Screening Property Rights for Innovation*

William Matcham† Mark Schankerman‡

May 12, 2026

Abstract

We develop a dynamic structural model of patent screening incorporating incentives, intrinsic motivation, and multi-round negotiation. We use natural language processing to create a measure of patent distance, which, together with detailed data on examiner decisions, enables us to estimate the model and study strategic decisions by applicants and examiners. Using the estimated model, we quantify the effectiveness of the U.S. Patent Office and evaluate counterfactual policy reforms. We find that patent screening is moderately effective, given the existing standards for patentability. Examiners exhibit substantial intrinsic motivation that strongly improves screening quality. We quantify the annual social costs of patent screening at $15.38bn, equivalent to 5% of total private sector R&D in the U.S. and show that reforms limiting the number of rounds of negotiation significantly reduce social costs.

Keywords: Patents, innovation, incentives, screening, intrinsic motivation
JEL Classification: D73, L32, O31, O34, O38

*We are grateful to Kate Ho and three referees for their constructive comments and suggestions that greatly improved the paper. We thank Jakub Drabik for excellent research assistance and contributions. We have benefited from comments on earlier drafts of the paper by Dietmar Harhoff, Marco Ottaviani, Florian Schuett, and seminar and conference participants at many universities including Tel Aviv, Ben Gurion, Bologna, CEMFI, Boston University, Leuven, Hebrew University of Jerusalem, CEPR IO Conference, and the Zvi Griliches Memorial Conference. We thank Martin Rater, Daniel Ryman, and Andrew Toole at the U.S. Patent and Trademark Office for assistance in obtaining part of the data, and Janet Freilich and Michael Meurer for discussions on procedural aspects of patent prosecution. This project was partly financed by research grants from the Suntory-Toyota Centres for Economic and Related Disciplines (STICERD) at the London School of Economics, the Edison Fellowship Program at George Mason University, and the European Research Council.

†Department of Economics, Royal Holloway University of London, william.matcham@rhul.ac.uk
‡Department of Economics, London School of Economics, m.schankerman@lse.ac.uk

1 Introduction

Public institutions play a central role in promoting innovation. The two most important channels are government support for public and private research and the allocation of property rights. Support for research includes direct funding and indirect subsidies, while property rights, in the form of patents, enhance innovation incentives for private-sector R&D. In 2015, the U.S. federal government financed 24% of total R&D expenditures, or $120 billion (in 2025 USD). At the same time, the U.S. Patent and Trademark Office (hereafter, Patent Office or USPTO) issued 325,000 new patents. Patent rights promote innovation by increasing the private returns to R&D, facilitating access to capital markets, and underpinning the market for technology, especially for small, high-technology firms (Hall and Lerner, 2010; Galasso and Schankerman, 2018). The aggregate economic impact of these policies is magnified by the extensive knowledge spillovers they generate (Bloom, Schankerman, and Van Reenen, 2013).

Despite the importance of innovation-supporting public institutions, little is known about whether they allocate resources efficiently, and how organizational changes would affect their performance. The contribution of this paper, as part of a broader research program, is to use structural modeling to study the efficiency of resource allocation by innovation-supporting public agencies. Our context is the U.S. patent system, with a focus on the quality of screening by the Patent Office.

The effectiveness of the U.S. patent system is a hotly debated policy issue. Academic scholars and policymakers have argued that patent rights have increasingly become an impediment, rather than an incentive, to innovation. These concerns have been prominently voiced in public debates (Federal Trade Commission, 2011), U.S. Supreme Court decisions (eBay Inc. v. MercExchange L.L.C., 547 U.S. 338, 2006), and significant statutory reforms of the patent system, such as the 2011 America Invents Act. Critics of the patent system claim that the problems arise in large part from ineffective Patent Office screening, in which patents are granted to inventions that do not represent a substantial inventive step, especially in emerging areas such as business methods and software (Jaffe and Lerner, 2004). The issue is important because granting excessive patent rights imposes static and dynamic social costs: higher prices and deadweight loss on patented goods, greater enforcement (litigation) costs, and higher transaction costs of R&D, along with the potential for retarding cumulative innovation (Galasso and Schankerman, 2015).

We develop a dynamic structural model of patent screening in the U.S. that reflects the actual patent application and examination process. An applicant is endowed with patent claims that are heterogeneous in their true private value and their true distance to prior art (which consists of any knowledge in the public domain, including patents). These claims delineate the scope of the property rights sought by the applicant, and the applicant chooses how much to exaggerate

them beyond what is covered by the underlying invention, if at all. Exaggerating the scope of claims increases potential returns but also increases the risk of a lengthy, costly negotiation with the assigned examiner.

Once the application is assigned to a patent examiner, and in order to decide on whether to grant or reject the patent, the examiner searches the prior art to gauge whether the submitted application represents a sufficient advance to warrant a patent. The patent examiner does not observe the actual distance of each claim from prior patents. However, through their prior art search, the examiner obtains an error-ridden assessment of distances for each claim in the submitted application.

At each stage of the multi-round negotiation that follows their search, the examiner acts first, deciding whether to grant or reject the patent application. The examiner has grounds to reject the patent application if they assess any claims to have a distance to prior art below the patentability threshold. Nonetheless, an examiner with grounds for rejecting the application will choose to grant a patent if doing so maximizes their expected payoff. Upon receiving a rejection, the applicant decides whether to abandon their application or continue the negotiation. Continuing the examination involves narrowing the scope of claims, which increases their distance from prior art but, at the same time, reduces their private value.

The examiner's payoff from each decision includes an extrinsic incentive, known as credits, which form part of their performance assessment and consideration for a bonus. The examiner also incurs an intrinsic utility cost from granting claims with a distance below the patentability threshold. This component captures the idea that workers may care about behaving in a way consistent with the mission of the public agency. Hence, we incorporate the concept of intrinsic motivation from Besley and Ghatak (2005)—the alignment of workers' objectives with the public agency's mission.

Modeling examiner-applicant negotiations (or multi-stage bargaining in other contexts) is generally challenging, as it typically involves at least one agent forming beliefs about unobserved payoff-relevant variables at each stage of the negotiations. This complication is also present in the general version of our model, where the examiner would need to update their beliefs of the true distance and value of each claim based on the applicant's actions in each round. Empirical implementation of such a model is practically infeasible.

To make progress, we derive necessary and sufficient conditions on the functional forms of key elements of the model that ensure the examiner does not need to form beliefs about the underlying unobserved true distances and claim values. This step converts the model to one that we can solve by backward induction. Our empirical analysis adopts intuitive functional form choices within

the permissible class, and we examine the robustness of our estimates to alternatives within the class. We argue that the conditions we specify for simplifying the equilibrium are consistent with the specific institutional features of our context. There may also be other contexts in which our conditions reasonably apply, but our approach should not be viewed as a general approach to the estimation of dynamic multi-round negotiations.

We estimate the model using data on examiner decisions and patent claim texts. The Patent Office collects detailed data on all applications, not just granted patents, and records examiner decisions over negotiation rounds. The decision dataset we create covers approximately 55 million decisions on 20 million patent claims on applications filed between 2011 and 2013. The claim text data we use contains about 105 million claims granted in 1976–2020. We apply modern natural language processing methods to the text data to develop a novel measure of distance between patent claims, a key ingredient of the model. We use the distance measure to estimate, for the first time, the patentability threshold expressed in terms of the minimum distance from prior patents required for patent eligibility, representing the inventive step. Using the estimated patentability threshold, we can quantify the extent to which invalid claims are granted (false grants) and valid claims are not granted (false rejections). This information is at the heart of evaluating screening effectiveness.

We conduct external validation tests that confirm our claim distance measure provides an informative signal. However, since these "data" arise from the output of a neural network, we acknowledge that the distance metric may still contain measurement error. This potential problem occurs across all studies that use natural language processing and other AI methods to generate input variables.[1]

There are four primary empirical findings. First, patent screening is relatively effective, *given* the judicial standards of patentability that the Patent Office is mandated to enforce. While more than 80% of patent claims have an initial distance below the patentability threshold and should be rejected, screening weeds out or narrows them during negotiation rounds so that only about 6% of granted claims are below the threshold. Still, 13% of granted patents contain at least one claim that does not meet the threshold, implying that type 1 errors do occur.

Second, inventors substantially exaggerate the scope of their invention in the initial patent applications. On average, this raises claim values by about 20%, but there is substantial heterogeneity. Importantly, since the decision on how much to exaggerate is endogenous in the model, it changes

[1]Developing methods that account for using AI-generated variables as data remains an important topic for future research. A recent example is Battaglia, Christensen, Hansen, and Sacher (2024).

in response to reforms to the patent prosecution process. Third, abandonment of valid claims is more common than the grant of invalid claims, with 16% of abandoned claims meeting the threshold for patentability. This manifestation of imperfect screening has been largely ignored in the policy discourse. Finally, we estimate large and heterogeneous examiner intrinsic motivation. These estimates provide the first structural quantification of intrinsic motivation in a public agency, where one would expect worker motivation to be especially relevant (Besley and Ghatak, 2005).

We conduct counterfactual reforms, including changes to patent applicant fees, restrictions on the number of negotiation rounds, removing the intrinsic motivation of patent examiners, and limiting examiner credits. We study the effects of these reforms on three dimensions of performance: examination speed, measured by the equilibrium number of rounds; and two types of screening errors, measured by the frequency of granting claims that do not meet the patentability threshold ("type 1" errors) and not granting claims that do pass the threshold ("type 2" errors). Both errors impose social costs. Incorrect grants impose ex post welfare costs (deadweight loss) from higher prices and litigation costs associated with enforcing these patents. Failure to grant valid claims dilutes innovation incentives and discourages the development of new inventions that create positive social value.

A key feature of our counterfactuals is that reforms typically involve a trade-off between type 1 and type 2 errors: policies that make prosecution stricter lead to fewer grants of invalid claims but to increased abandonment of valid claims. The policy conclusions from reforms are thus ambiguous, as they will depend not only on the frequency of each type of error but also on the magnitude of social costs associated with those errors. Therefore, we develop a methodology to quantify the social costs in the current environment and under various counterfactual reforms. We estimate the total social costs of patent screening at $15.38bn per annual cohort of applications, which is equivalent to 5% of total R&D performed by business enterprises in the United States.

We show that the social costs of screening are affected by institutional design features. First, restrictions on the number of negotiation rounds (absent in the current U.S. patent system) significantly reduce the social costs of screening, up to 37% when only one round is allowed. Second, removing intrinsic motivation increases the frequency with which examiners grant invalid patents approximately sevenfold, further demonstrating that intrinsic motivation strongly affects the accuracy of patent screening. This finding highlights the importance of designing human resource policies to select examiners with high intrinsic motivation and maintaining this motivation over their careers.

Finally, we find that extrinsic incentives in the form of examiner credits, by themselves, have

little impact on any counterfactual outcomes. We interpret this result as indicating that the high levels of intrinsic motivation we estimate leave little scope for extrinsic incentives. However, removing credits when examiners have no intrinsic motivation does increase social costs, which indicates that extrinsic and intrinsic motivation are substitutes rather than complements in this context.

Related Literature We contribute to the literature on intrinsic motivation and design of incentives in mission-oriented agencies. Theoretical papers examine how extrinsic rewards can crowd out intrinsic motivation (Benabou and Tirole, 2003; 2006), while Besley and Ghatak (2005) emphasizes how intrinsic motivation—defined as the alignment between worker and agency objectives—induces welfare-improving sorting of workers and affects the optimal design of incentives and authority. Empirical studies have typically used field experiments to analyze how intrinsic motivation affects public agency performance, using various proxies for motivation (leading examples are Ashraf, Bandiera, Davenport, and Lee, 2020 and Khan, 2025). By contrast, our paper is the first to estimate intrinsic motivation in a structural model of a public agency, and our results confirm the importance of human resource policies that sustain high intrinsic motivation.

Recent papers study how screening mechanisms affect the performance of public agencies. Adda and Ottaviani (2024) develops a model of nonmarket allocation of resources, such as the awarding of research grants, and shows how informational constraints affect the optimal allocation rules. Li and Agha (2015) analyzes the allocation of research grants at the National Institutes of Health (NIH) and shows that peer review increases the effectiveness of grants in terms of post-grant citations. Azoulay, Graff Zivin, Li, and Sampat (2018) studies the economic impact of NIH grants, linking screening outcomes to publication citations and other innovation outcomes. Finally, Qiu (2023) develops a dynamic model of grant funding in the National Institutes of Health and studies the optimal allocation between young and veteran investigators.

In the patent literature, empirical papers have shown that patent examiner characteristics and extrinsic incentives affect the quality of granted patents (Cockburn, Kortum, and Stern, 2003; Frakes and Wasserman, 2017b). As a result of these findings, in our model we incorporate heterogeneity in patent examiner characteristics, including intrinsic motivation, that can affect the quality of patent screening.

The most closely related paper on patent screening is Schankerman and Schuett (2022), which develops an integrated framework to study patent screening, encompassing the patent application decision, examination, post-grant licensing, and court litigation. Their model is calibrated on U.S. data and used to evaluate a wide range of counterfactual patent and court reforms. While they estimate the effectiveness of patent examination, they treat it as an exogenous parameter, and

they do not model the prosecution process itself. Our dynamic equilibrium model of the patent examination process allows us to study how reforms to the negotiation process and agents' incentives affect screening quality. As such, our paper complements Schankerman and Schuett (2022), but an integration of the two approaches remains for future research.

2 Model of the Patent Screening Process

Before describing the model, we highlight two aspects of our approach. First, in a departure from most existing literature, we model patents as collections of claims that are heterogeneous both in their private value and in their distance from previous inventions. The patent document is composed of independent claims that delineate the scope of the property rights. In reality, and in the model, the examiner assesses the patentability of each claim separately by searching prior art and then decides whether to accept or reject the patent application *as a whole*. We provide more details about this negotiation process later.

Second, we analyze screening of a patent application, taking as given that the underlying invention has been developed. The validity of the structural model does not require a model of the potential applicant's decision to invest in developing their idea into an invention. However, to quantify the social costs associated with the screening system, as we do in Section 7, we develop a complementary model of the development decision.

