	4
Pending*	790019	107.1	59	90	133	2.73	2	3	3
All	3992719	105.8	54	86	130	2.99	2	3	3
Grants									
At Publication	2113273	111.4	58	90	137	3.08	2	3	4
At Grant (previously published)	2113273	155.9	93	136	195	2.70	1	2	3
At Grant (not previously published)	634235	141.0	82	121	176	3.12	2	3	4
At Grant (1976-2000)	2203409	155.6	92	137	198	2.43	1	2	3

* *As of December 31, 2016*

Table 2. Applications at publication by application characteristics (2001-2014)

	IC Length			IC Count		
	Later Issued	**Later Abandoned**	**Difference**	**Later Issued**	**Later Abandoned**	**Difference**
Small entity status						
Large	111.03	94.07	16.96	3.09	3.08	0.02
Small or Micro	112.24	94.03	18.21	3.03	2.94	0.08
Technology Center						
1600 Biotech, Organic Chem	110.23	81.79	28.44	3.75	3.98	-0.23
1700 Chem & Mat Engineering	97.75	84.37	13.38	2.74	2.66	0.08
2100 Comp Architecture	107.80	95.68	12.11	3.60	3.50	0.10
2400 Comp Networks	107.73	95.60	12.13	3.59	3.50	0.10
2600 Communications	109.21	95.68	13.53	3.47	3.19	0.27
2800 Semiconductors, Electrical	110.99	95.65	15.35	2.89	2.62	0.27
3600 Trans, Constr, E-Comm, Ag	125.45	106.86	18.60	2.80	2.78	0.02
3700 Mech, Mfg, Products	117.04	99.93	17.11	2.84	2.67	0.17
NBER category						
1 - Chemicals	102.07	95.20	6.87	2.91	2.84	0.07
2 - Comp & Comm	109.71	97.43	12.28	3.43	3.36	0.07
3 - Drugs & Medical	107.28	78.80	28.47	3.54	3.72	-0.18
4 - Electrical	110.82	95.41	15.41	2.85	2.63	0.23
5 - Mechanical	123.43	105.25	18.18	2.66	2.49	0.17
6 - Others	114.17	95.72	18.45	2.83	2.58	0.25
Parent application type						
Foreign	122.95	101.84	21.10	2.69	2.66	0.03
PCT - foreign	119.92	97.11	22.81	2.66	2.81	-0.15
PCT - US	109.40	87.94	21.46	3.39	3.60	-0.21
CIP of US app	107.15	95.77	11.37	3.58	3.51	0.07
CON of US app	112.09	94.85	17.24	3.27	3.43	-0.17
DIV of US app	108.99	94.25	14.74	3.16	3.12	0.04
No parent	98.73	91.91	6.82	3.32	2.97	0.36
US provisional	98.83	83.19	15.64	3.67	3.44	0.23

IC Length is defined as the length of an application's shortest Independent Claim

Table 3. Publication-Patent Pairs (2001-2014)

	IC Length			IC Count		
	At publication	At issuance	Difference	At publication	At issuance	Difference
Small entity status						
Large	111.03	154.97	43.94	3.09	2.75	-0.34
Small or Micro	112.24	160.03	47.79	3.03	2.50	-0.53
Technology Center						
1600 Biotech, Organic Chem	110.23	121.39	11.16	3.75	2.27	-1.48
1700 Chem & Mat Engineering	97.75	138.64	40.88	2.74	2.21	-0.54
2100 Comp Architecture	107.80	175.53	67.73	3.60	3.32	-0.28
2400 Comp Networks	107.73	183.72	75.99	3.59	3.34	-0.25
2600 Communications	109.21	159.73	50.53	3.47	3.26	-0.21
2800 Semiconductors, Electrical	110.99	145.36	34.36	2.89	2.66	-0.23
3600 Trans, Constr, E-Comm, Ag	125.45	179.36	53.90	2.80	2.58	-0.22
3700 Mech, Mfg, Products	117.04	168.23	51.18	2.84	2.53	-0.31
NBER category						
1 - Chemicals	102.07	135.32	33.25	2.91	2.20	-0.71
2 - Comp & Comm	109.71	165.59	55.88	3.43	3.19	-0.24
3 - Drugs & Medical	107.28	138.32	31.04	3.54	2.47	-1.07
4 - Electrical	110.82	148.19	37.37	2.85	2.60	-0.25
5 - Mechanical	123.43	167.59	44.16	2.66	2.44	-0.22
6 - Others	114.17	165.71	51.54	2.83	2.53	-0.30
Parent application type						
Foreign	122.95	166.83	43.88	2.69	2.49	-0.20
PCT - foreign	119.92	168.06	48.14	2.66	2.22	-0.44
PCT - US	109.40	150.78	41.38	3.39	2.58	-0.81
CIP of US app	107.15	149.30	42.15	3.58	3.02	-0.56
CON of US app	112.09	141.68	29.59	3.27	2.89	-0.38
DIV of US app	108.99	140.48	31.49	3.16	2.37	-0.79
No parent	98.73	150.02	51.30	3.32	3.03	-0.29
US provisional	98.83	146.68	47.85	3.67	3.04	-0.63

Note: 10,311 of 2,113,273 publication-patent pairs were lost due to data availability issues for application characteristics. IC Length is defined as the length of an application's shortest Independent Claim

Table 4a. Validation Results

	Fully Maintained			New Subclass		
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
ICC at Grant	-0.0229***		-0.0169***	-1.87e-05		-7.28e-06
	(0.000435)		(0.000444)	(2.01e-05)		(2.05e-05)
ICC at Publication		0.0127***	0.0113***		2.81e-05**	2.75e-05**
		(0.000169)	(0.000172)		(9.61e-06)	(9.80e-06)
Constant	0.518***	0.449***	0.479***	0.0230***	0.0229***	0.0229***
	(0.00167)	(0.00159)	(0.00177)	(0.000170)	(0.000169)	(0.000173)
Observations	1,448,038	1,448,177	1,448,038	4,937,731	4,937,997	4,937,731
R-squared	0.017	0.019	0.020	0.006	0.006	0.006
Model Type		OLS			OLS	
Years	1994-2004	1994-2004	1994-2004	1976-2014	1976-2014	1976-2014