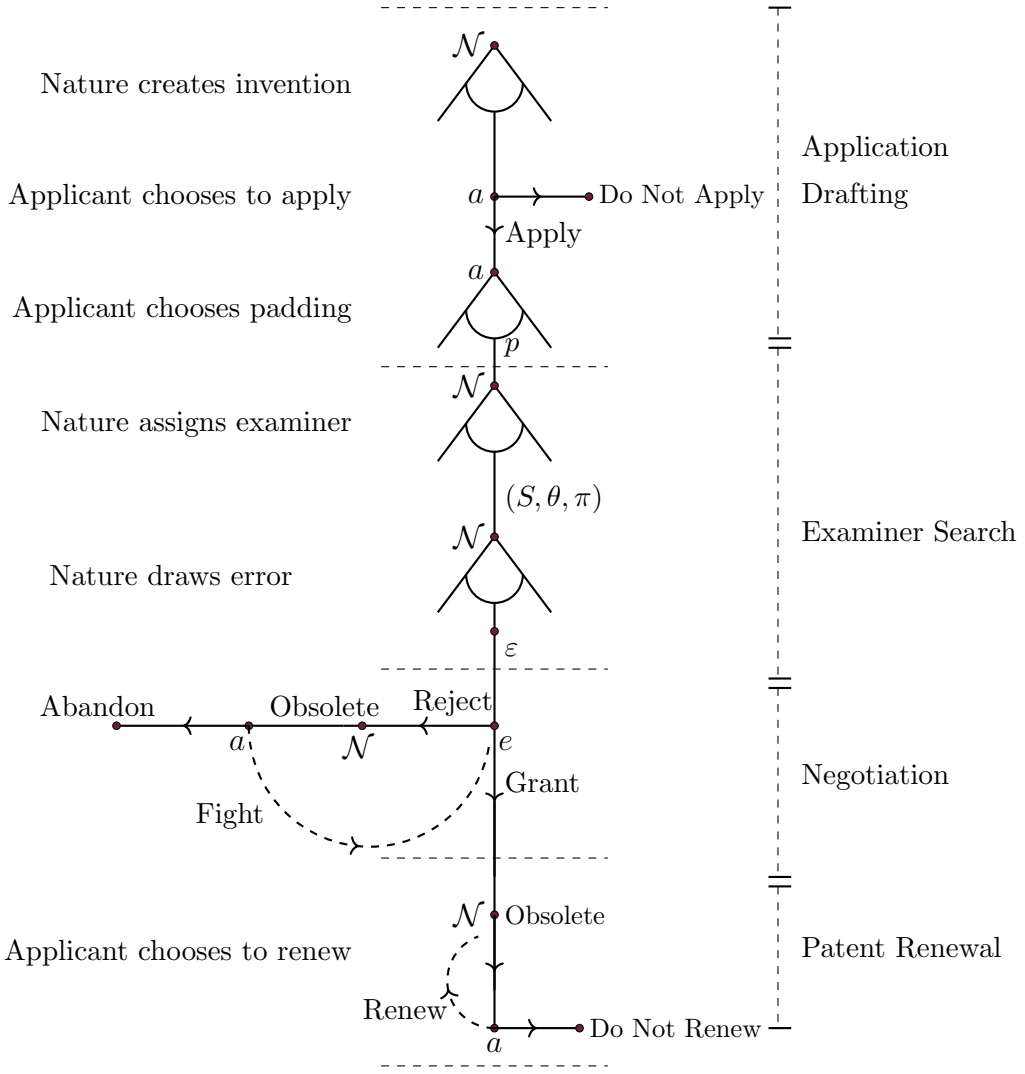
2.1 Model Description

We model the patent screening process as a dynamic game between an applicant, a, and an examiner, e, in technology area T. Both the applicant and the examiner are risk-neutral expected utility maximizers who discount future payoffs by the factor β each period. The model features four stages: (i) Application Drafting, (ii) Examiner Search, (iii) Negotiation, and (iv) Patent Renewal. We present the structure and timing of the model, focusing first on the actions of the applicant and examiner in each stage of the game, and then on their corresponding payoffs.

Figure 1 depicts the model's extensive form, starting with the "Application Drafting" stage. An applicant is endowed with an invention comprising M_0 components, each of which constitutes an independent claim in a patent application.[2] We characterize each claim by the pair (D_j^*, v_j^*),

[2]Note two points. First, we do not endogenize the division of the invention into claims. Second, a patent application includes both independent and dependent claims. The former delineates the main boundaries of the asserted property right, which determine the value of the patent to the applicant. Dependent claims clarify content of independent claims but do not expand the boundaries. In this paper we focus on independent claims.

FIGURE 1. EXTENSIVE FORM OF THE MODEL



where D_j^* is the distance of the true version of claim j to the nearest claim in any existing invention in the public domain ("prior art"), and v_j^* denotes the initial flow returns (or "value") that would be generated by the true version of claim j once commercialized. We define the returns v_j^* as relative to the applicant's outside option, for example, protecting by trade secrecy.

We emphasize that distance and value represent distinct dimensions that should not be conflated. A smaller claim distance indicates that the claim is more similar to existing patents from a technological perspective, but this alone does not mean it is low value.[3]

[3]For further discussion about technological distance and value, see the end of Section 3.4.

2.1.1 Applicant's Patenting and Padding Decisions

The applicant's first decision is whether to apply for a patent. If they do not, the game ends. Applying involves filing a patent application, which is a written description of the property rights associated with the invention. The applicant must choose the extent to which they exaggerate (or understate) the true scope of the claims in the patent application. We call this choice the level of *padding*, denoted by p. Padding obfuscates the true scope of the invention by concealing the true inventive step and thereby expands the property right.

The basic trade-off for the applicant in deciding how much to pad is between increased value from obtaining a large scope for the claim against the increased risk of rejection of the application by the examiner. Padding raises the applicant's revenue, but it moves the application closer to the prior art and thus increases the likelihood of examiner rejections during the examination process because the claim is too close to prior art. To capture the trade-off, we define initial padded (flow) returns to claim j, $\tilde{v}_j^1 = \mathcal{V}(v_j^*, p)$, which is increasing in v_j^* and p, and initial padded distance $\tilde{D}_j^1 = \tilde{D}(D_j^*, p)$, which is increasing in D_j^* but decreasing in p.

Figure 2 illustrates the trade-off relating to padding. The orange checkerboard semicircle in the top left corner represents the closest existing invention to the claim j, which is the small full blue circle in the bottom right corner. The applicant pads the true claim to create the larger (and higher value) cross-hatched circle. The distance between the true claim and the nearest existing invention is D_j^*, whereas the distance between the padded claim and the closest point is $\tilde{D}_j$. In the model and empirical work, the applicant pads all claims by the same proportion.

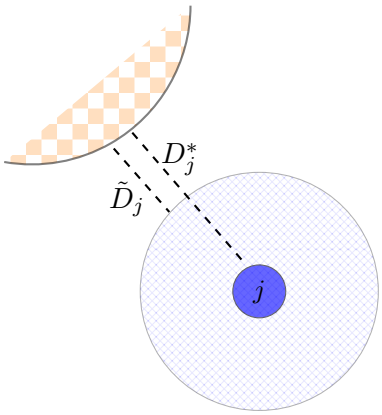
Finally, padding also involves additional drafting work for the applicant's attorney and thus creates a direct cost to the applicant at the point of application, denoted $F^{\text{app}}(p)$, in addition to the fixed Patent Office application fee, ϕ^{app}.

2.1.2 Examiner Assignment, Search, and Assessment

Moving to the "Examiner Search" stage of Figure 1, the Patent Office assigns the received application randomly to an examiner within the relevant technology area. The evidence broadly supports the quasi-random assignment of applications to examiners (Sampat and Williams, 2019; Feng and Jaravel, 2020). Once assigned, the examiner searches the existing prior art to assess the validity of the application.

Claims on patent applications are meant to meet two key statutory requirements for patentability under the U.S. federal code: novelty (U.S. Code 35, Section 102) and nonobviousness (Section 103). Novelty requires that the claim has not been in use for one year before filing. Nonobviousness requires that the claim makes an inventive step beyond the closest existing invention (or other

FIGURE 2. CLAIMS AND PADDING



publicly available prior art) that would not be self-evident to someone skilled in the relevant area. In this paper, we interpret the inventive step in terms of the distance between a patent claim and prior art, with the novelty and nonobviousness requirements satisfied if distance exceeds a threshold, denoted τ.[4]

After searching prior art, the examiner assesses the nonobviousness of each claim as written, that is, using the padded claim. Denote their initial assessment of claim j's padded distance by $\hat{D}_j^1 = \mathcal{D}(D_j^*, p, \varepsilon)$, where ε represents the stochastic examiner error in assessing nonobviousness. The function $\mathcal{D}$ is strictly increasing in D_j^* and ε and decreasing in p. It is increasing in ε because a larger error means that the examiner fails to identify some relevant prior art and thus over-estimates distance. To make the model empirically tractable, we assume that the examiner's assessment error is uniform across claims in the patent, remains constant during the examination, and is independent of the true claim distance D_j^* and p.

It is the examiner's mandate to reject the patent if any claim is judged to be below the patentability threshold (i.e., too small an inventive step relative to prior art). Thus, we say that the examiner has *grounds* to reject the patent application if the assessed padded distance of any of the constituent claims falls below the patentability threshold. However, as we explain in Section 2.1.4 where we describe the examiner's payoffs for granting/rejecting a patent, having grounds

[4]There is also the enablement requirement (Section 112), mandating the written claim be clear in identifying the boundaries of claimed property rights and precise, so that someone skilled in the arts could make and use the invention. Using the Office Action Research Dataset described in Section 3.1, we find that 80% of examiner decisions containing a Section 112 rejection also contain a 102/103 rejection. Thus, we focus on novelty/nonobviousness (hereafter referred to as nonobviousness). Finally, there is also a subject matter eligibility criterion (Section 101), a more procedural question of whether the content is "patentable", but this does not depend on distance from prior art and thus we do not consider it in this paper.

for rejecting a patent will not always compel an examiner to *choose* to reject the patent. This is because the examiner will make their decisions on whether to grant a patent to maximize their expected utility, which depends on the structure of incentives they face. This point is a crucial one, as it implies that examiners' decisions in the model, and in the data, may not align with decisions made solely on legal grounds.

2.1.3 Structure of Negotiations

Once the examiner has made their initial assessment of claim distances, the game moves to the negotiation stage. This stage is a finitely repeated version of the stage game shown in the "Negotiation" section of Figure 1. In each round of negotiation, the examiner chooses whether to grant a patent or reject the patent application. If granted, all claims are awarded, and the game moves to the renewal stage (described in Section 2.1.5). If the examiner rejects the patent, then in the model, the examiner automatically and mechanically rejects all claims j whose assessed distance is below the threshold.

After the examiner rejects the application, there is an exogenous probability that the invention becomes obsolete. If obsolescence occurs, flow returns become zero permanently.[5] Hence, if the patent is rendered obsolete at any round, the applicant immediately abandons the patent application and the game ends. We model obsolescence as a Markov process, with state variable ω_r equal to one if obsolescence occurs in or before round r and zero otherwise. Formally, if $\omega_r = 1$, then $\omega_{r+1} = 1$ (1 is an absorbing state). Otherwise, if $\omega_r = 0$, ω_{r+1} is a Bernoulli random variable with parameter P_{ω}^{pre}. In the model, obsolescence is independent of all other random variables.

After the realization of obsolescence, the applicant decides whether to continue negotiations or abandon the patent application. Abandoning ends the negotiation game. Continuing to the next round ("fighting") entails the applicant narrowing the scope of the claims that the examiner has assessed to be too close to prior art (i.e., claims with assessed distance below the patentability threshold), and then resubmitting the application. Narrowing involves increasing the padded distance of the claim from prior art and, at the same time, reducing the padded value (see Figure 2 for clarification). We treat the extent of narrowing for each such claim in round r as exogenous, denoted by η_r.[6]

[5] Obsolescence can arise from several sources, such as the arrival of a superior competitive invention, a negative demand shock or high development costs that render commercialization unprofitable. The patent returns fall to zero because the Patent Office publishes all applications, which undermines subsequent appropriation by trade secrecy as an alternative.

[6] We could extend the model to allow the applicant to choose whether to narrow padded distance by η with some probability or respond by arguing that the examiner is in error and not narrow at all. However, data on

Once the applicant has narrowed, the negotiation proceeds to the next round, $r+1$. The value of claims is now $\tilde{v}_j^{r+1}$. The examiner reassesses the distance of each claim the applicant narrowed, again using the prior art identified in their original search, obtaining a new distance assessment $\hat{D}_j^{r+1}$ (we formalize the determination of $\tilde{v}_j^{r+1}$ and $\hat{D}_j^{r+1}$ at the start of Section 2.3). Once again, on this basis, the examiner decides whether to grant or reject the patent. This process continues until the examiner grants the patent or the applicant abandons. In the final round R, if the examiner does not grant, the applicant must abandon.[7]

To illustrate how this negotiation structure and narrowing play out in practice, in Appendix B we provide examples of two actual patents that went through multiple rounds and highlight how changes in the patent text reflect the narrowing of patent scope.

2.1.4 Negotiation Game Payoffs

Examiner Payoffs The examiner's payoff for each of their actions consists of an extrinsic and intrinsic component. The extrinsic incentive takes the form of examiner credits. The Patent Office utilizes a point system that gives the examiner a specified number of credits for various decisions/outcomes at each negotiation round. Different credits are awarded when the examiner grants/rejects the patent, and also when the applicant abandons/amends the application. In practice, the examiner's performance is evaluated chiefly along two dimensions: the number of accumulated credits relative to production targets and the timely management of their portfolio of applications (Foit, 2018).[8] Bonuses are based on the extent to which the examiner exceeds their credit production target (falling short triggers reviews and potential penalties). This bonus structure incentivizes examiners to maximize the credits they obtain. In the model, we capture both dimensions of examiner performance evaluation through, first, credits associated with their actions, and second, *delay costs* that reflect the incentives for timely portfolio management. Below, we describe these two components in more detail.

Let $g_{GR}^r(S,T)$ denote the credits for granting a patent in round r for examiner with seniority S in

patent word counts imply that this extension is empirically unimportant. To examine this, we use the data in Marco et al. (2019) to identify word counts for patents granted with one rejection after publication, and calculate the proportion of cases in which the applicant resubmits an application with the same word count. This happens only 11% of the time, so we view the choice to ignore the possibility of no narrowing as a simplifying assumption in the baseline.

[7] We limit to six rounds in the empirical implementation, as 96% of applications last at most six rounds.

[8] We abstract from the examiner's intertemporal incentives that link different applications in their portfolio, such as meeting quarterly targets. Instead, we focus on the interaction between the applicant and the examiner on a specific application. A model in which examiners optimize decisions over all examinations in their docket is not necessary to meet the aims of our model and would introduce substantial complications.

technology area T, $g^r_{REJ}(S,T)$ the credits for rejecting the patent, and $g^r_{ABN}(S,T)$ the examiner credits if the applicant abandons the patent. The examiner also receives a credit of $g^r_{FIGHT}(S,T)$ if the applicant continues the negotiation (Appendix E provides details on credit schedule). All of these credits decline with the examiner's seniority level, S, and vary across technology areas, T, presumably reflecting higher productivity and differences in the complexity of the technology, respectively.[9]

The examiner's payoff from granting a patent also includes a second component reflecting their intrinsic incentive. This component derives from their level of intrinsic motivation, denoted by the parameter θ. Intrinsically motivated workers incur a disutility from awarding patents containing claims that do not meet the patentability standard, based on their assessment of claim distances. This intrinsic motivation moderates any incentive to maximize credits by granting patents prematurely. Formally, let M_r denote the number of claims in round r that the examiner thinks are invalid, i.e., claims with $\hat{D}^r_j < \tau$. The examiner's intrinsic utility cost from granting the patent is given by the function $\mathcal{R}(M_r, \theta)$, which is increasing in both arguments and satisfies $\mathcal{R}(0,\theta) = \mathcal{R}(M_r, 0) = 0$. For an examiner with any intrinsic motivation, granting a patent that contains claims they believe are invalid goes against the organization's mission statement, thereby reducing the utility of granting.[10]

Putting the extrinsic and intrinsic components together, the stage game payoff to the examiner from granting a patent in round r is

$$\mathcal{G}^r = g^r_{GR}(S,T) - \mathcal{R}(M_r, \theta). \tag{1}$$

We do not include an intrinsic utility cost to the examiner from rejecting the patent or from the applicant abandoning the patent. This exclusion follows from the fact that (i) the examiner's mandate is to reject the application if any claim is perceived to be invalid, irrespective of how many claims are valid, and (ii) the applicant always has the option to narrow the claims that are too close to prior art. Hence, the examiner's stage game payoff from rejecting a patent is

[9] For example, the most junior examiner gets 2.5 times as many credits for each action as the most senior examiner, and an application in Computer Architecture Software and Information Security provides 56% more credits than in Mechanical Engineering, Manufacturing and Products. See Appendix Section E for more information.