Standard errors in parentheses. The models above include disposal year and USPC fixed effects.
*** p<0.001, ** p<0.01, * p<0.05

Table 4b. Validation Results

	Forward Citations			Number of 4-digit CPCs		
VARIABLES	(7)	(8)	(9)	(10)	(11)	(12)
ICC at Grant	-0.00297***		0.0183***	-0.0219***		-0.0189***
	(0.000667)		(0.000608)	(0.000372)		(0.000377)
ICC at Publication		0.0372***	0.0379***		0.00831***	0.00669***
		(0.000157)	(0.000157)		(0.000158)	(0.000163)
Observations	2,068,106	2,068,231	2,068,106	4,666,314	4,666,557	4,666,314
Model Type		Poisson			Poisson	
Years	2000-2011	2000-2011	2000-2011	1976-2014	1976-2014	1976-2014

Standard errors in parentheses. The models above include disposal year and USPC fixed effects.
*** p<0.001, ** p<0.01, * p<0.05

Figure 1

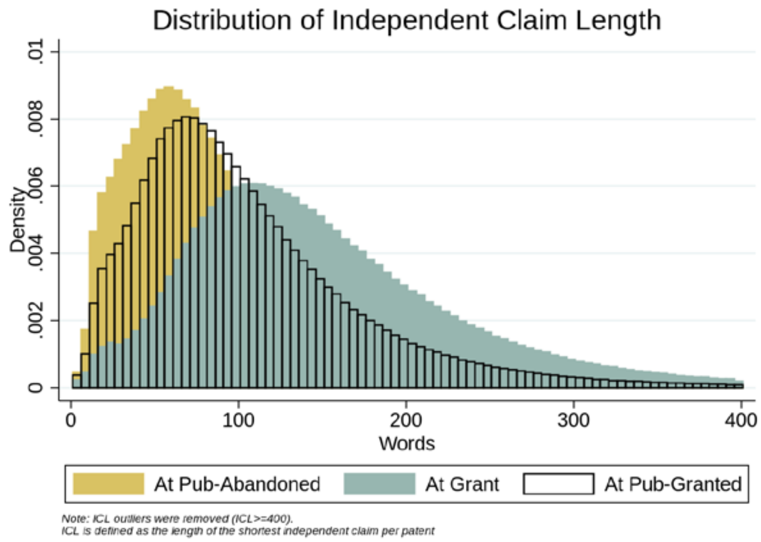


Figure 2

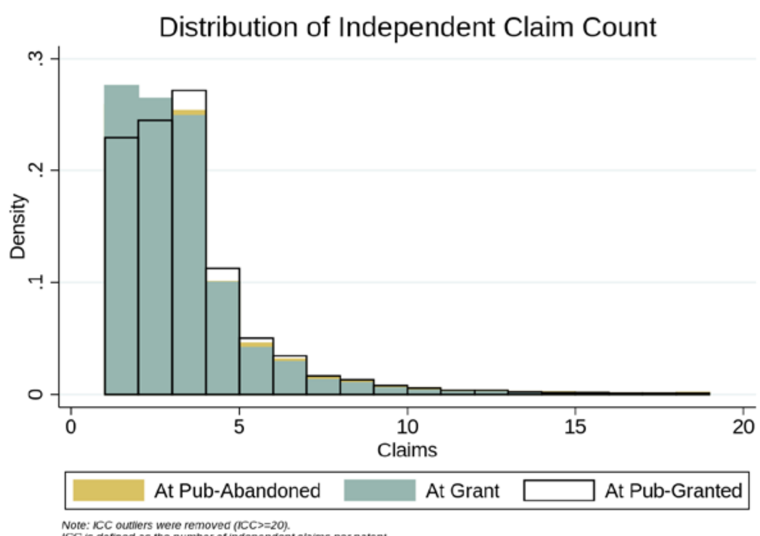


Figure 3

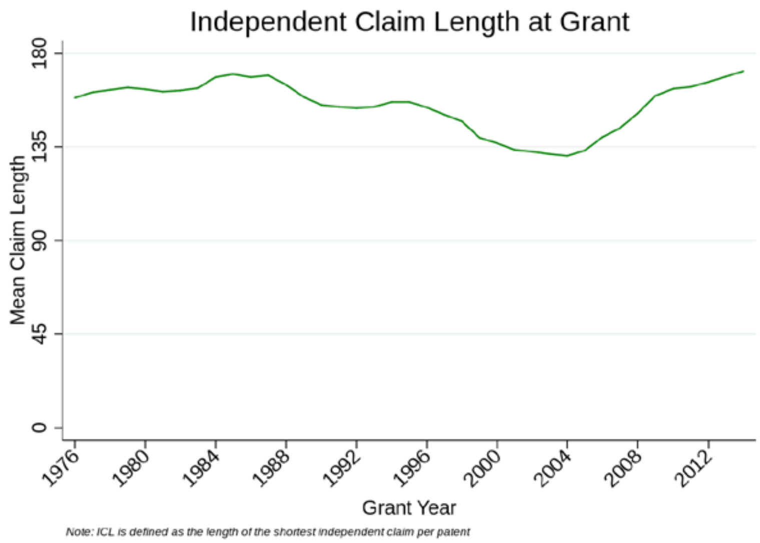


Figure 4

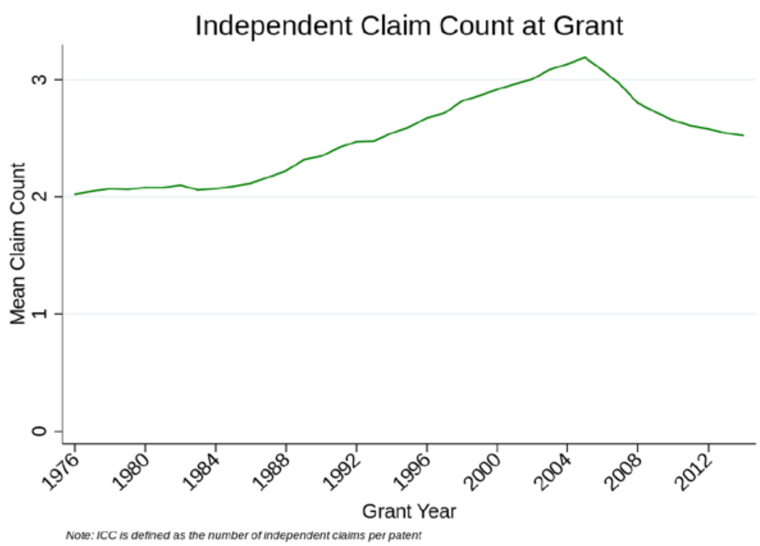


Figure 5

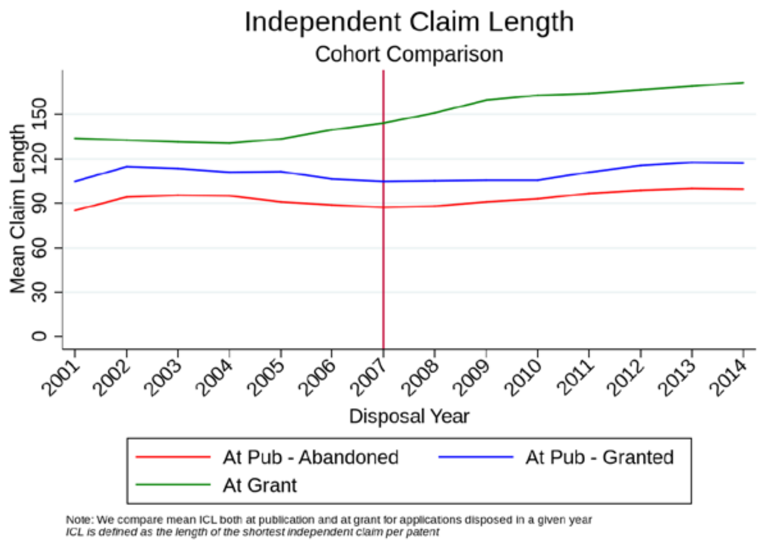


Figure 6

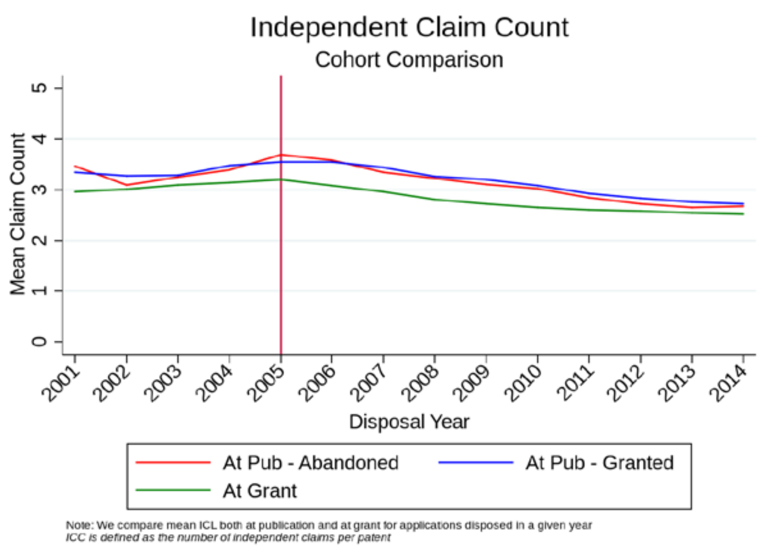


Figure 7

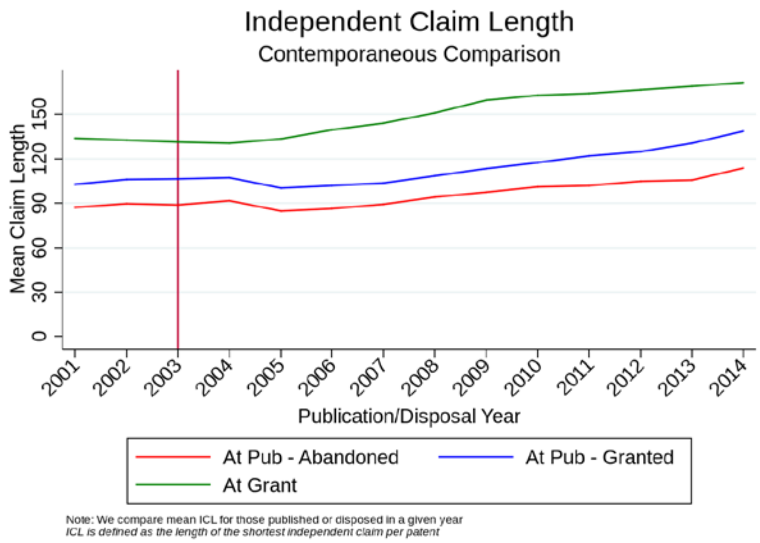


Figure 8

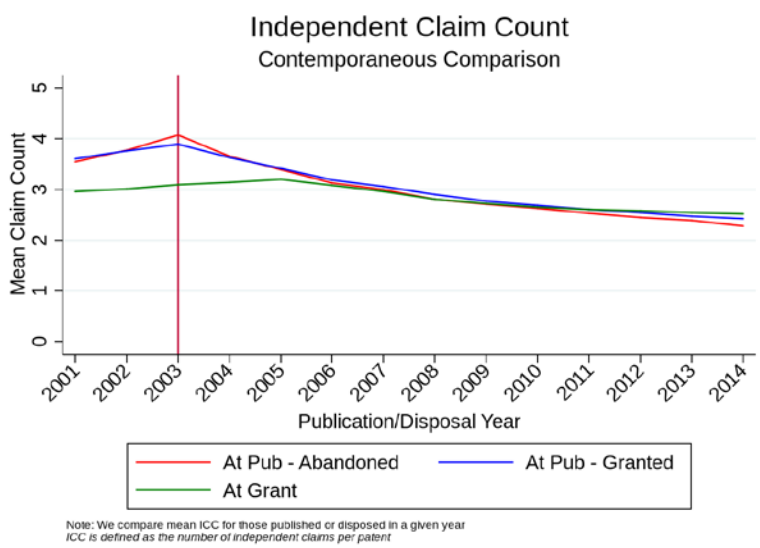


Figure 9

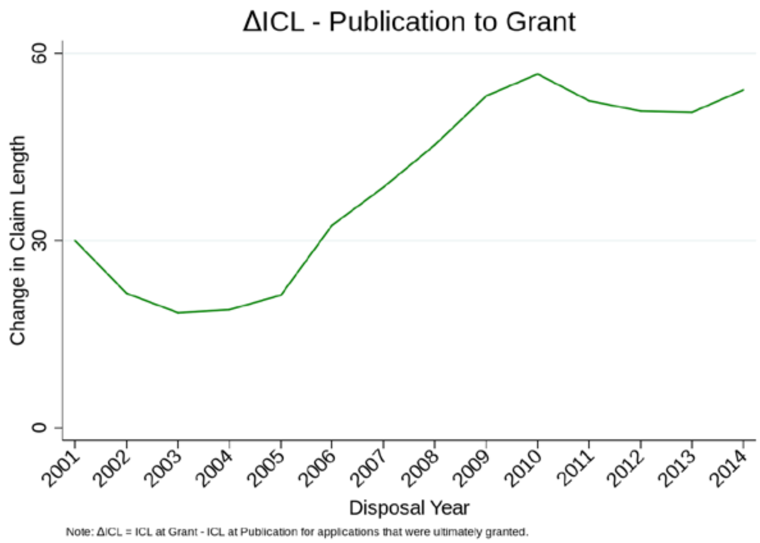


Figure 10

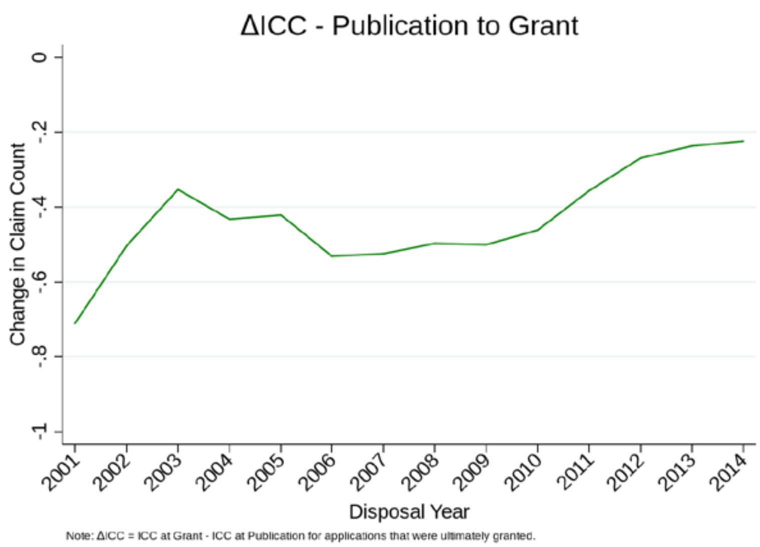


Figure 11

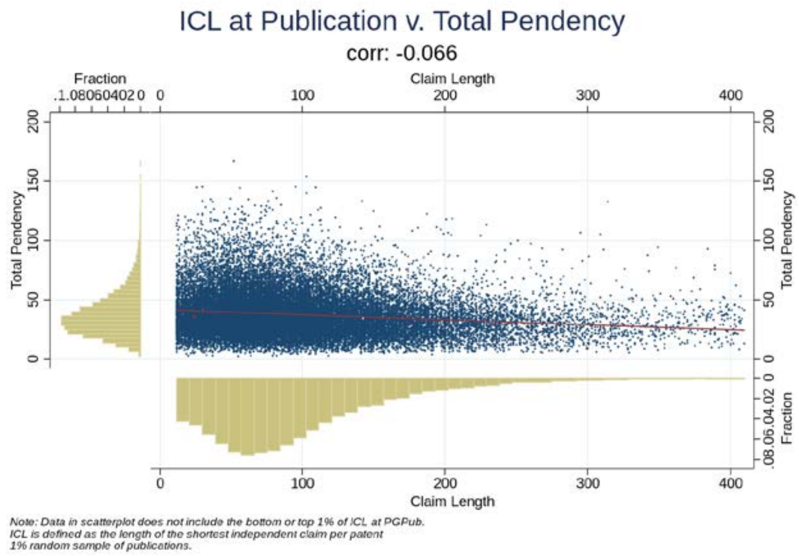