[10] One might be concerned that our specification of intrinsic motivation also captures examiner career concerns within the Patent Office. Their internal career prospects are supposed to depend on the frequency with which they grant invalid claims (Foit, 2018). For each junior examiner, a review of at least one grant/rejection decision per quarter is conducted by the supervising examiner. In addition, for the Office of Patent Quality Assurance, a senior panel conducts "random reviews" of examiners' decisions. However, these reviews are infrequent, do not come with explicit punishments, and are frequently successfully appealed by the head examiner in the art unit.

only the extrinsic incentive from the credit, $g^r_{REJ}(S,T)$. Likewise, the immediate payoff to the examiner from applicant abandonment is only $g^r_{ABN}(S,T)$.

There is an additional cost to the examiner if the applicant continues the negotiation to the next round, which we call the delay cost, denoted π. Delaying creates pressure on examiners, as they are evaluated, in part, on the effective and timely management of their portfolio of applications. This parameter will also reflect the examiner's productivity: more productive examiners incur a greater opportunity cost of continuing to another round (in terms of earning credits on other patent applications).

The structure of the examiner's stage game payoffs implies that, in the face of delay costs, examiners with insufficient intrinsic motivation may grant patents containing claims they believe to be invalid. It is in this sense that examiners with grounds to reject patent applications may go against the mandate of the Patent Office and grant a patent nevertheless.

Applicant Payoffs If the patent is granted in round r, the stage game payoff to the applicant is $V^r - \phi^{\text{iss}} - F^{\text{iss}}$, where ϕ^{iss} and F^{iss} are the Patent Office and attorney issuance fees, respectively, and V^r is the expected net returns from owning the patent, which is a function of the flow returns associated with each narrowed claim at the point of round r and the renewal decisions by the applicant after grant. We derive the expression for V^r in Equation (4) of Section 2.4, where we analyze the model.

The applicant's stage game payoff from abandoning the application is normalized to zero. If the applicant fights, they incur two kinds of fighting costs: attorney fees for amending the application, F^{amend}, and a Patent Office fee, denoted ϕ^{amend}_r, which varies by round.

2.1.5 Patent Renewal

If the examiner grants the patent, we enter the "Patent Renewal" stage of the model, the final stage of Figure 1. Our renewal model adapts Schankerman and Pakes (1986) (which studies European patents) to the U.S. context, adding stochastic, post-grant obsolescence in addition to deterministic depreciation. For a patent granted in round r, the returns for each granted claim j start at $\tilde{v}^r_j$ and depreciate at rate δ each period after the grant. With probability P^{post}_ω, the invention becomes obsolete, at which point all returns shrink to zero permanently. To maintain the patent rights, at ages (i.e., years after grant) $t = 4$, 8, and 12, the applicant must pay Patent Office renewal fees ϕ^{renew}_t, along with the associated attorney fees F^{renew} to implement patent renewal. The renewal decision occurs at the patent level, not the individual claim level. If renewed for the full term, the patent ends 20 years after the applicant applies, at which point the invention enters the public domain.

2.2 Information Structure

Applicant At each stage of the negotiation, the applicant knows (i) the true distance of each claim from prior art, D_j^*, (ii) the true private value of each claim, v_j^*, (iii) the complete set of attorney and Patent Office fees, (iv) the complete set of examiner credits across all rounds, (v) their choice of padding, p, and (vi) the values of narrowing, η_r the examiner will require if they are rejected.

Before applying for a patent, the applicant knows the set of examiners who might be assigned and their characteristics (S, θ, π), but does not know which examiner the Office will assign to their application. After the examiner is assigned, the applicant can calculate the examiner's search error ε exactly from the examiner's report of their distance assessment. Finally, after applying, the applicant knows the current and all prior realizations of obsolescence but does not know future realizations.

Examiner The examiner does not observe either true claim distances D_j^*, actual padded claim distances $\tilde{D}_j$, true claim values v_j^*, or the extent of padding p at any stage of the negotiation. At all points of the examination, the assigned examiner knows (i) the applicant's patent attorney and thus their fighting costs, (ii) all prior and current realizations of obsolescence but no future realizations, (iii) the structure of credits and Patent Office fees for the applicant, (iv) all of their prior and current assessments of claim distance, $\hat{D}_j^r$; and (v) the *initial* padded value of all claims, $\tilde{v}_j^1$ (but not the constituent components p and v_j^*). All other model parameters are common knowledge to the applicant and the examiner.

2.3 Simplifying the General Form

In this section, we address the informational challenge in the general form of the multi-round negotiation model. In general, in these contexts one or more agents must form and update beliefs on other agents' types based on observed actions. To solve this problem, we derive conditions on functional forms that obviate the need for belief updating.

For this task, we require a more general notation for narrowed padded values and narrowed distance assessments. Denote a given vector of narrowing across all potential rounds by $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_R)$, where R is the final round. The corresponding narrowed padded value of claim j is given by $\tilde{v}_G(v_j^*, p, \boldsymbol{\eta})$. The subscript G refers to the fact that $\tilde{v}_G$ is the most general functional form for narrowed padded value, and we will provide more restrictive functional forms in our results that follow.

This notation is sufficiently general to cover narrowed padded values of each claim j in every

round r of negotiation—whether further narrowing is required or not—for any possible level of narrowing. To see this, note that the initial padded value before any narrowing, $\tilde{v}_j^1 = \mathcal{V}(v_j^*, p)$ is equal to $\tilde{v}_G(v_j^*, p, \mathbf{0})$ where $\mathbf{0}$ is the R-dimensional zero vector. If claim j requires no narrowing after round one, then $\tilde{v}_j^s = \tilde{v}_j^1$ for all $s \geq 1$. Otherwise, claim j is amended and its narrowed padded value in round two is $\tilde{v}_j^2 = \tilde{v}_G(v_j^*, p, \boldsymbol{\eta}^1)$ with $\boldsymbol{\eta}^1 = (\eta_1, 0, \ldots, 0)$. The values $\tilde{v}_j^r$ are defined analogously for $r \geq 3$. A similar notation applies for the examiner's assessment of claim j's narrowed distance, $\hat{D}_G(D_j^*, p, \varepsilon, \boldsymbol{\eta})$. Hence, the examiner's initial assessment of distance for claim j is $\hat{D}_j^1 = \mathcal{D}(D_j^*, p, \varepsilon) = \hat{D}_G(D_j^*, p, \varepsilon, \mathbf{0})$ and it evolves as narrowing proceeds, analogously to padded value.

Under our information structure, the applicant does not need to form beliefs about examiner characteristics or actions. The applicant knows all current payoff-relevant variables and can calculate all future payoff-relevant variables.[11] However, since the examiner does not observe padding, true value, or true distance, the examiner cannot calculate future narrowed padded values or future narrowed assessments of distance, which are the payoff-relevant variables for the two agents. For example, since the examiner does not know p or v_j^*, the examiner cannot forecast $\tilde{v}_j^2 = \tilde{v}_G(v_j^*, p, \boldsymbol{\eta})$ in the first round, even if they know $\tilde{v}_j^1$ and $\boldsymbol{\eta}$. As a result, the examiner cannot predict the applicant's future actions that follow from their own decisions. In this context, the examiner must formulate beliefs over padding as well as true distances and values of all claims, and the examiner would need to update these beliefs at each round of negotiation based on the applicant's decision to continue fighting. Empirical implementation of such a general model is virtually infeasible.

The issue we just described arises because we specify unrestricted forms for the narrowed padded value, $\tilde{v}_G$, and narrowed distance assessment function, $\hat{D}_G$. To solve the problem, we derive restrictions on functional forms $\tilde{v}_G$ and $\hat{D}_G$ under which the examiner does not need to form beliefs. We begin by formalizing what it means for examiners not to require beliefs on the applicant's unobserved types. For any observed vector of narrowing $\boldsymbol{\eta}$, the examiner can calculate all future narrowed padded values, using only their observation on initial padded value, if there

[11] Note two additional points. First, we assume that the applicant and examiner never condition on payoff-irrelevant variables. Second, the fact that the applicant can anticipate future narrowing by the examiner raises a conceptual concern. The applicant could immediately narrow to the full extent required and avoid future fighting costs and risk of pre-grant obsolescence. This matter mirrors issues with early dynamic bargaining models, which could not generate negotiation in equilibrium precisely because players could anticipate future bargaining (e.g., Rubinstein, 1982). To remove this conundrum, one could introduce a stochastic element to future narrowing, but this would significantly complicate implementation of the model. We thank a referee for pointing this out.

is a function $\mathcal{W}_v$ such that

$$\tilde{v}_G(v_j^*, p, \boldsymbol{\eta}) = \mathcal{W}_v(\tilde{v}_j^1, \boldsymbol{\eta}), \tag{2}$$

for all $(v_j^*, p, \boldsymbol{\eta})$. Similarly, the examiner can calculate all possible future assessments of distance, using only their initial assessment of distance, if there is $\mathcal{W}_D$ such that

$$\hat{D}_G(D_j^*, p, \varepsilon, \boldsymbol{\eta}) = \mathcal{W}_D(\hat{D}_j^1, \boldsymbol{\eta}), \tag{3}$$

for all $(D_j^*, p, \varepsilon, \boldsymbol{\eta})$. If the examiner can predict all future narrowed padded values and distance assessments for a given narrowing vector, they will not require beliefs on the applicant's types because they can calculate all future values of payoff-relevant variables with the content of their information set. Hence, the existence of $\mathcal{W}_v$ and $\mathcal{W}_D$ satisfying Equations (2) and (3) defines our condition for the examiner not to require beliefs.

By the definitions of initial padded value $\tilde{v}_j^1 = \mathcal{V}(v_j^*, p)$ and of initial distance assessment, $\hat{D}_j^1 = \mathcal{D}(D_j^*, p, \varepsilon)$ it follows from Equations (2) and (3) that for the examiner not to need beliefs, $\tilde{v}_G$ must be weakly separable in (v_j^*, p) and $\boldsymbol{\eta}$, and $\hat{D}_G$ must be weakly separable in (D_j^*, p, ε) and $\boldsymbol{\eta}$. Hence, these separability conditions are *necessary* conditions to circumvent examiner beliefs. While this result restricts the class of potential functions for the empirical implementation to those weakly separable in narrowing, it does not tell us whether a specific choice of weakly separable $\tilde{v}_G$ and $\hat{D}_G$ will remove the need for beliefs.

However, Proposition 1 in Appendix C.1 shows that, under additional mild restrictions, weak separability is not just necessary but also sufficient to characterize the class of functional forms for $\tilde{v}_G$ and $\hat{D}_G$ under which examiner's beliefs are not required. Imposing the restrictions allows us to find the subgame-perfect equilibrium by backward induction and to formulate an empirical implementation of the model.[12]

The weak separability restriction implies that the marginal rate of substitution (MRS) between the true initial claim value and padding in generating claim value is independent of narrowing. Similarly, the condition implies that the MRS between true claim distance, padding, and examiner error, in generating assessed distance, is independent of narrowing. This property implies that the examiner cares about padded distance and its relationship to the patentability threshold, but not about true distance or padding individually. This is consistent with the examiner's mandate set by the Patent Office, which is to ensure that the inventive step as revealed in the padded application is sufficient to be patented.

[12] These restrictions ensure that there is an equilibrium not requiring beliefs in our setting. For a more general approach to belief-free equilibria, see Ely, Hörner, and Olszewski (2005) and Hörner and Lovo (2009).

Finally, we highlight that while weak separability seems reasonable in our context, it may not apply in other contexts. As an example, consider the case of a referee screening a scientific article. The referee presumably cares about the true scientific contribution of the article relative to prior art (D^*), independently of the extent to which the contribution is padded by the author. As stated above, this would violate the separability restriction, which rules out the referee caring about D^* by itself.

2.4 Analysis of the Model

Renewal Decisions We characterize equilibrium path actions using backward induction, starting with the renewal decisions. The applicant decides whether to renew a patent based on the expected returns from retaining patent rights. Renewal decisions are made at ages 4, 8, and 12. The expected returns of holding the patent from age t_1 to t_2 are[13]

$$\mathbb{E}_{\boldsymbol{\omega}} V_{t_1,t_2} = \sum_{t=t_1}^{t_2} (1-\delta)^t \left[\beta(1-P_\omega^{\text{post}})\right]^{t-t_1} \sum_j \tilde{v}_j.$$

Suppose the application is granted in round r. Then, conditional on surviving to age 12, the applicant renews at age 12 if $V_{12}^r \coloneqq \mathbb{E}_{\boldsymbol{\omega}} V_{12,20-r} - \phi_{12}^{\text{renew}} - F^{\text{renew}} > 0$. The applicant renews at the age eight if $V_8^r \coloneqq \mathbb{E}_{\boldsymbol{\omega}} V_{8,11} - \phi_8^{\text{renew}} - F^{\text{renew}} + I_{12}\beta^4(1-P_\omega^{\text{post}})^4 V_{12}^r > 0$, where I_t is equal to one if the applicant anticipates that they will renew at age t, and zero otherwise. An analogous decision rule holds for renewal at age four, which defines V_4^r. Finally, we define the ex post expected net benefits from patent rights, when granted in round r, as

$$V^r = \mathbb{E}_\omega V_{1,3} + I_4\beta^4(1-P_\omega^{\text{post}})^4 V_4^r. \tag{4}$$

Stage Game Decisions Let x_a^r and x_e^r denote the actions by the applicant and examiner at round r of the negotiation if the invention is not obsolete. The value function for the examiner after rejecting in round r, denoted W_e^r, satisfies

$$W_e^r = \begin{cases} g_{ABN}^r & \text{If} \quad x_a^r = \text{ABN or } \omega_r = 1 \\ g_{FIGHT}^r + \beta\left[-\pi + \max\left\{\mathcal{G}^{r+1}, g_{REJ}^{r+1} + \mathbb{E}_{\omega_{r+1}}\left(W_e^{r+1}\right)\right\}\right] & \text{Otherwise} \end{cases}$$

where, as a reminder, the examiner's payoff from granting $\mathcal{G}$ is defined in Equation (1). The examiner grants in round r, that is, $x_e^r = \text{GR}$, if $\mathcal{G}^r > g_{REJ}^r + \mathbb{E}_{\omega_r}(W_e^r)$. This inequality says that the examiner grants if the period payoff from granting exceeds the credits from rejecting plus the expected continuation value from the point of having rejected in round r.