Figure 12

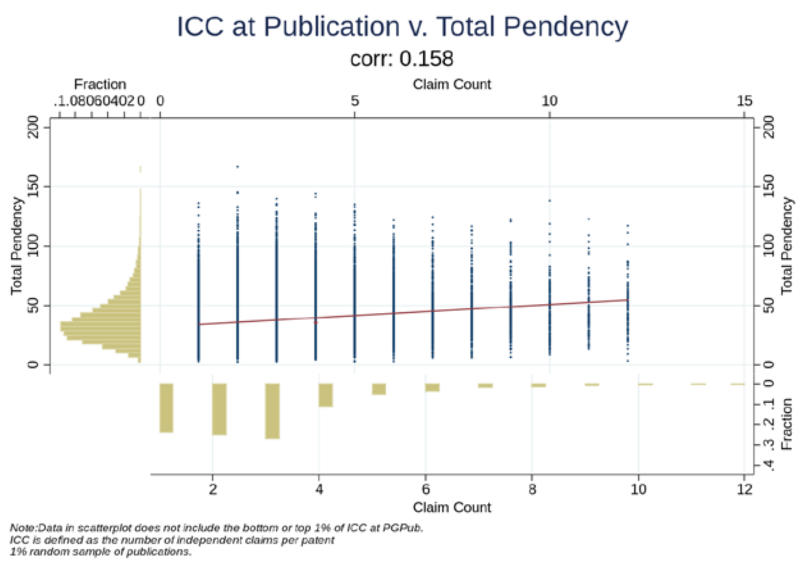
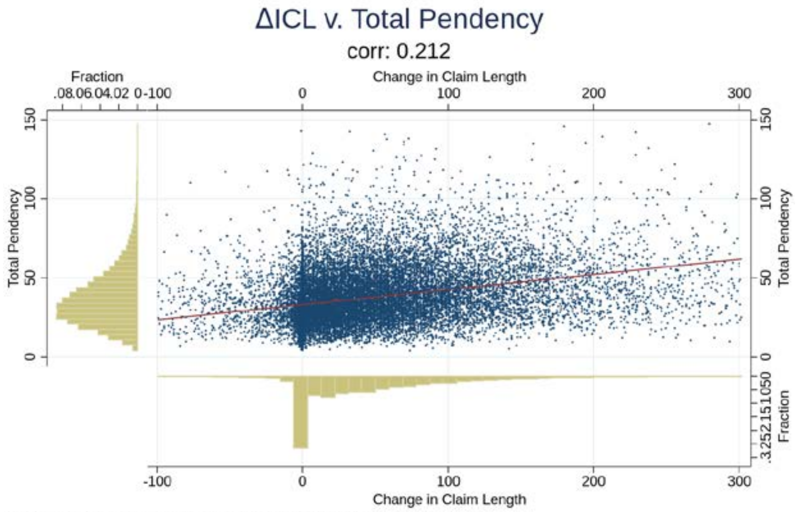
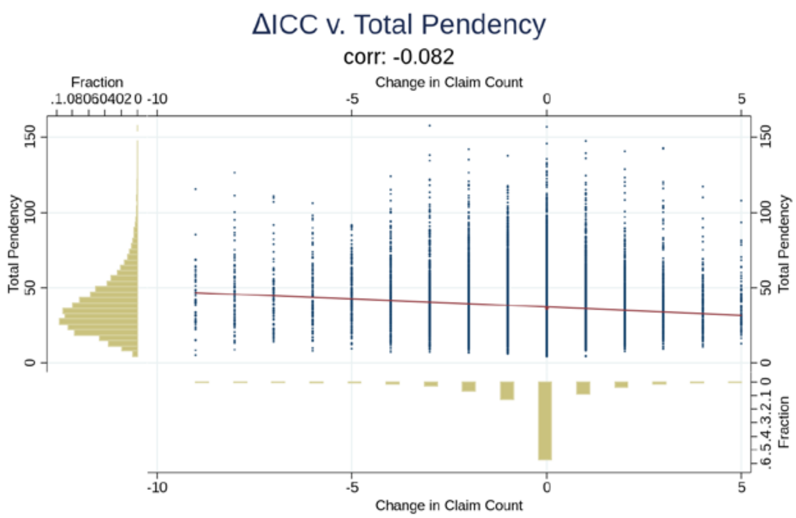


Figure 13



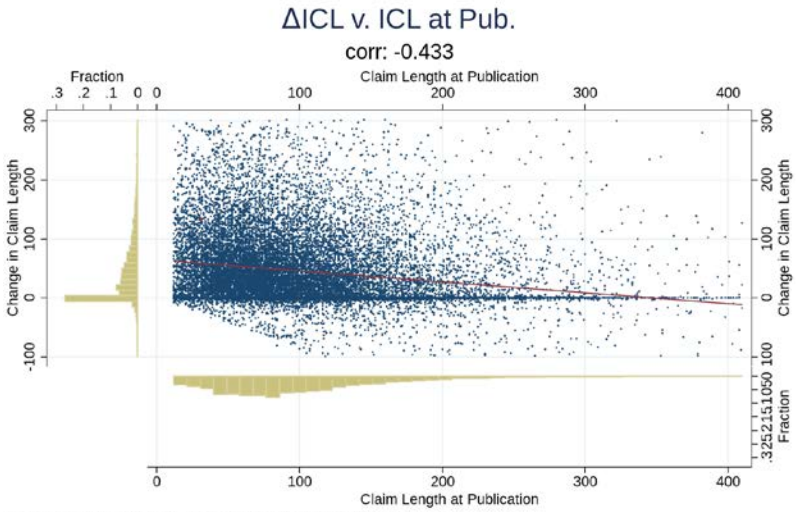
Note: Data in scatterplot does not include the bottom or top 1% of ΔICL from PGPub to grant.
ΔICL is defined as the change in length of the shortest independent claim from PGPub to grant
1% random sample of publication-patent pairs.

Figure 14



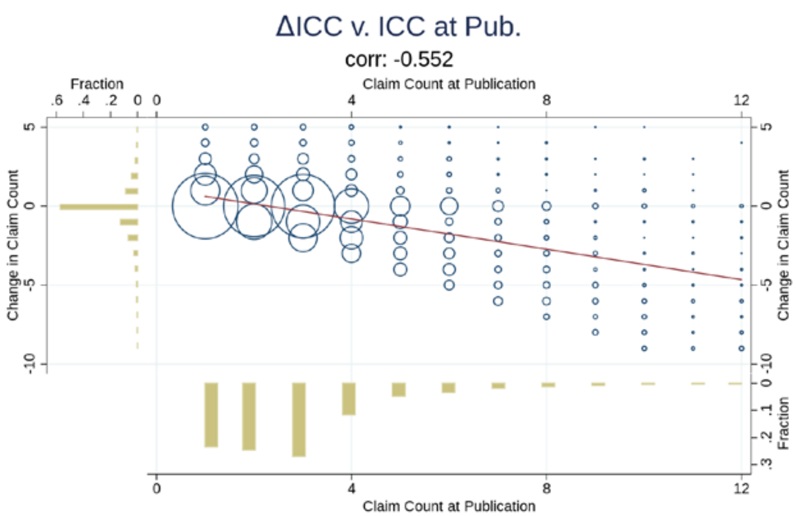
Note: Data in scatterplot does not include the bottom or top 1% of ΔICC from PGPub to grant.
ΔICL is defined as the change in length of the shortest independent claim from PGPub to grant
1% random sample of publication-patent pairs.

Figure 15



Note: Data in scatterplot does not include the bottom or top 1% of both ICL at publication and ΔICL.
ΔICL is defined as the change in the number of independent claims from PGPub to grant
1% random sample of publication-patent pairs.

Figure 16



Note: Data in scatterplot does not include the bottom or top 1% of both ICC at publication and ΔICC.
ΔICC is defined as the change in length of the shortest independent claim from PGPub to grant
1% random sample of publication-patent pairs.

Appendix A: Example of Dependent to Independent Claim "Roll Up"

In section 2, we provided an example of a published application (US 20050065799) in which dependent claims five and six at publication were "rolled up" into the first independent claim at grant (U. S. Patent 7769690). The application and patent text is provided below. Most, but not all, of the application's dependent claims five and six are incorporated into the granted patent's first independent claim. As you can see, the inclusion of the dependent claims into the independent claim narrows the scope of the independent claim.

U.S. Patent Application (US 20050065799 – filed 10/21/2002)

Independent Claim

1. A method for supply of data relating to a described entity to a relying entity, the method comprising:

generating a first digital certificate signed with an electronic signature by a first signing entity and including:

one or more attributes of the described entity;

one or more attributes of the first digital certificate which include one or more attributes identifying the first signing entity;

an indication of data relating to the described entity which is to be supplied;

an indication of one or more sources for the data to be supplied; and

one or more attributes identifying one or more relying entities to which the data is to be supplied;

the relying entity forwarding the first digital certificate for processing; and

a source supplying the data indicated in the first digital certificate.

Dependent Claims

5. The method of claim 1, wherein some or all of the data relating to the described entity is supplied by a second digital certificate to the relying entity, the second digital certificate signed with an electronic signature by a second signing entity and including:

one or more attributes of the described entity including the data which is to be supplied;

one or more attributes of the second digital certificate which include one or more attributes identifying the second signing entity; and

one or more attributes identifying one or more relying entities to which the data is to be supplied.

6. The method of claim 5, wherein the first digital certificate authorises the relying entity to use the first digital certificate to obtain a second digital certificate.

U.S. Patent (US 7769690 – Granted 8/3/2010)

Independent Claim

1. A method for supply of data relating to a described entity to a relying entity, the method comprising:

generating, using a computer device, a first digital certificate signed with an electronic signature by a first signing entity and including:

one or more attributes of the described entity;

one or more attributes identifying the first signing entity;

an indication of data relating to the described entity which is to be supplied;

an indication of one or more sources for the data to be supplied; and

one or more attributes identifying one or more relying entities to which the data is to be supplied;

the relying entity forwarding the first digital certificate for processing; and

after the processing, the one or more sources supplying the data indicated in the first digital certificate to the relying entity,

wherein some or all of the data relating to the described entity is supplied by a second digital certificate to the relying entity, the second digital certificate signed with an electronic signature by a second signing entity and including:

one or more attributes of the described entity including the data which is to be supplied;

one or more attributes of the second digital certificate which include one or more attributes identifying the second signing entity; and

one or more attributes identifying one or more relying entities to which the data is to be supplied, and

wherein the first digital certificate authorizes the relying entity to use the first digital certificate to obtain the second digital certificate.

Appendix B: Methodology

We applied a natural-language Python[57] algorithm to identify whether a claim is independent or dependent, and to parse each individual claim for the full text of each claim. To do so, we assumed that all dependent claims contain some dependency language referring to (and thereby incorporating limitations from) earlier claims, rather than actually reciting the language of limitations of the claims from which they depend.

In the `document_stats` dataset, we aggregated the individual claims-level data into patent/application-level summary statistics.[58] Each observation contains, for each application at publication and for each patent at grant, the number of independent and dependent claims, the average number of words in all independent claims, and a count of the number of words in the shortest independent claim. Since this paper's principal focus is the analysis of patent application and granted patent claims and filing characteristics, the dissemination and analysis of other patent-prosecution-related characteristics, such as data on RCE filings, numbers and types of continuations generated, appeals, etc., will be left to future dataset releases and analyses.

This Appendix details the data sources, methodology, descriptive statistics, and some general trends that can be observed in the `claims_stats`, `claims_fulltext`, and `document_stats` datasets. It is our hope that researchers will be able to use this data to enhance understanding of the examination process, including but not limited to assessing patent scope and how it changes during examination.

Data Sources

Our primary data sources for the claims-level datasets include the Patent Application Publication Full-Text and Patent Grant Full Text files provided by the U.S. Patent and Trademark Office (USPTO).[59] The Patent Application Publication Full-Text data, provided in XML format and disseminated as separate

---

[57] The Python code used to generate the USPTO's Patent Claims Research Datasets will be made available soon on GitHub.