[13] We use the notation $\mathbb{E}_{\boldsymbol{\omega}}$ to denote expectations taken over the vector of obsolescence shocks that are not yet realized. The notation $\mathbb{E}_{\omega_r}$ refers to an expectation over ω only in round r.

We next define the value function for the applicant upon being rejected in round r, denoted W_a^r. If the invention becomes obsolete, so that $\omega_r = 1$, then $W_a^r = 0$. Otherwise, we have that $W_a^r = \max\{0, \mathcal{U}_{r+1}^{\text{fight}}\}$, where:

$$\mathcal{U}_{r+1}^{\text{fight}} = -\phi_{r+1}^{\text{amend}} - F^{\text{amend}} + \beta\Big(I(x_e^{r+1} = \text{GR})\big[V^{r+1} - \phi^{\text{iss}} - F^{\text{iss}}\big] + I(x_e^{r+1} = \text{REJ})\mathbb{E}_{\omega_{r+1}} W_a^{r+1}\Big),$$

$I(A)$ is the indicator function, equal to one if statement A is true and zero otherwise, and V^{r+1} defines the ex post, expected net benefits from patent rights if granted in round $r+1$, as given in Equation (4). The applicant's decision rule after rejection follows directly from the statement of the value function above (fight if and only if $\mathcal{U}_{r+1}^{\text{fight}} > 0$).

Choice of Padding and Decision to Apply The applicant decides the initial level of padding without knowing the identity of the examiner who will be assigned to the application. The applicant chooses initial padding to maximize expected utility less legal costs, with the expectation taken over the roster of potential examiners $e = 1, \ldots, E$ (random assignment of applications implies an equal chance of each examiner in the relevant technology center), over the examiner error $\varepsilon \sim G_{e,\varepsilon}(\cdot)$, and over potential obsolescence of their invention. Formally, the applicant's optimal padding choice p^* maximizes the ex ante value of patent rights, $\Gamma(p)$:

$$\Gamma(p) = \frac{1}{E}\sum_{e=1}^{E}\left[\int \mathcal{Z}_a^0(e, \varepsilon, p)\, dG_{e,\varepsilon}(\varepsilon)\right] - \phi^{\text{app}} - F^{\text{app}}(p),$$

where the expected utility for the applicant, before applying, from padding choice p, when assigned examiner e who makes error ε is

$$\mathcal{Z}_a^0(e, \varepsilon, p) = I(x_e^1 = \text{GR})[V^1 - \phi^{\text{iss}} - F^{\text{iss}}] + I(x_e^1 = \text{REJ})\mathbb{E}_{\omega_1} W_a^1.$$

Finally, because the outside option of not patenting is normalized to zero, the applicant applies if the expected utility of the subsequent negotiation game is nonnegative:

$$\Gamma^* := \Gamma(p^*) \geq 0. \tag{5}$$

3 Data and Descriptive Analysis

3.1 Data Sources

Patent Claims Text We exploit the USPTO Granted Patent Claims Full Text Dataset, which contains the text for over 105 million independent and dependent claims in U.S. patents granted between 1976 and 2020 (United States Patent and Trademark Office, 2026a;b). In Section 3.2, we describe how we use these data to train an algorithm to construct a distance measure for granted claims.

Prosecution Rounds Estimating a model of the patent prosecution process over multiple rounds requires comprehensive round-level data on the patent process. We use the Transactions History data in the USPTO Patent Examination (PatEx) Research Dataset to create a dataset on the round-by-round evolution of patent applications (Graham, Marco, and Miller, 2018). For all patent applications, these data include examiner and applicant decisions at each examination round.

We match the round-level data to two datasets on patent applications. The first is the Application Data in the PatEx Dataset, which contains information on the applicant and examiner, the patent art unit (narrow technology classifications), and a binary indicator of the size of the applying firm (below or above 500 employees). We aggregate art units into the broader technology classifications used by the USPTO, known as "technology centers." Second, since we focus on novelty/nonobviousness rejections, we require data on the types of rejections of each claim. These data are available from the USPTO Office Action Research Dataset for Patents (Lu, Myers, and Beliveau, 2017). We use applications in these data between 2011 and 2013. After merging datasets, we obtain a sample of approximately 55 million claim-round decisions across 20 million claims.

Patents Renewal Rates and Fees We use renewal rates for grant cohort 2011 as reported in United States Patent and Trademark Office (2023). Mandatory patent office fees during prosecution and renewal of granted patents are from United States Patent and Trademark Office (2013).

Legal Fees We use data from the 2017 and 2019 American Intellectual Property Law Association Report of the Economic Survey (AIPLA, 2017; 2019). The survey reports statistics of the distribution of attorney fees for different tasks, including preparing and filing an application, paying renewal fees, amending applications, and mediating disputes. The statistics are split by three broad technology areas (biotechnology/chemical, electrical/computer, and mechanical). We use these data to estimate the distributions of attorney costs for each patent application, adjusted for inflation.

Examiner Credit Adjustments We obtained data on examiner seniority from Frakes and Wasserman (2017a), which provides a panel of General Schedule (GS) grades for examiners over time. Using this dataset, we calculated the seniority of the examiner at the time of each application. Finally, we obtained (unpublished) information on examiner credit adjustments from the Patent Office at the highly disaggregated US patent classification level, which we aggregated to the technology center level in our data (Rater et al., 2020).

3.2 Claim Distance Metric

The distance measure is the cornerstone of our empirical analysis of patent screening. It is essential for evaluating the performance of the public screening agency, and its use in a structural modeling context is novel.

Our approach calculates distances between claims by representing the text of each patent claim as a numerical vector and computing a metric over that vector space. Previous studies have used variations of the standard bag-of-words method to represent patent claim text as a numerical vector (Kelly, Papanikolaou, Seru, and Taddy, 2021). This approach, which looks for word overlap, has two significant limitations: it ignores word order and semantics. Word overlap is particularly troublesome in the context of patent applications, as attorneys strategically seek to describe the invention differently from the prior art.

We adopt the Paragraph Vector approach of Le and Mikolov (2014), which improves the bag-of-words approach by training a neural network (in our case, on patent claim texts after 1976) to "learn" the meaning of words by studying the context in which they appear and forming a vector representation for each word, picking up the meaning of paragraphs as a by-product. This approach is particularly suited for highly specialized texts such as legal documents and patents.[14]

Our approach involves four steps. First, we standardize the text and remove words that do not convey information. Second, we use the paragraph vector approach to represent the text of a patent claim as a numerical vector (Řehůřek and Sojka, 2010). We train the model to create a 300-dimensional dense vector representation of each independent claim.[15] We use the distributed memory method, which learns to predict a target word given the words in its context.

The third step involves taking every granted patent claim vector and calculating its distance to every previously granted claim. We use cosine similarity (CS) and angular distance, which are standard in the natural language processing literature. We calculate the angular distance metric between non-negative vectors x and y as $AD(x, y) = 2 \cdot \arccos(\text{CS}(x, y))/\pi$, which provides a normalized distance on the interval $[0, 1]$. The final step uses this distance measure to identify the closest, previously granted claim to the focal claim, which is the relevant measure in the model. For robustness, we experiment with using the mean of five closest distances. The resulting

[14] At the time this paper was developed, this was the state-of-the-art approach, but there is a fast-moving frontier. See Ash and Hansen (2023) for details on text algorithms and Ganguli, Lin, Meursault, and Reynolds (2024) for a recent study that compares various algorithms.

[15] The distance algorithm we train is unable to assess the distance between chemical formulas and also drawings, which play an important role in patents in technology center 1600 ("Biotechnology and Organic Fields"). Hence, our analysis using the distance metric excludes this technology center.

distribution of distances is similar.

3.3 Descriptive Statistics

Table 1 presents summary statistics of the data. First, 63% of applications result in a patent being issued. Second, the mean duration of patent prosecution is 2.51 years, and the mean number of rounds is 2.17, but there is substantial variation. This fact implies that some applications involve lengthy negotiation between the applicant and examiner. Third, while the modal number of independent claims on an application is 3, the number of claims varies, with a 99^{th} percentile of 9. This fact holds across technology areas; 97% of the variation in the number of claims is within technology centers. Lastly, in the sample, 25% of applications are filed by firms with fewer than 500 employees ("small entities").

Further empirical features are worth noting. First, 43% of granted patents are renewed to the statutory limit, and only 14% are not renewed at the first renewal date (age four). Second, as shown in Appendix Figure A.1, the distribution of granted claim distances is bell-curved with a left skew.

The majority of examiner decisions on claims is rejections: in the first round, the examiner rejects a mean of 83% of an application's independent claims. The most common outcome in the first round is for the examiner to reject all claims, but 11% of applications are granted in the first round, in which case the examiner rejects no claims. This bimodality is similar across technology areas, with 97% of the variation in round-one rejection rates occurring within technology centers. These facts suggest that distances to prior art are correlated across claims within an application, and this is confirmed by the fact that 83% of the total variation in claim distances is between patent applications, rather than within applications. Further, the distributions of claim distances are similar across technology areas, with 98% of the variation in distances within technology centers.

We complement these summary statistics with regression analysis, documented in Appendix Table A.1. Patent grant rates in our large sample vary sharply across technology centers and examiner seniority, with senior examiners granting more often. Also, the frequency of multi-round negotiation is much lower for senior examiners, it varies across technology centers, and small entities are less likely to negotiate. Finally, we decompose the variation in examiner-specific outcomes (such as their grant rate) into within- and between-technology center-seniority pairs: 78% of the variation in examiner grant rates and 82% of the variation in the average number of rounds is within technology center-seniority pairs.

These descriptive findings highlight the heterogeneity in the sample and confirm that the pre-

Table 1. Summary Statistics

Variable	Mean	Median	S.D.	1%	99%
Application Granted	0.63	1.00	0.48	0.00	1.00
Years of Prosecution	2.51	2.41	1.04	0.58	5.50
Negotiation Rounds	2.17	2.00	1.18	1.00	6.00
Independent Claims	2.59	2.00	1.75	1.00	9.00
Small Entity	0.25	0.00	0.43	0.00	1.00

Notes: "Small Entity" is equal to 1 if the applying firm has fewer than 500 employees. "Application Granted" is equal to 1 if a patent is issued.

dominant variation in outcomes is within technology areas, not between them. In the empirical implementation of the model, we incorporate sources of patent- and claim-level heterogeneity to account for the variations in the data.

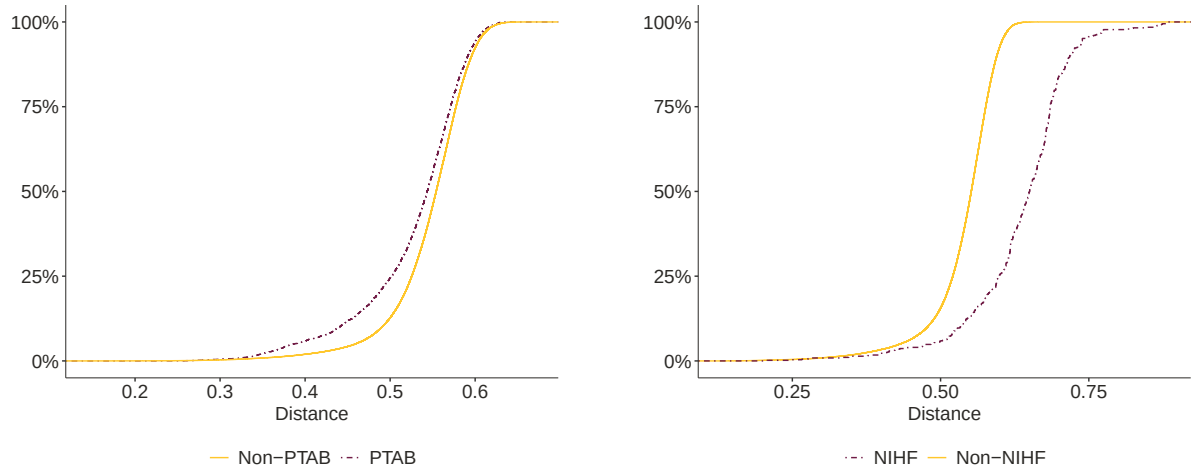
3.4 External Validation of the Distance Measure

Since there is no ground truth (non-AI-based) data on the padded distance between claims and prior art, we cannot evaluate the performance of the distance algorithm on an out-of-sample test set. Instead, we provide two external validation tests, which show that our distance measure produces reasonable results in contexts independent of its construction.

Selection Into PTAB Challenges We use data from the Patent Trial and Appeal Board (PTAB) in an external validation test of our claim distance measure (United States Patent and Trademark Office, 2025). The PTAB is an administrative mechanism within the USPTO that serves as a second layer of post-grant screening. Third parties can initiate challenges against granted claims on grounds of lack of novelty/nonobviousness, which are adjudicated by a panel of senior examiners.

The first validation test compares the distance to prior art for patents challenged in the PTAB with patents not challenged. On average, we would expect claims in patents challenged in the PTAB to have smaller distances to prior art than those not involved in challenges. To test this hypothesis, we take our sample of granted claim distances for patents applied for between 2011 and 2013 and locate all claims in the sample that feature in a PTAB challenge. Figure 3a plots the empirical CDFs separately by those challenged and not challenged in the PTAB. The figure confirms that unchallenged claim distances stochastically dominate PTAB claim distances.

FIGURE 3. EXTERNAL VALIDATION DISTRIBUTIONS



(A) PTAB (B) NIHF

As further evidence, the first two rows of Table 2 present the percentage of PTAB and non-PTAB claims with distance to prior art below the patentability threshold, far above the threshold (more than two standard deviations), and especially far above the threshold (more than three standard deviations). Claims in patents challenged at the PTAB (5705 in total in our merged sample) are twice as likely to be below the threshold ($p < 0.001$) and are also much less likely to be far above the threshold.[16]

National Inventors Hall of Fame (NIHF) The USPTO and National Inventors Hall of Fame maintain a record of about 650 inventors whose inventions are deemed to represent "world-changing technological achievements" (National Inventors Hall of Fame, 2025). As such, we would expect NIHF patents to be more distant from prior patents, and well above the patentability threshold.

To begin, Figure 3b presents the empirical CDFs of claims for NIHF and non-NIHF patents. The plots confirm that claim distances in NIHF patents strongly stochastically dominate those for non-NIHF claims. A regression of the log of claim distance on a dummy for NIHF, controlling for grant-year and technology-center dummies, shows that the distance is 14% larger for claims in NIHF patents, on average ($p < 0.001$).

[16]These findings also hold when we control for the technology area and granting year: regressing the log of claim distance on a dummy for a PTAB challenge and fixed effects for grant-year and technology center shows that claims in PTAB challenges are 3.23% closer to prior art than those not challenged ($p < 0.001$).