[58] The data were obtained from USPTO Electronic Bulk Data Products (http://www.uspto.gov/learning-and-resources/electronic-bulk-data-products)

[59] Full-text of patents and patent applications is available at http://patft.uspto.gov/. Bulk data is available at http://www.uspto.gov/learning-and-resources/electronic-bulk-data-products.

files by years or ranges of years, contains the full-text of all patent applications published from December 2000 to December 31, 2014. The Patent Grant Full-Text files, provided in multiple file formats (XML, SGML, and APS), contain the full-text of all patents issued from 1976 to December 31, 2014. These files were cleaned, parsed, and appended to create the `claims_fulltext` datasets (one each for PGPubs and patents), which includes the patent or application number, the full-text of each claim, and an indicator variable to distinguish between independent and dependent claims, in a STATA® data file format.[60] In the `claims_stats` datasets (again, one each for PGPubs and patents), we include claim-level statistics (e.g. word count, number of "or"s, etc.) but not the full text of each claim. This allows researchers to analyze claim-level data in a more manageable dataset size.

For our analysis, but not included in our data release, we merged an in-house USPTO patent application database with the `document_stats` dataset to link certain filing and prosecution information at the application/patent-level for publicly available (published and/or granted) applications with our measure of patent scope. This information includes the nature of any parent application for the subject application (e.g., having a parent that was a foreign or PCT application) and the relationship to the parent of the subject application if the parent is a regular utility application (e.g., the subject application is a divisional application of that parent) and any filing priority information relating to the parent application. The USPTO in-house database includes various post-filing prosecution characteristics such as disposal type (`disp_ty`) and disposal date (`disp_dt`), among others.[61] We also used certain prosecution characteristics. For example, we use the disposal date for an application (which includes the time evaluating any requests for continued examination (RCEs) in the same application) to determine total pendency from filing to abandonment or grant ("disposal"[62]), and post-first-action pendency to measure the time from first-action to disposal. While the dataset does not include claim counts or claim lengths at the time of an abandonment, the our merged data on publications included a variable to distinguish whether the application matured into a granted

[60] Cancelled claims were identified in claims_fulltext but were not included in independent claim count and length summary statistic calculations.
[61] For a fuller description of all of the prosecution characteristic variables that were available for coding, please see the variable descriptions in Appendix C.
[62] There are two types of disposals: abandonment or grant. For more information on disposals and patent prosecution, please see http://www.uspto.gov/patents-getting-started/general-information-concerning-patents#heading-22. Please note that abandoned applications can sometimes be reinstated.

patent or ultimately went abandoned (disp_ty). Accordingly, many of our analyses distinguish characteristics at publication of applications that result in grants from applications that result in abandonments. Of course, abandonment (or grant) of a particular application does not mean that prosecution ended on the invention described in the abandoned (or granted) application, as various forms of continuation applications may have been filed prior to final disposition of any particular application.

Data Limitations

Relying on publicly available information on claims as captured from existing databases limits our sample in several ways. First, we can observe the claim text only at the time of publication and at the time of grant. This reliance also restricts the time period, because pre-grant publication of patent applications has been practiced by the USPTO only for applications filed after November 29, 2000.[63] Since that time, and without a non-publication request (which requires foregoing international protection on the patented innovation), publication has been required by statute 18 months after the filing priority date requested in relation to the earliest related parent application.[64] Applications filed prior to November 29, 2000 are unpublished. Thus, although our source patent dataset (grants) extends back to 1976, the bulk patent application data contains applications filed only during and after 2000. We have calculated that since November 29, 2000, approximately ten percent of filed applications have opted out of publication.

Further, in contrast to the captured data on claims from granted applications (at publication and at issue), machine-readable claim text is not readily available for abandoned applications (after publication). That is, we cannot observe the change in claims between publication and abandonment. Consequently, we limit our analysis of difference variables (dif_wrd_min, dif_wrd_avg, and dif_clm_ct) to publication-patent pairs (i.e., to applications that resulted in granted patents).

Although it is possible for claims in a particular application to change between filing and publication, we believe this is a relatively infrequent event. Our analysis shows that only 8.11 percent of total applications in the dataset have a preliminary claims amendment filed after their actual (not priority) filing date but before the publication date. Normal office practice is to incorporate preliminary

---

[63] *See* 35 U.S.C. 122(b).

[64] *See* 35 U.S.C. § 122(b)(2)(B); 37 C.F.R. § 1.213(a)(1)-(4).

amendments into the claims when they are published, and thus these claim amendments (except for the possible few that are filed too close to publication to be incorporated) are reflected in the publication data. Since the percentage of applications with preliminary amendments submitted between filing and publication is relatively small, we have treated for analysis the claims at publication as a reasonable approximation of the claims at filing.[65]

As can be expected in any dataset of this size, the source data files (the Patent Application Publication Full-Text and Patent Grant Full Text files) have some errors. Specifically, some claim language was excluded from the text and word length counts. The general (introductory) claiming language (e.g., "I claim" or "What is claimed is") has been excluded from the `claims_fulltext` datasets.[66] Similarly, we have not included the numeral associated with any claim in the claim length counts; rather, we have included only the language following the numeral for any particular claim (although the numeral is included in the dataset).[67] For example, U.S. patent 4,788,349[68] was issued with three claims of word lengths fourteen, two, and two, respectively. Excluding the general claiming language – which is not included in the datasets and consists of the words, "I claim:" – and the numeral assigned to the claims thus allows for one word claims such as chemical compounds. The exclusion of the general claiming language and numerals from the claim counts slightly biases the individual, average, and minimum independent claim length downwards.

Claim Identification and Measurements

As stated above, we used full-text claims data for patents and applications (`claims_fulltext`) to create patent-level summary statistics for both PGPubs and patents. We computed the summary statistics by applying a Python-based algorithm developed to distinguish independent claims from dependent claims and to compute various measures of claim length and claim count, among other variables.[69] The algorithm identifies independent from dependent claims by

[65] It should be noted that not all preliminary amendments are included in an application's publication. See MPEP 1121.