Since NIHF patents represent major technological achievements, the most appropriate validation test in the NIHF context concerns the right tail of distance distribution. The evidence in the final two rows of Table 2 confirms expectations. Seventy-two percent of claims in NIHF patents are more than two standard deviations above the threshold—nine times more than the non-NIHF equivalent of 8%—and 43% of claims in NIHF patents are more than three standard deviations above the threshold, compared to 0.01% of those claims for non-NIHF patents. Finally, claims in NIHF patents are 22% less likely to be below the threshold, compared to other non-NIHF patents. While around 7% of claims in NIHF patents are below the threshold, there is no reason to expect that *every* claim in an NIHF patent should be above the threshold, even if the patent as a whole (or specific claims in it) represent a significant technological achievement.

This evidence confirms that our distance measure is an informative signal of the padded distance between claims and the prior art. Of course, we cannot rule out that the distance measure may still contain some measurement error arising from the natural language processing algorithm. This feature is not specific to our context; it arises in any study that uses AI-based outputs as data.

Interpretation of the Distance Metric In our model, the relevant distance metric is the technological (dis)similarity of a patent claim from prior art – an indicator of the inventive step of the invention. We assume that technological distance and private value are distinct, uncorrelated variables. One can have an invention representing a small technological advance, but which is highly valuable, and a large technological advance that has a small potential market and thus limited private value.[17]

To illustrate this point with our distance metric, we consider a patent in our data set on a battery where the leading claim specifies "at least 4 Watt-hours of capacity" (US patent no. 8,617,259). We created a synthetic claim with identical text except that we doubled the battery capacity to "at least 8 Watt-hours of capacity". Our algorithm for measuring patent claim distance scores these patents as nearly identical in terms of the distance between them. Yet, clearly, the value of the synthetic patent would be larger due to the doubling of capacity. Of course, this example is artificial in that the extension to a larger capacity might require separate technological advances,

[17] A well-known example of the first is Amazon's patent family relating to one-click shopping. There is little dispute about its value, but its patentability was disputed. Amazon obtained a patent at the U.S. Patent Office but were made to amend in re-examination and faced rejections and refusals at the European Patent Office. The possibility of technologically significant inventions that are not valuable is illustrated by orphan drugs (indeed, that was the motivation behind the U.S. Orphan Drug Act of 1983).

Table 2. External Validation of the Distance Measure

	% $\tilde{D}$ Below τ	% $\tilde{D}$ Two S.D. Above τ	% $\tilde{D}$ Three S.D. Above τ
PTAB	19.09	16.79	0.09
Non-PTAB	8.67	20.99	0.50
NIHF	6.80	71.60	42.80
Non-NIHF	8.70	7.85	0.01

Notes: All differences between NIHF against non-NIHF and PTAB challenges against non-challenged are significant with p-value less than 0.001.

but in this case, our distance metric would reflect this, since the claim text of the synthetic patent would now differ.

Finally, we want to highlight a broader point: the appropriate metric depends on the focus of the study. In other contexts, distance metrics in other dimensions might be relevant, such as product market space to capture the degree of demand-side substitutability among inventions, or knowledge space to capture the channel through which knowledge spillovers are realized (Bloom, Schankerman, and Van Reenen, 2013). As such, the distance measure in this paper may not be suitable for all other patent-related contexts.

4 Empirical Implementation of the Model

4.1 Functional Forms and Distribution Choices

This section describes our functional form and distribution choices. Appendix Table A.2 summarizes all parameters, along with associated distributional assumptions.

Functional Forms

Before providing our functional form choices, we emphasize that some key objects of interest—true claim distance, true claim value, and padded claim value—are not observable to the econometrician. This challenge means that we cannot ground our choice of functional forms using external data on these objects. Instead, we conduct extensive robustness checks of our functional form choices, detailed after estimation in Section 5.3.

Padded Distance As described in Section 2.3, we require two key functions: a first, $\mathcal{D}(D_j^*, p, \varepsilon)$, that maps padding, true distance, and examiner search error into initial assessed padded distance,

and a second, $\mathcal{W}_D(\hat{D}_j^1, \boldsymbol{\eta})$, mapping the examiner's assessment of initial claim distance and the vector of narrowing into their assessment of narrowed padded distance. For the first, we specify the initial padded distance as $\tilde{D}(D_j^*, p) = D_j^* p^{-1}$ in the baseline and assume a multiplicative examiner error, so that the assessed distance is $\mathcal{D}(D_j^*, p, \varepsilon) = D_j^* \varepsilon p^{-1}$. We interpret the examiner error as a proportion of the initial padded distance (e.g., $\varepsilon = 1.1$ represents a 10% positive error).

For the second function (the examiner's distance assessment after r degrees of narrowing), we use the form $D_j^* \varepsilon p^{-1} / \left(\prod_{s=1}^{r}(1-\eta_s)\right)$ This form is weakly separable in assessed distance and narrowing, as required to ensure the examiner does not need to update beliefs (as discussed in Section 2.3 and proved in Proposition 1 in Appendix C.1). In the baseline model, we use constant narrowing over rounds, $\eta_s = \eta$ for all s, simplifying the expression to $D_j^* \varepsilon p^{-1} / (1-\eta)^r$. Our results are robust to more demanding specifications that allow narrowing to vary across rounds (see Section 5.3).

Padded Value Analogous to padded distance, as described in Section 2.3, we require two functions for padded value: one mapping padding and true value into initial padded value, $\mathcal{V}(v_j^*, p)$, and a second function mapping the applicant's initial padded value and narrowing into a narrowed padded value, $\mathcal{W}_v(\tilde{v}_j^1, \boldsymbol{\eta})$. For the first function, it is reasonable to assume that $\mathcal{V}$ exhibits super-modularity (complementarity) in padding and true value, hence we start with a Cobb-Douglas form: $\tilde{v}_j^1 = \mathcal{V}(v_j^*, p) = {v_j^*}^{\Upsilon} p^{\zeta}$. Because padded value is unobservable, we can rescale it and use ${v_j^*}^{\chi} \; p$, where $\chi = \Upsilon/\zeta$. Moreover, since we will assume that true claim value is log-normally distributed (see "Distributions" below), we can set $\chi = 1$ without loss of generality and arrive at the choice of pv_j^* for padded value.[18] This choice makes padding p a proportional increase in value. Finally, as with distance, we specify padded value for a claim narrowed for r times as $pv_j^* \prod_{s=1}^{r}(1-\eta_s)$, which reduces to $pv_j^*(1-\eta)^r$ under constant narrowing.

Legal Costs of Padding The cost of padding in terms of attorney fees is increasing and symmetric in the absolute value of padding. We specify $F^{\text{app}}(p) = f^{\text{app}}(1+ \mid p-1 \mid)$ where f^{app} is the application drafting fees per unit of padding. The idea is that it takes the attorney longer to draft an application that meets the required patentability standards when there is either positive or negative ($p < 1$) padding.[19]

[18] If $v_j^* \sim LN(\mu_v, \sigma_v^2)$, then ${v_j^*}^{\chi} \sim LN(\chi\mu_v, \chi^2\sigma_v^2)$, so estimates of the claim value parameters contain χ.

[19] Overstating the scope of the invention involves deciding on what elements to add as well as crafting the application to avoid the risk of not meeting the enablement requirement. Understating the patent scope involves deciding what elements to drop while still adequately revealing the invention to allow replication.

Intrinsic Motivation The model formulates the intrinsic motivation utility cost to the examiner as an increasing function $\mathcal{R}(M_r, \theta)$, where M_r is the number of claims the examiner grants that they believe to be invalid (i.e., below the patentability threshold). Our baseline specification makes this utility cost a function of the proportion of such invalid claims in the application: $\mathcal{R}(M_r, \theta) = \theta \frac{M_r}{M_0}$, where M_0 is the number of claims in the application. In robustness analysis, we experiment with an alternative specification, $\mathcal{R}(M_r, \theta) = \theta M_r$. There are no theoretical grounds for preferring one specification over the other, but using the proportion fits the data better.

Distributions

Applicant Variables We start by discussing true unpadded distances, D_j^*, and true unpadded flow returns, v_j^*. We assume that D_j^* and v_j^* are independent. Given that both these variables are unobservable in our data, it is unclear how to identify their correlation or test this simplifying assumption. We specify true claim distance D_j^* as Beta distributed with parameters (α_D, γ_D). The Beta distribution is a natural choice as it provides a flexible distribution on the interval $[0, 1]$, which coincides with the domain of our distance metric. Further, we use a multivariate normal distribution copula to allow for correlated claim distances within an application (for details, see notes to Appendix Table A.2). Motivated by Schankerman and Pakes (1986), the log of initial claim flow returns is assumed to be normally distributed with mean μ_v and variance σ_v^2.

Finally, all attorney legal fees (F^{amend}, F^{iss}, F^{renew}, and f^{app}) are log-normally distributed. For amendment, F^{amend}, and application drafting per unit padding, f^{app}, we specify different μ and σ parameters for simple applications (those with one independent claim) and complex applications in chemical, electrical, and mechanical fields. We focus our discussion on the estimates and robustness of simple application attorney costs, with associated parameters $\mu_{f^{\text{app}}}^{\text{simple}}$ and $\sigma_{f^{\text{app}}}^{\text{simple}}$.[20]

Examiner Variables Intrinsic motivation is log-normally distributed, and we allow for different μ parameters for junior (pre-GS-14) and senior grade examiners, denoted μ_θ^J and μ_θ^S, respectively. We constrain the σ_θ parameter of the log-normal distributions to be the same for juniors and seniors, but this does not imply equal variances (only equal coefficients of variation).

[20] We have data on the quantiles of the distributions of amendment, maintenance, and issuance hourly fees charged by lawyers. Since these moments directly correspond to the elements of applicant fighting costs and do not identify any other parameters in the model, we estimate the means and variances of the log of fighting costs using an external two-step generalized method of moments estimation procedure for each of these three types of attorney costs. This does not apply to application costs, which are linked to padding, and thus are estimated within the model.

In the baseline, we treat examiner delay costs, π, as constant, but in the robustness analysis, we allow this parameter to vary by round.

Examiner errors are normally distributed, with a mean and variance that are inversely related to the degree of intrinsic motivation. We microfound the relationship between moments of examiner search errors and intrinsic motivation in Appendix D. The basic idea is as follows. The examiner decides how intensively to search the prior art. Search is costly but increases the probability of uncovering relevant prior art. The utility cost of a search error increases with intrinsic motivation. Thus, the examiner's optimal search intensity rises, and errors decline, with their intrinsic motivation. As a result, intrinsic motivation affects endogenous outcomes through two channels: (i) making conscious errors more costly in the examiner's payoff function, and (ii) intensifying optimal search. Consistent with the microfoundations, we specify the error mean as $1 + 1/(\varrho\theta)$ and the variance as $\sigma_\varepsilon^2/\theta$. Because θ is log-normally distributed, $\varrho = 1$ without loss of generality.[21]

4.2 Estimating the Distance Threshold

We first detail how we estimate the distance (patentability) threshold $\hat{\tau}$ and then describe the formal conditions under which it is a consistent estimator of the true threshold, τ. Estimation is external to the model, using only observations on claim distances and examiner grant decisions.

For every examiner e, we calculate the minimum value of the granted (padded) distances among claims they grant in all years in our sample. We denote this quantity by $\tau_e = \min_{j \in M_e^{GR}} \tilde{D}_j$, where M_e^{GR} is the set of claims granted across all applications by examiner e and $\tilde{D}_j$ is the narrowed padded distance at the point of grant. We estimate the threshold as $\tau = \max_e \tau_e$. The intuition for the estimator is as follows. If an examiner is infinitely intrinsically motivated and does not make errors, they would never grant a claim with distance below the true threshold, so the minimum distance of their granted claims will, in the limit, be τ. However, for all other examiners, the minimum distance of their granted claims will, in the limit, fall below τ, whether because they choose to grant below the threshold or because they make an assessment error.

Proposition 2 in Appendix C.2 provides the formal conditions for consistency of $\hat{\tau}$. The key conditions are: first, that examiner errors are normally distributed with a mean that converges to one and a variance that converges to zero, as the level of intrinsic motivation converges to infinity; and second, that there is at least one examiner whose intrinsic motivation is sufficiently large.

We have already assumed the normality of examiner errors in Section 4.1. The requirements

[21] If $\theta \sim LN(\mu_\theta, \sigma_\theta^2)$, then $\varrho\theta \sim LN(\mu_\theta + \ln(\varrho), \sigma_\theta^2)$ for $\varrho > 0$.

on the asymptotic value of the mean and variance of examiner errors are consistent with the microfoundations in Appendix D and our empirical specification in Section 4.1. While we cannot directly test this assumption because intrinsic motivation is not directly observable, we are able to provide evidence consistent with it.

In Appendix D, we show that an examiner with higher intrinsic motivation invests more time in search and evaluation for any given patent, and so makes fewer examination decisions. Thus, the assumption implies that examiners who complete fewer decisions should exhibit a lower mean and variance of examiner errors. We compute the mean and standard deviation of the size of type 1 errors (grants of invalid claims) by junior and senior examiners, where the size of the error is the difference between the estimated threshold and the distance of incorrectly granted claims. Since examiners complete 39% more examination rounds on average as a senior than a junior, we should find that the mean and standard deviation of the type 1 errors are higher for seniors than juniors. The evidence is consistent with this: the mean is 11% higher (5.69% vs 5.13%), and the standard deviation is 3% higher (5.50% vs 5.32%) for senior examiners. Of course, one might be concerned that senior examiners are more productive and, on this account, less prone to errors. However, this would make our test conservative and strengthen the conclusion.

Finally, we compute the distance thresholds separately for each technology center. Since the inventive step is based on statutes and judicial decisions applicable to all technology fields, it is reassuring that our estimates of the threshold are very similar across technology centers, ranging from 0.47 to 0.49, on the $[0, 1]$ interval of the distance metric (and Beta distribution).

4.3 Model Estimation and Identification

We first summarize the model variables that are observable in the data. For the applicant, we observe the number of claims, (moments of) fighting costs, abandonment/fighting decisions by round, padded distances at grant, and renewal decisions. We do not observe pre- or post-grant obsolescence events, claim values, unpadded distance, narrowing, or padding. For the examiner, we observe seniority, technology center, credits, application grants/rejections by round, claim rejections by round, and examiner decision errors (based on our estimation of the patentability threshold). We do not observe examiners' intrinsic motivation θ, delay costs π, or distance assessment errors ε.

We estimate the model using simulated method of moments, minimizing the objective function $(\boldsymbol{m}(\boldsymbol{\psi}) - \boldsymbol{m}_{\mathcal{S}})' \, \Omega \, (\boldsymbol{m}(\boldsymbol{\psi}) - \boldsymbol{m}_{\mathcal{S}})$, where $\boldsymbol{m}(\boldsymbol{\psi})$ is the vector of simulated moments computed from the model when the parameter vector is $\boldsymbol{\psi}$, $\boldsymbol{m}_{\mathcal{S}}$ is the vector of sample moments, and Ω is a weight

matrix.[22] The number of available moments in the model far exceeds the number of parameters. To select our preferred subset of moments for estimation, we followed a data-driven methodology based on the sensitivity of parameter estimates to the inclusion of specific moments (described briefly below, and in Appendix F). The procedure pruned the set of moments to 46 that assist in estimating the parameters.

The selected moments corresponding to examiner outcomes are the proportion of applications granted by seniority and round, standard deviation of examiner rejection rates by seniority, the mean and standard deviations of the size of type 1 errors by seniority, and the proportion of patents granted containing an invalid claim by seniority and round. The moments corresponding to applicant outcomes are the proportion of abandonments by seniority and round, renewal rates, means and standard deviations of granted claim distances by grant round, means and medians of legal application fees by technology class, and the 75^{th} and 90^{th} percentiles of the distribution of initial returns from patents rights in the U.S., which we construct by estimating an external patent renewal model.

4.3.1 Sensitivity Analysis

As is common with complex nonlinear models, we cannot prove non-parametric identification of the model primitives or point identification of our parameters. Instead, we implement the approach developed by Andrews, Gentzkow, and Shapiro (2017), which proposes a sensitivity matrix that quantifies how changes in the value of a moment affect the estimates of each parameter. In our setting, the sensitivity matrix is $\Lambda = (\mathcal{M}'\Omega\mathcal{M})^{-1}\mathcal{M}'\Omega$ where $\mathcal{M} = \partial m(\psi)/\partial\psi$ is the Jacobian of the simulated moments, which we evaluate at our estimates. The element Λ_{ij} reflects how changes in the value of the moment in column j affect the estimate of the parameter in row i. Because our moments and parameters are not on the same scale, we convert the elements to elasticities to make them comparable and normalize the sensitivity elasticities by the sum of their absolute values across all moments. After normalization, matrix elements reveal the relative importance of each moment as a source of variation for estimating a given parameter.

The results show that most parameter estimates are driven primarily by a few key moments. For 18 of the 20 parameters in the model, at most three moments account for over half of the parameter total sensitivity. Further, for 13 of the parameters, over half of the total sensitivity is accounted for by one or two moments. For 10 of the parameters, two moments account for

[22] We use a diagonal weight matrix that transforms moments to a uniform scale. We cannot use the optimal two-step procedure because we do not have application-level data on fighting costs required to compute the correlation between fighting cost moments and others.

more than 75% of the sensitivity. Moreover, data moments that materially affect at least one parameter ($\geq$ 20% of their sensitivity) tend to move only one or two parameters, with only one moment materially affecting more than two parameters.

We next show that the specific small set of moments that drive each of the parameters has an intuitive interpretation. We highlight the primary moments driving parameters of interest, together with brief intuition.

(i) The parameters of the distribution of initial claim values (μ_v, σ_v) are driven almost entirely (94% and 95%) by the 75^{th} and 90^{th} percentiles of the distribution of initial returns to patents at grant, as externally estimated from a patent renewal model. These percentile moments drive only μ_v and σ_v. Intuition: Higher percentiles of the patent value distribution for those patents granted is most naturally achieved through higher initial claim flow returns.

(ii) The distance distribution parameters, (α_D, γ_D), are primarily driven by the mean and standard deviation of granted claim distances (42% and 52%, respectively). Intuition: Observed claim distances are clearly determined by unpadded claim distances.

(iii) Applicant fighting cost parameters for each technology area are driven almost entirely by the mean and median of relevant attorney fees of the corresponding fighting cost (varying between 82% and 98%, depending on technology area), and fighting cost moments drive only their corresponding fighting cost parameters.

(iv) The narrowing parameter, η, is driven primarily by abandonment rates (29%) and the error rates (26%). Intuition: Higher abandonment is generated by higher narrowing requirements, and larger errors can be explained by less narrowing of invalid claims in the screening process.

(v) The intrinsic motivation parameters (μ_θ^J, μ_θ^S, σ_θ) are driven chiefly by the mean and standard deviation of granted claim distances (57%, 67%, and 46%, respectively) and also by examiner errors (22%, 14%, and 14%, respectively). Intuition: A lower level of granted claim distance would manifest when examiners are more willing to grant invalid claims, as is the case with lower intrinsic motivation; more examiner errors will occur when examiners have lower intrinsic motivation.

(vi) The post-grant obsolescence probability, P_ω^{post}, is driven primarily by the four patent renewal moments, which account for 89% of the sensitivity. Intuition: Higher attrition at early renewal stages will come about with more frequent post-grant obsolescence (too high post-grant obsolescence is ruled out by the fact that nearly 50% of patents are renewed for

the full term).

(vii) The pre-grant obsolescence probability, P_{ω}^{pre}, is driven mainly by the abandonment moments (42%) Intuition: Through backward induction logic, applicants that foresee their abandonment in latter rounds will abandon immediately. Hence, the only way to generate higher abandonment rates after round one is through increased pre-grant obsolesence.

(viii) For the delay cost π, the dominant moment is type 1 error rates (50%). Intuition: Higher grant rates of invalid patents would come about from examiners facing increased costs of rejecting applications that incentivize grants of invalid applications.

(ix) The examiner error variance parameter, σ_{ε}, is driven by error moments (23%) and the mean and standard deviation of granted claim distances (29%). Intuition: Higher error moments are generated by increased examiner search error; higher variation in granted *padded* claim distances result from increased examiner search errors that subsequently lead to increased granting of invalid patents in earlier rounds.

5 Empirical Results and Robustness

5.1 Parameter Estimates

Table 3 presents our main parameter estimates for the applicant (Panel A) and examiner (Panel B). We report bootstrapped standard errors, which are negligible due to the large number of observations used to compute our data moments.

Applicant Parameters We estimate the per-round proportion of narrowing by the examiner as 36% per round. Thus, for a claim granted in the mean number of rounds (2.08), screening reduces the property rights granted to about 60% of what was initially sought. Therefore, the percentage of applications eventually granted, which has nearly reached 70% in the U.S., overstates the extent of property rights actually obtained by inventors.

Our estimates of the distribution of claim distance imply that 81% of application claims have initial distances below the threshold. Nonetheless, patent prosecution substantially narrows applications so that most are valid when eventually granted (see Section 5.2).

The distribution of initial returns from an unpadded claim is highly skewed, consistent with previous literature: the mean claim value is \$30,554, while the median is \$16,094 (both 2023 USD). The median initial unpadded returns from a patent *application* are around \$50,000 (2023 USD). Our estimates are broadly in line with previous estimates of U.S. patent values (Bessen,

TABLE 3. PARAMETER ESTIMATES

Parameter	**Symbol**	**Estimate**	**S.E.** ($\times 10^{-3}$)
Panel A: Applicant			
Per-round narrowing	η	0.36	0.05
Initial distance alpha	α_D	3.88	0.68
Initial distance gamma	γ_D	7.01	1.30
Initial returns log-mean	μ_v	9.51	0.94
Initial returns log-sigma	σ_v	1.13	0.35
Pre-grant obsolescence	P_ω^{pre}	0.17	0.10
Post-grant obsolescence	P_ω^{post}	0.04	0.03
Simple application fighting cost log-mean	$\mu_{f^{\text{app}}}^{\text{simple}}$	8.69	2.78
Simple application fighting cost log-sigma	$\sigma_{f^{\text{app}}}^{\text{simple}}$	0.85	3.58
Panel B: Examiner			
Junior intrinsic motivation log-mean	μ_θ^J	4.02	0.76
Senior intrinsic motivation log-mean	μ_θ^S	2.61	0.45
Intrinsic motivation log-sigma	σ_θ	1.00	0.33
Delay cost	π	1.29	2.69
Error standard deviation constant	σ_ε	0.16	0.19

Notes: This table provides the model parameters. Standard errors are bootstrapped. Table A.3 provides application attorney cost parameters by technology area.

2008), though the comparison is not perfect.[23]

Pre-grant obsolescence is high, at 17% per negotiation round (typically one year long). The post-grant obsolescence rate is 4% per year, which is similar to estimates in the literature (Pakes, 1986; Lanjouw, 1998). Pre-grant obsolescence is higher for two reasons: applicants are more likely to discover their invention is obsolete earlier in its life cycle (e.g., finding out that commercialization costs make the project unviable), and abandonment during prosecution is driven by obsolescence, making granted patents a selected sample.

[23]The comparison is not exact because we estimate the distribution of initial returns for all applications and unpadded claims, whereas the estimates in the literature are for padded value of granted patents. We are the first to distinguish between padded and unpadded value.

Applicants bear high legal costs for drafting an application. Application costs are as high as $41,690 (2023 USD) at the 90^{th} percentile of padding and fighting costs. Appendix Tables A.3 and A.4 contain parameter estimates of other attorney costs by technology area. These legal costs constitute part of the social costs of patent screening.

Examiner Parameters Our estimates indicate a high degree of intrinsic motivation, but with substantial variation across examiners: the estimated value of σ_θ implies a coefficient of variation of 131%. Junior examiners are more intrinsically motivated than seniors: the median junior examiner is four times more motivated than the median senior, but there is also considerable variation within each group. Two countervailing forces drive the relationship between intrinsic motivation and seniority. Senior examiners should have lower intrinsic motivation if they become "jaded" with experience, but selection implies that less motivated examiners are more likely to move to the private sector to receive higher remuneration. Our estimates suggest that the jading effect is stronger than the selection effect.

Intrinsic motivation utility costs are large relative to extrinsic rewards for both seniority groups. To illustrate, for a junior examiner (GS-9) in the chemical technology center with the median level of intrinsic motivation, the utility cost for knowingly granting a patent with all claims invalid is 2.01 raw credits (not normalized for seniority and technology area), which is equivalent to the credits the examiner gets from granting a patent. By contrast, the estimated examiner delay costs of going to another round are small. The maximal delay cost across all examiner seniorities and technology centers is 0.09 raw credits per round. This finding suggests that pressure to resolve applications promptly is ineffective (or unnecessary), despite docket management being one of the stated grounds for examiner evaluation in the Patent Office.

Recall that examiner assessment errors, ε, are normally distributed with mean $1 + 1/\theta$ and variance $\sigma_\varepsilon^2/\theta$. In the baseline specification, we constrain σ_ε to be the same for junior and senior examiners. Thus, our findings on intrinsic motivation imply that junior examiners have lower bias and variance in their search errors, relative to seniors. At the median level of intrinsic motivation, a junior examiner's mean error is 1.80% and standard deviation is 2.08%; for senior examiners, the estimates are 7.33% and 4.20%, respectively. Overall, examiner errors in assessing distance are modest for juniors, the 95^{th} percentile being 5.22%, but for seniors, the errors are larger, with a corresponding figure of 14.24%.

5.2 Simulated Padding and Examiner Errors

Padding is not observable in the data, so we simulate the model to calculate the implied distribution of optimal initial padding for applicants who (endogenously) apply. At the mean, padding

increases claim value by about 20%, rising to 31% at the 70^{th} percentile and 57% at the 90^{th} percentile.

We compute two key performance metrics with the simulated model: type 1 and type 2 errors. Type 1 errors occur when an examiner grants a patent with invalid claims. At the extensive margin, approximately 13% of grants contain at least one claim that should not have been approved. However, overall, only 6% of all granted claims are invalid. Further, most of these type 1 errors occur on claims close to the threshold. For example, only 2% of all granted claims are "egregious" errors, in the sense of being more than one standard deviation below the patentability threshold. Given our earlier finding that 81% of claims initially have unpadded distances below the threshold, this demonstrates that the prosecution process is relatively effective at screening out invalid claims.

Type 2 errors denote cases in which an applicant abandons an application that contains valid claims. At the extensive margin, more than 31% of abandonments include at least one valid claim, and among all abandoned claims, 16% are valid. As with type 1 errors, most type 2 errors also occur in cases of marginal validity: only 13% of abandoned applications contain a claim with a distance over one standard deviation above the threshold, and only 5% of abandoned claims are at least one standard deviation above the threshold.

5.3 Model Fit and Robustness Analysis

Our simulated model moments, calculated at the estimated parameters, successfully match most of the data moments used for estimation (see Appendix Figure A.2). The real test of model fit is the ability to match data moments external to the estimation procedure. To do this, we compute (i) percentiles on granted distances in each round; (ii) mean distance for fourth, fifth, and sixth rounds; and (iii) means and percentiles of round one rejection rates across seniority categories. Appendix Figure A.3 displays these moments. We match the set of external moments closely.

We also conduct extensive robustness checks on our baseline model. In what follows, we summarize the findings from the checks; the complete set of estimates is in Appendix Table A.5. First, we generalize the functional form linking padded distance to true distance and padding. Rather than proportionality, we specify $\tilde{D}_j = (D_j^*)^{\vartheta}/p$. The estimated ϑ is 1.77. Other model parameters are similar to the baseline estimates, apart from the parameters of the Beta distribution for unpadded distance, of course, which change so that the padded distance in the model still matches that in the data.

Second, in the baseline model, we compute the threshold as the maximum, across examiners, of the closest distance among each examiner's granted claims. For robustness, we experiment

with the first percentile rather than the closest distance for each examiner, in case the threshold is sensitive to outliers in finite samples. The parameter estimates from using the alternative construction of the distance threshold are very similar.

Third, we relax the assumption of constant claim narrowing in two ways. First, we allow the narrowing parameter η to differ between the first round and subsequent rounds. Narrowing is larger in the first round, estimated at 35%, as compared to 25% for later rounds. The other parameters remain robust. Second, we allow narrowing to vary by examiner seniority. The estimated per-round narrowing is higher for senior examiners than for juniors (37% vs. 29%, respectively). Other parameter estimates are similar.

Fourth, we consider two alternative specifications for the examiner's intrinsic motivation utility cost. The first specification allows the cost to be nonlinear in the proportion of wrongly granted claims so that $\mathcal{R}(M_r, \theta) = \theta \left(\frac{M_r}{M_0}\right)^{\varsigma}$. The estimated exponent is close to 1 ($\varsigma = 1.09$), and the other model parameters are similar to the baseline. The second version makes the cost a function of the number of wrongly granted claims rather than the proportion: $\mathcal{R}(M_r, \theta) = \theta M_r$. The estimated parameters are generally robust. It is worth noting that granted patents typically have at most one invalid claim, making it harder to pin down the intrinsic motivation parameter in the second specification. For this reason, we use the proportional specification (with $\varsigma = 1$) in the baseline.

Fifth, we allow examiner delay costs to differ after the second round of negotiation. We find that the delay cost is higher in the first two rounds as compared to later rounds, contrary to expectations: 2.21 vs. 1.70. In any case, the impact of delay costs remains very small, equivalent to at most 0.10 credits.

Lastly, we allow the variance of examiner distance assessment errors to differ by seniority directly (aside from the effect of intrinsic motivation). Our estimates, $\hat{\sigma}_{\varepsilon}^{J}$ =0.23 and $\hat{\sigma}_{\varepsilon}^{S}$ =0.09, indicate that, at the same level of intrinsic motivation, junior examiners have a substantially larger variance in search errors than seniors. This implies that junior examiners are less consistent in their evaluations, which may reflect less experience. However, evaluated at the median level of intrinsic motivation, the mean assessment error is still much lower for juniors than seniors (2.13% and 7.07%, respectively), and the 95% upper bound is 7.65% for juniors versus 11.00% for seniors. These implications are similar to those we obtained from the baseline model.