[66] There are exceptions in the claims_fulltext data set: (1) the first claims of twenty-two utility patents begin with the general (introductory) claiming language, "I claim"; and (2) claims in ten patents, such as patent 6,901,209, begin with the words, "I Claim." For example, claim 5 states, "I claim the access system of claim 4 characterized by the addition of data manager means to allow a user to access the program." This list is not exhaustive.

[67] The Claim number can be found as a separate field in the claims_fulltext data set (claim_no).

[68] See https://www.google.com/patents/US4788349

[69] See Python code in Appendix D

assuming that dependent claims will reference independent and other dependent claims, but not vice-versa.[70] Specifically, if claim language contains a direct reference to another claim or a group of claims, we designated the referring claim as a dependent claim (and coded it as such in the database). If the claim contains no such language, we designated it as an independent claim. We repeated this process for all applications at publication and for those that are granted at issue. To measure the independent claim length (ICL), we used a simple count of the number of words in each independent claim.[71] To create a patent-level metric, we measured ICL by using the minimum claim length among all independent claims of an application or granted patent. Our metric for the number of independent claims is a simple independent claim count (ICC). [72] We did not include in `document_stats` the minimum claim length for dependent claims. [73]

Following our assumption that patent scope depends on the length and number of independent claims, it is important to provide the arithmetic difference in the length and number of independent claims between publication and grant. These differences from publication to grant provide an approximation of the changes in breadth of the independent claims from filing to grant and thus of the change in the scope of the applications during prosecution. For example, as a direct result of

---

[70] It may be the case that a claim will contain referents to other claims that do not incorporate the other claims' limitations. However, we believe this to be a rare event.

[71] Because the algorithm uses natural language processing, claims that separate portions of words with spaces are automatically read as including separate words, which may thereby artificially increase the claim's word count. For example, chemical formula sometimes are written as a single word without spaces, but occasionally may contain many spaces, which would artificially increase the word count by as many spaces as are added. See US Patent 3,262,977, claim 4 ("N – [1' -phenyl-propyl-('1)] – 1,1 diphenyl-propyl-(3)-amine").

[72] Our algorithm also identifies specific words or phrases (e.g., "or" and "selected from") that are more likely to have the potential to broaden the scope of an independent claim by addition of other words, to permit robustness checks.

[73] To measure the dependent claim length (DCL), we would need to start with a simple count of the number of words in each dependent claim, and then add the count of the limitations language of the claim(s) from which the dependent claim depends and eliminate the count of the referential language in the dependent claim (as such language would then become duplicative and unnecessary). Nevertheless, the data in `claims_fulltext` are coded with the claim number(s) from which each dependent claim directly depends. Accordingly, some automated counts to approximate the number of words of dependent claims are possible to perform, e.g., by tracing the chains of dependency and adding the simple count of the words of each dependent claim and of the claim(s) from which it depends. (Such simple counts would be slightly over-weighted, by including counts of both the referential language and of the full text of the claim(s) to which those dependent claims refer). Some dependent claims, moreover, reference multiple independent or dependent claims that may have different lengths, which makes it more difficult to provide a count that is an accurate length for any such dependent claim. (Of course, each such multiply dependent claim could be decomposed into separate claims for further analysis.)

our assumption on patent scope, if the change in independent claim length (ICL) from publication to grant is positive, then it follows that the patent scope at grant should (generally) be narrower than at publication (and filing). If the change in independent claim count (ICC) is positive, then the scope of the patent should (generally) be broader at grant than at publication (and filing).

Appendix C: Variable Codebook

claims_fulltext (Patents and PGPubs – 2 separate datasets) Dataset

Variable Name	Variable Description	Notes	Dataset
appl_id	Application identification number	U.S. patent application number issued to an applicant at filing.	PGPub only
claim_no	Claim number		Both
claim_txt	Claim text	This variable includes the full text of each claim, dependent or indepdent	Both
ind	Indicator of independent claim	See identifying algorithm in Appendix 12.1.3 for more information	Both
pat_no	Patent number	U.S. patent number issued to an applicant at grant.	Patents only

claims_stats (Patents and PGPubs – 2 separate datasets) Dataset

Variable Name	Variable Description	Notes	Dataset
appl_id	Application identification number	U.S. patent application number issued to an applicant at filing.	PGPub Only
claim_no	Claim number		Both
cns_ct	Count of “consisting” in each claim		Both
deps	Referenced claims to which the claim is dependent	This variable includes the all claim references within the text of the observed claim	Both
ind	Indicator of independent claim	See identifying algorithm in Appendix 12.1.3 for more information	Both
or_ct	Count of “or”s in each claim		Both
pat_no	Patent number	U.S. patent number issued to an applicant at grant.	Patents Only
sf_ct	Count of “selected from” in each claim		Both

document_stats Dataset

Variable Name	Variable Description	Notes

appl_id	Application identification number	U.S. patent application number issued to an applicant at filing.
pat_no	Patent number	U.S. patent number issued to an applicant at grant.
dif_clm_ct	Difference in the count of independent claims between publication and grant	
dif_wrd_avg	Difference in the average count of words in independent claims between publication and grant	
dif_wrd_min	Difference in the minimum count of words in independent claims between publication and grant	
pat_clm_ct	Number of independent claims at grant	
pat_sc_ct	Number of semicolons among all independent claims at grant	
pat_sc_min	Minimum count of semicolons among independent claims at grant	
pat_wrd_avg	Average count of words among independent claims at grant	
pat_wrd_ct	Number of word among independent claims at grant	
pat_wrd_min	Minimum count of words among independent claims at grant	
pub_clm_ct	Number of independent claims at publication	
pub_sc_ct	Number of semicolons among all independent claims at publication	
pub_sc_min	Minimum count of semicolons among independent claims at publication	
pub_wrd_avg	Average count of words among independent claims at publication	
pub_wrd_ct	Number of word among independent claims at publication	
pub_wrd_min	Minimum count of words among independent claims at publication	