6 Counterfactual Analysis

We conduct a series of counterfactual experiments to study the impact of various reforms on the speed and quality of the screening process. Table 4 presents the results. For reference, we provide the baseline results in the first row of the table. We report the bootstrapped confidence intervals for the counterfactual outcomes in Appendix Table A.7. The confidence intervals are tight across all outcomes, and all differences we describe below are statistically significant at the 5% level.

Fees In the baseline, we set applicant fees at the actual Patent Office levels, which are relatively low and do not include any per-round fees until round three. In the first counterfactual, we introduce a $25,000 fee for each negotiation round. The fee gives applicants a greater incentive to exit the patent process swiftly and a weaker incentive to apply in the first place. Imposing this fee reduces padding by 17% and increases the fraction of inventions for which patents are not sought by 30% relative to the baseline. The mean (padded) value of claims rises slightly, reflecting self-selection by applicants.

At the extensive margin, type 1 error falls slightly, but type 2 error increases, as applicants more readily abandon patents with valid claims to avoid the increased fees. Similar results hold for the intensive margin errors. The trade-off between these two types of errors is a prominent feature of many of the counterfactuals we analyze. Finally, in the baseline and fee counterfactual, around 35% of type 1 errors and 43% of type 2 errors are egregious in the sense that the claim distances are more than one standard deviation away from the estimated patentability threshold.

Rounds Restrictions We consider three caps on the number of negotiation rounds: three rounds, two rounds, and a single round (removing all negotiation between the applicant and the examiner).[24] Restrictions on rounds strongly affect screening quality and speed. A two-round cap more than doubles the proportion not applying and reduces mean padding by 63%. As expected, the mean number of rounds decreases relative to the baseline, both because of the mechanical effect of the cap but also because of the sharp reduction in padding. All three caps increase the mean claim value through the selection effect, with a 26% increase in the limitation to one round.

Across all three caps, the proportion of granted patents with invalid claims decreases. In the

[24]These counterfactuals are motivated by a U.S. federal court decision which ruled that the Patent Office did not have the authority to restrict the number of rounds beyond two (which it viewed as a "substantive" change) but noted that it could make "procedural" changes such as increasing fees (SmithKline Beecham Corp. v. Dudas, 541 F. Supp. 2d 805, 2008). Since one can achieve the same equilibrium number of rounds with an "equivalent" fee, the distinction is problematic from an economic point of view.

Table 4. Counterfactual Experiments

Counterfactual	**Not Apply** (%)	**Pad** (%)	**# Rounds**	$\tilde{v}_j$	**T1** (%)	**T1 Egr** (%)	**T2** (%)	**T2 Egr** (%)
Baseline	14.18	20.50	2.08	29.35	12.61	4.08	31.31	12.55
25,000 Round Fee	18.39	16.94	1.97	30.07	12.15	4.21	33.75	14.43
Three Rounds	17.73	15.56	1.96	29.94	12.37	3.86	36.18	15.67
Two Rounds	32.31	7.59	1.64	31.88	11.79	3.99	38.62	15.73
One Round	65.92	-4.75	1.00	36.97	4.19	1.10	75.67	59.71
Credit$\searrow$	14.08	20.42	2.09	29.39	11.85	3.24	31.53	12.70
5% IM	5.25	52.36	1.67	46.44	91.88	78.45	7.84	3.67
5% IM + Credit$\searrow$	4.38	69.07	1.58	53.43	92.70	78.80	6.54	2.57

"Not Apply' is the percent of inventors who do not apply for a patent. "Pad" is the mean level of padding. "# Rounds" is the mean number of rounds. $\tilde{v}_j$ denotes the average padded claim value at grant, in thousands of 2023 USD. "T1" represents the proportion of granted patents with some invalid claims, and "T1 Egr" represents the proportion of granted patents with at least one claim with a distance more than one standard deviation below the threshold. "T2" represents the proportion of abandoned applications with some valid claims, and "T2 Egr" the proportion of abandoned applications with some claims having a distance more than one standard deviation above the threshold.

case of only one round, type 1 error falls sharply: at the extensive margin falling from 13% in the baseline to 4%. The downside of rounds restrictions is that they increase the proportion of abandoned applications with valid claims. The one-round cap raises type 2 error at the extensive margin from 31% to 76%.

For each rounds cap, we compute an "equivalent" per-round fee in the sense of generating the same equilibrium mean number of rounds. Since the applicant has private information in this setting, one might expect fees to be a more efficient instrument than rounds restrictions. The fee equivalent to a two-round cap is approximately $115,000 per round, which would be politically infeasible to implement.

Removing Intrinsic Motivation We evaluate two changes to intrinsic motivation. First, we reduce the intrinsic motivation parameter, θ, for every examiner to 5% of its original value. Reducing the intrinsic motivation value lowers the examiner's utility cost of granting invalid claims and increases their errors in assessing distance. Knowing that examiners are more willing to grant invalid claims, the proportion of inventors not applying falls from 14% to 5%, and mean padding increases from 21% to 52%. Despite the increase in padding, the mean number of rounds falls by 20%. Not surprisingly, type 1 error jumps sharply. At the extensive margin, type

1 error jumps sevenfold, with about 85% of these errors being egregious. Type 2 error declines by 75%. This result underlines that extrinsic incentives (examiner credits) alone cannot sustain the current level of screening performance.

Second, we keep the values of the intrinsic motivation parameter θ fixed but remove the intrinsic motivation cost from the examiners' grant payoffs, i.e. setting $\mathcal{R}(M_r, \theta) = 0$. This counterfactual quantifies the importance of the direct impact of intrinsic motivation on examiner payoffs. The results are qualitatively similar (with some attenuation to the size of the effects) to the previous experiment, which only reduces the intrinsic motivation parameter. We conclude that the primary channel through which intrinsic motivation affects outcomes is the effect on payoffs rather than the effect on the distribution of errors. Together, these counterfactuals highlight the importance of intrinsic motivation for the quality of patent screening and its potential salience in other public agencies.

Removing Examiner Credits We remove all credits for the examiner on an application after the first round. This change could be justified on efficiency grounds as "marginal cost" pricing since we estimate small examiner costs for an additional round of negotiation. Removing credits has a small impact on all outcomes. Intrinsic motivation is sufficiently strong for examiners to avoid granting invalid patents, even when they will receive no further extrinsic reward for doing so. This finding is inconsistent with the hypothesis that extrinsic incentives crowd out intrinsic motivation in our context.

We also analyze the effect of removing credits after the first round in addition to reducing intrinsic motivation to 5% of its original level, relative to just reducing intrinsic motivation alone. In this case, we find material impacts of credits consistent with economic intuition. Padding increases from 52% to 69% because applicants know that examiners have even more incentive to grant, and once again, despite the increase in padding, the equilibrium number of rounds falls by 5%. Type 2 error declines by 17% because the increased padding makes abandonments less likely to include valid claims. Hence, credits only work as an effective incentive when examiners are not (strongly) intrinsically motivated.

Final Remarks As is standard in counterfactual analyses, we emphasize that the maintained assumption in these experiments is that all other model (exogenous) parameters remain unchanged. However, one can envision scenarios where this might not be the case. We illustrate with two examples. First, in the event of a rounds cap, there are fewer opportunities for narrowing to be achieved, so the per-round narrowing η might increase. We would expect a further decline in the fraction of inventors applying for a patent and a further decrease in type 1 error.

Second, a reform that makes patent screening less rigorous, for example, policies that reduce intrinsic motivation, might be expected to produce "marginal quality" patents and thereby increase post-grant obsolescence, P_{ω}^{post}. We would expect this rise in obsolescence to offset some of our estimated increase in the fraction of inventors applying for a patent, and it would make applicants more likely to abandon, thereby exacerbating type 2 errors. At this stage, these repercussion effects remain speculative. We would require an expanded model that endogenizes what are treated as structural parameters of interest in order to pin down how these parameters change and be more confident about their implications for screening outcomes.

To conclude, we highlight that none of our reforms unambiguously improves both prosecution speed and quality. Policies that make prosecution stricter speed up the process and lead to fewer grants of invalid patents but also result in increased abandonments of valid applications. Therefore, evaluating reforms requires measuring the social costs of screening under each scenario, which we implement in the next section.

7 Quantifying the Social Costs of Patent Screening

7.1 Methodology

We summarize our methodology and calibration here; Appendix G provides details.

Type 1 Costs There are two sources of social costs from type 1 errors: deadweight loss from patent royalties and litigation costs from challenges against invalid patents.

To compute deadweight loss, we assume that the patentee charges the Arrow royalty equal to the unit cost saving from the invention. The deadweight loss from royalties depends on the market structure for licensees. Our baseline specification is perfect competition among licensees, with linear demand and constant unit cost.

We express the deadweight loss as $DWL = \frac{\lambda}{2}\frac{\Delta\wp}{\wp}\bar{V}$, where $\wp$ is the initial price (without the royalty associated with the patent), $\bar{V} = q\Delta\wp$ denotes total royalty payments, and λ is the absolute value of the elasticity of product demand. We set $\lambda = 2$ (results are robust to $\lambda = 1, 3$). See Appendix G.1 for details on our calibration of $\frac{\Delta\wp}{\wp}$.

The second component is litigation costs on patents with some claims below the patentability threshold. We assume that courts are perfect in that they invalidate patents if and only if they contain claims below the patentability threshold. However, not all invalid patents are exposed to litigation because their private value is not large enough to justify the litigation expense. Letting VSL denote the value at stake in litigation, patents are exposed to litigation if $VSL \geq \check{V}$,

where $\check{V}$ is a litigation exposure threshold. We calculate $\check{V}$ to match the proportion of patents not exposed to litigation in Schankerman and Schuett (2022), given by $\check{v} = 89.6\%$. Hence, $\check{V} = G_{VSL}(\check{v})$, where $G_{VSL}(\cdot)$ denotes our distribution of the private value of the patent right at stake in litigation. Exposed patents are challenged in court with probability of 16.4% from Schankerman and Schuett (2022).[25]

The social cost for invalid patents not exposed to litigation is the deadweight loss from royalties only. For exposed patents that are not challenged in court, we assume there is an underlying dispute that is settled by costly mediation incurred by both parties, in addition to the deadweight loss from royalties. For exposed patents that are challenged in court, we assume that the courts are perfect and thus always invalidate wrongly granted claims. The social cost in this case is the sum of litigation costs for both the patentee and challenger.[26]

Type 2 Costs From an ex post perspective, there is no social cost from a type 2 error, since the innovation has already been produced and publicized through the patent document, and the R&D cost is sunk. Thus, we analyze the social cost of type 2 errors from the ex ante (incentive) perspective. Type 2 errors reduce the expected value of patent protection to the inventor, thereby deterring some inventors from developing welfare-enhancing inventions. The social cost is the social value of welfare-enhancing inventions that are *not* developed when there is the possibility of type 2 error but *would* be developed in the absence of type 2 error. For this task only, we require a simple model of development.

The decision to develop an idea depends on three elements: the ex ante value of patent rights, Γ^*, development cost, κ, and the value of the invention without patent protection, Π. Our model calculates Γ^* (Equation (5) in Section 2.4). For development cost, we take random draws from the distribution of κ estimated by Schankerman and Schuett (2022). To derive the invention value without patent rights, we use $\Pi = \Gamma^*/\Psi$, where Ψ is the proportional increase in private value due to patent protection (the "patent premium"), which we calibrate based on Bessen (2008).

To compute social costs of type 2 errors, let $\mathcal{B}_i = \Pi_i + \max\{0, \Gamma_i^*\}$ denote the private benefit of development, which includes all profits associated with the padded invention: the gains arising from the true invention, plus the gains from padding. The latter gains represent transfers either

[25] Patentees with invalid patents can pre-empt a challenge by charging a royalty payment equal to the cost of litigation for the challenger. For these cases, the social cost is only the deadweight loss since the payment is a pure transfer from the licensee to the patentee.

[26] We assume litigation costs are linear in the value at stake and calibrate the coefficients using data from the American Intellectual Property Law Association (AIPLA). The AIPLA also provides mediation costs by value at stake, which we associate to each developed invention. See Appendix Section G.1 for details.

from other firms or consumers, as they do not represent gains from any technological improvement. As such, they are a "business stealing" effect. An inventor invests to develop an idea i in the presence of type 2 error if their private (net) benefits $\mathcal{B}_i - \kappa_i$ are non-negative.

Social benefits include private gains from the true invention plus the externality but exclude pure transfers from padding. We assume that externalities arise from the true invention (but not from any padding associated with it). Thus, the social net benefit is given by $SNB_i = \frac{\rho^S}{\rho^P}\left(\frac{\mathcal{B}_i}{p_i}\right) - \kappa_i$ where $\mathcal{B}_i/p_i$ represents (to a first order approximation) the true invention and ρ^S/ρ^P is the externality multiplier given by the ratio of the social and private rates of return. An idea is socially beneficial if $SNB_i \geq 0$. For the baseline, we use a conservative estimate of $\rho^S/\rho^P = 2$ from Bloom, Schankerman, and Van Reenen (2013).

Finally, to calculate the set of ideas that would be developed in the absence of type 2 error, we simulate the outcomes from a counterfactual experiment in which, at the point of patent abandonment, the inventor obtains the value of all valid claims in the patent. In this scenario, all remaining abandoned claims are invalid, so there is no type 2 error. Let $\mathcal{B}'$ denote the private benefit of development in this new scenario. Idea i would be developed in this scenario if $\mathcal{B}' - \kappa_i \geq 0$. We then compute type 2 costs as the sum of social net benefits SNB_i, across ideas with $\mathcal{B}_i - \kappa_i < 0$ but $\mathcal{B}'_i - \kappa_i \geq 0$.

Patent Prosecution Costs Patent prosecution costs are the sum of Patent Office administrative costs and all applicant legal costs, including costs of drafting the patent application (F^{app}) and amending it (F^{amend}) during each round of prosecution. For administrative costs, we multiply the number of claim rounds in the simulation by the average USPTO cost per round and per claim (United States Patent and Trademark Office, 2005; details in Appendix G).

Benefits of Errors Finally, weighed against these social costs, we account for the potential benefits from both types of errors. The benefit of type 1 errors is that they increase incentives for inventors to develop and patent their ideas. This benefit is analogous to the cost of a type 2 error. We compute type 1 benefits as the sum of social development benefits from welfare-enhancing projects that would not be developed without type 1 error, but that are developed with type 1 error. The benefit from type 2 errors is the deadweight loss avoided by not having granted the patent right. There is no benefit associated with litigation cost savings since, under our assumption of costly but perfect courts, valid patents that are granted would not be challenged in equilibrium.

7.2 Estimated Social Costs of Patent Screening

Table 5 summarizes the social costs of screening per annual cohort of potential inventions (in 2023 USD), for the baseline model and counterfactual reforms. We report the 95% percentile bootstrapped confidence intervals for the total social cost; Appendix Table A.7 provides confidence intervals for each component of social costs.

Social costs from type 1 errors are \$3.01bn, \$0.25bn from type 2 errors, and \$12.11bn from prosecution costs. It is striking that prosecution costs dominate, and the bulk of those costs are associated with applicant legal costs rather than Patent Office administrative costs. The total social cost of patent screening is \$15.38bn, equivalent to approximately 5% of all R&D performed by business enterprises in the U.S. in 2011, the starting year of our dataset for model estimation (National Center for Science and Engineering Statistics, 2025). These calculations use a 5% patent premium. This choice is at the upper limit of estimates in Bessen (2008). If we use the lower estimate of 2.5%, total social costs increase to \$18.20bn, equivalent to 6% of total R&D expenditures.

Introducing a \$25,000 per-round fee reduces prosecution costs by discouraging applications and reducing padding. The fee has a moderate effect on type 1 costs, but it increases type 2 costs as applicants are more likely to abandon with some valid claims in a scenario with high negotiation fees. Total social costs decline by 8% with the per-round fee. If this fee raises extra revenue that is reinvested in more intensive (and hence more accurate) examination, then social costs would decline further (Schankerman and Schuett, 2022).

The impact of caps on the number of negotiation rounds is generally larger than that of the per-round fee. Restricting the process to one round reduces social costs by 37%. The non-overlapping confidence intervals confirm the statistical significance of these impacts. While the total social cost declines, type 2 costs rise with rounds restrictions. This increase may induce political opposition to such a reform from the patent community in the absence of some compensation, such as an adjustment to the R&D tax credit.

Reducing intrinsic motivation to 5% of its original level increases total social cost by 75%, and all components of social cost rise: there is no trade-off in this experiment. When examiners have almost no intrinsic motivation, they are willing to grant applications fast, even if the applications are substantially invalid. The resulting decrease in prosecution costs on each application is offset by the marked increase in the number of inventors applying for patent rights. Moreover, the willingness to grant patents with invalid claims increases type 1 costs almost fourfold. This finding confirms the importance of intrinsic motivation in this public agency.

Table 5. Net Social Costs of Patent Prosecution

Counterfactual	T_1 Cost	T_2 Cost	T_3 Cost	Total	Total C.I.
Baseline ($Bn)	3.01	0.25	12.11	15.38	[15.25, 15.43]
25,000 Round Fee	2.52	0.51	11.16	14.19	[13.80, 14.70]
Three Rounds	2.55	0.87	11.21	14.63	[14.58, 14.70]
Two Rounds	2.28	3.40	8.46	14.14	[13.93, 14.31]
One Round	0.40	5.74	3.61	9.76	[9.38, 10.15]
Credit$\searrow$	2.35	0.16	12.14	14.66	[14.62, 14.68]
5% IM	10.95	1.15	14.82	26.91	[25.87, 27.80]
Credit$\searrow$ + 5% IM	12.07	0.05	16.31	28.44	[27.86, 28.93]

Notes: "T_1 Cost" denotes total type 1 net social costs, "T_2 Cost" denotes total type 2 net social costs, and "T_3 Cost" denotes patent prosecution costs. "Total" sums the three costs. "C.I." refers to the 95% confidence interval. All numbers are measured in billions of 2023 USD. The table is based on $\lambda = 2$, $\frac{\rho^s}{\rho^p} = 2$, and $\Psi = 0.05$. Appendix Table A.6 provides results for $\frac{\rho^S}{\rho^P} = 1.5$ and patent premium $\Psi = 0.025$.

Finally, removing both intrinsic motivation and examiner credits beyond round one increases social costs by 6% relative to only removing intrinsic motivation. This result suggests that credits can be effective for screening when intrinsic motivation is low.

Before concluding, we highlight one qualification regarding our empirical quantification of the social costs of screening. While our positive analysis—specifically, the screening and development model—does not presuppose that the patentability threshold is at the optimal level, our specific calculations presented in this section do rest on this assumption.[27] To see this, suppose the prevailing threshold were lower than optimal, so that some patents are considered "valid" and granted, but should not be under the optimal threshold. In this case, we would understate type 1 error (some invalid claims would be incorrectly classified as valid) and thus understate type 1 costs. An analogous argument shows that, in this case, we would overstate type 2 costs. Whether the current patentability threshold is optimal remains an open research question we are pursuing.

[27]For discussion of the optimal level of patent eligibility, see Schankerman and Schuett (2022) Section 2.2.

8 External Validity

A natural question is whether our key findings about the efficacy of patent screening and the impact of counterfactual reforms in the U.S. apply to other leading patent offices, with different structures, negotiation rules, and incentives. As highlighted in the introduction, our objective in this research agenda is to make a quantitative comparative evaluation of institutional designs in innovation-supporting public agencies. Thus, there is no reason to expect our quantitative findings on patent screening in the U.S. to apply to other settings *unless* those settings have similar institutional structures and incentives.

In fact, other major patent offices (such as those in Korea, China, and Europe) do differ in their design. One distinction is who controls the termination of the application. In the U.S., the examiner can only terminate the process by granting a patent, since the applicant can request repeated re-examinations, subject to payment of fees. In Korea, the applicant is limited to one request for re-examination, after which the examiner can terminate. In contrast, in the European Patent Office (EPO) and China, the examiner has the right to issue a final decision with no re-examination (subject to appeal, as in all patent offices). Assigning these control rights affects the applicant's incentives to pad and to self-select into patenting in the first place. A second major difference is that the EPO uses a panel of three examiners, whereas the other major offices utilize a single examiner. In general, one would expect this feature to reduce type 1 and type 2 errors and their associated social costs, but any gains would, of course, have to be weighed against the higher administrative costs.

However, we would expect our *qualitative* findings about key counterfactual reforms to apply to other major patent offices because the economic logic behind them, which we provided in Section 6, is not specific to the context. For example, restricting the number of allowable negotiation rounds (or imposing equivalent fees for additional rounds) induces greater self-selection into patenting, leading to less padding and an increase in abandonment of valid claims. These outcomes reflect the change in applicant incentives created by the restrictions and should not depend on specific institutional features of the screening process, though their quantitative effects may well differ.

9 Conclusion

In this paper, we study the allocation of property rights for innovation by estimating a structural model of patent screening in the U.S. The model incorporates incentives, intrinsic motivation, and multi-round negotiation between the examiner and applicant. We adopt a new measure of patent distance using a natural language processing algorithm, which plays a key role in our

empirical analysis and illustrates how such methods can be usefully combined with structural modeling. We find that patent screening is moderately effective, given the prevailing standards for patentability within which the Patent Office must operate. Counterfactual analysis shows that restrictions on the number of allowable rounds of negotiation (or equivalent fees) significantly reduce the social costs of screening. We estimate the total social costs of patent screening at $15.38 billion per year, equivalent to five percent of enterprise-performed R&D in the United States.

There are various directions for future research. Given that our ultimate objective is to compare the effectiveness of different institutional designs, a first direction is to estimate patent screening models tailored to the other major international patent offices. Second, given the fast-moving frontier in natural language processing methods, it will be beneficial to update and improve the training of algorithms that measure patent distance, and to develop econometric techniques to address algorithmic error in such analyses. Third, our analysis of patent screening is conditional on the prevailing standard for patentability. An important challenge is to develop a methodology to identify the socially optimal patent standard. Finally, we believe there are opportunities to develop empirical frameworks to study innovation-supporting institutions such as the National Institutes of Health, the National Science Foundation, and similar institutions in other settings.

References

Adda, J. and M. Ottaviani (2024): "Grantmaking, Grading on a Curve, and the Paradox of Relative Evaluation in Nonmarkets," *The Quarterly Journal of Economics*, 139, 1255–1319.

AIPLA (2017): *American Intellectual Property Law Association 2017 Report of the Economic Survey*, Arlington, VA, https://www.aipla.org/detail/journal-issue/economic-survey-2017. Last accessed 9 April 2026.

——— (2019): *American Intellectual Property Law Association 2019 Report of the Economic Survey*, Arlington, VA, https://www.aipla.org/detail/journal-issue/2019-report-of-the-economic-survey. Last accessed 9 April 2026.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 132, 1553–1592.

Ash, E. and S. Hansen (2023): "Text Algorithms in Economics," *Annual Review of Economics*, 15, 659–688.

Ashraf, N., O. Bandiera, E. Davenport, and S. S. Lee (2020): "Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services," *American Economic Review*, 110, 1355–94.

Azoulay, P., J. S. Graff Zivin, D. Li, and B. N. Sampat (2018): "Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules," *The Review of Economic Studies*, 86, 117–152.

Battaglia, L., T. M. Christensen, S. Hansen, and S. Sacher (2024): "Inference for Regression with Variables Generated from Unstructured Data," cemmap Working Paper CWP10/24.

Benabou, R. and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *The Review of Economic Studies*, 70, 489–520.

——— (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

Besley, T. and M. Ghatak (2005): "Competition and Incentives with Motivated Agents," *American Economic Review*, 95, 616–636.

Bessen, J. (2008): "The value of U.S. patents by owner and patent characteristics," *Research Policy*, 37, 932–945.

Bloom, N., M. Schankerman, and J. Van Reenen (2013): "Identifying Technology Spillovers and Product Market Rivalry," *Econometrica*, 81, 1347–1393.

Cockburn, I. M., S. Kortum, and S. Stern (2003): "Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigation Outcomes," in *Patents in the Knowledge-Based Economy*, ed. by W. M. Cohen and S. A. Merrill, Washington, DC: The National Academies Press, 19–53.

Ely, J. C., J. Hörner, and W. Olszewski (2005): "Belief-Free Equilibria in Repeated Games," *Econometrica*, 73, 377–415.

Federal Trade Commission (2011): *The Evolving IP Marketplace: Aligning Patent Notice and Remedies with Competition*, Washington D.C.: Government Printing Office.

Feng, J. and X. Jaravel (2020): "Crafting Intellectual Property Rights: Implications for Patent Assertion Entities, Litigation, and Innovation," *American Economic Journal: Applied Economics*, 12, 140–81.

Foit, L. (2018): "Understanding the USPTO Examiner Production System," *Midwest IP Institute*.

Frakes, M. and M. Wasserman (2017a): "Replication data for: "Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents?: Evidence from Micro-Level Application Data"," https://doi.org/10.7910/DVN/ABE7VS.

Frakes, M. D. and M. F. Wasserman (2017b): "Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data," *The Review of Economics and Statistics*, 99, 550–563.

Galasso, A. and M. Schankerman (2015): " Patents and Cumulative Innovation: Causal Evidence from the Courts," *The Quarterly Journal of Economics*, 130, 317–369.

——— (2018): "Patent rights, innovation, and firm exit," *The RAND Journal of Economics*, 49, 64–86.

Ganguli, I., J. Lin, V. Meursault, and N. F. Reynolds (2024): "Patent Text and Long-Run Innovation Dynamics: The Critical Role of Model Selection," Working Paper 32934, National Bureau of Economic Research.

Graham, S. J., A. C. Marco, and R. Miller (2018): "The USPTO Patent Examination Research Dataset: A Window on Patent Processing," USPTO Economic Working Paper 2015-4, United States Patent and Trademark Office, http://dx.doi.org/10.2139/ssrn.2702637.

Hall, B. and J. Lerner (2010): *The Financing of R&D and Innovation*, vol. 1, Elsevier.

Hörner, J. and S. Lovo (2009): "Belief-Free Equilibria in Games With Incomplete Information," *Econometrica*, 77, 453–487.

Jaffe, A. and J. Lerner (2004): *Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to Do About It*, Princeton University Press.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021): "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 3, 303–320.

Khan, M. Y. (2025): "Mission Motivation and Public Sector Performance: Experimental Evidence from Pakistan," *American Economic Review*, 115, 2343–75.

Lanjouw, J. O. (1998): "Patent Protection in the Shadow of Infringement: Simulation Estimations of Patent Value," *The Review of Economic Studies*, 65, 671–710.

Le, Q. and T. Mikolov (2014): "Distributed representations of sentences and documents," in *International conference on machine learning*, PMLR, 1188–1196.

Li, D. and L. Agha (2015): "Big names or big ideas: Do peer-review panels select the best science proposals?" *Science*, 348, 434–438.

Lu, Q., A. Myers, and S. Beliveau (2017): "USPTO Patent Prosecution Research Data: Unlocking Office Action Traits," USPTO Economic Working Paper 10, United States Patent and Trademark Office, http://dx.doi.org/10.2139/ssrn.3024621.

Marco, A. C., J. D. Sarnoff, and C. A. deGrazia (2019): "Patent claims and patent scope," *Research Policy*, 48, 103790.

National Center for Science and Engineering Statistics (2025): "U.S. R&D Totaled $892 Billion in 2022; Estimate for 2023 Indicates Further Increase to $940 Billion," Publication NSF 25-327. https://ncses.nsf.gov/pubs/nsf25327. Last accessed 4 April 2026.

National Inventors Hall of Fame (2025): "Inductee List," https://www.invent.org/inductees/list. Last accessed 4 April 2026.

Pakes, A. (1986): "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54, 755–784.

Qiu, Y. J. J. (2023): "The Matthew effect, research productivity, and the dynamic allocation of NIH grants," *The RAND Journal of Economics*, 54, 135–164.

Rater, M., D. Ryman, and A. Toole (2020): "Examiner Credit Corrections by Patent Class," United States Patent and Trademark Office. Unpublished data. Aggregated values by technology center are included in the replication package.

Řehůřek, R. and P. Sojka (2010): "Software Framework for Topic Modelling with Large Corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Rubinstein, A. (1982): "Perfect Equilibrium in a Bargaining Model," *Econometrica*, 50, 97–109.

Sampat, B. and H. L. Williams (2019): "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome," *American Economic Review*, 109, 203–36.

Schankerman, M. and A. Pakes (1986): "Estimates of the Value of Patent Rights in European Countries during the Post-1950 Period," *Economic Journal*, 96, 1052–1076.

Schankerman, M. and F. Schuett (2022): "Patent Screening, Innovation, and Welfare," *The Review of Economic Studies*, 89, 2101–2148.

United States Patent and Trademark Office (2005): "Fiscal Year 2005 Performance and Accountability Report," Tech. rep., United States Patent and Trademark Office, https://www.uspto.gov/sites/default/files/about/stratplan/ar/USPTOFY2005PAR.pdf. Last accessed 9 April 2026.

——— (2013): "Setting and Adjusting Patent Fees During Fiscal Year 2013: Supplemental Table of Patent Fee Changes," Final rule under Section 10 of the America Invents Act, dated 18 January 2013. https://www.uspto.gov/sites/default/files/aia_implementation/AC54_Final_Table_of_Patent_Fee_Changes.pdf. Last accessed 4 April 2026.

——— (2023): "Fiscal Year 2023 Performance and Accountability Report," Tech. rep., United States Patent and Trademark Office, https://www.uspto.gov/sites/default/files/documents/USPTOFY23AFR.pdf. Last accessed 4 April 2026.

——— (2025): "PTAB Trials and Appeals Data," https://data.uspto.gov/ptab/trials/proceedings. Link last accessed 4 April 2026.

——— (2026a): "PatentsView Granted Patent Long Text Data," Product Identifier: PVGPATTXT. https://data.uspto.gov/bulkdata/datasets. Last accessed 3 April 2026.

——— (2026b): "PatentsView Pre-Grant Publication Long Text Data," Product Identifier: PVPGPUBTXT. https://data.uspto.gov/bulkdata/datasets. Last accessed 3 April 2026.