Screening Property Rights for Innovation*

William Matcham† Mark Schankerman‡

January 13, 2025

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 new measure of patent distance, which, together with detailed data on examiner decisions, enables us to estimate the model and thereby study strategic decisions by applicants and examiners. Our results show that patent screening is moderately effective, given the existing standards for patentability. Examiners exhibit substantial intrinsic motivation that significantly improves screening quality. We quantify the annual social costs of patent screening at $24.7bn, equivalent to 6.3% of U.S. private sector R&D. Reforms that limit negotiation rounds significantly reduce social costs.

Keywords: Patents, innovation, incentives, screening, intrinsic motivation

JEL Classification: D73, L32, O31, O34, O38

*We thank Jakub Drabik for excellent research assistance and contributions, especially on the patent distance metric. We thank Dietmar Harhoff, Marco Ottaviani, Florian Schuett, and seminar and conference participants at Tel Aviv University, Humboldt University of Berlin, Ben-Gurion University, CEMFI, University of Bologna, Hebrew University of Jerusalem, Boston University, CEPR IO Conference, and the Zvi Griliches Memorial Conference, for comments on earlier drafts of the paper. Martin Rater, Daniel Ryman, and Andrew Toole, at the U.S. Patent and Trademark Office, are thanked for assistance in obtaining part of the data. We also thank Janet Freilich and Michael Meurer for helpful discussions on procedural aspects of patent prosecution. This project was partly financed by a research grant from the Suntory-Toyota Centres for Economic and Related Disciplines at the London School of Economics. William Matcham also thanks the Sunwater Institute and the Edison Fellowship Program at George Mason University for financial assistance and the opportunity to discuss the project with legal scholars.

†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, such as patents, enhance innovation incentives for private sector R&D. In 2015, the U.S. federal government financed 36.9% of overall R&D expenditures, or $164.6 billion (in 2023 U.S.D.). At the same time, the U.S. Patent and Trademark Office (hereafter, *Patent Office*) issued almost 350,000 new patents. Intellectual Property 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 research investments and property rights for innovation 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-related 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 (The Economist, 2015; 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, where patents are granted to inventions that do not represent a substantial inventive step. These comments are particularly common with reference to 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 that reflects the actual patent application and examination process. An applicant is endowed with patent claims that are heteroge-

neous in their value and true distance to prior patents. These claims delineate the scope of the property right sought by the applicant. To craft the application, the applicant must choose how much to exaggerate the true scope of their claims. Exaggerating scope increases the potential returns but increases the risk of a lengthy and costly negotiation with the assigned examiner.

The patent examiner does not observe true claim distances and, before they can make a decision on the patent, must search the existing patent literature to gauge whether the submitted application represents a sufficient advance to warrant a patent. Through their search, the examiner obtains an error-ridden assessment of distances on the submitted application. In each stage of the multi-round negotiation that follows, the examiner first decides whether to grant or reject the patent application. Upon receiving a rejection, the applicant decides whether to abandon their application or continue the negotiation. Continuing the examination involves narrowing claims' scope, increasing distance but reducing 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 if they award more patent scope than is warranted by the invention. 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 Besley and Ghatak (2005) concept of *intrinsic motivation*—alignment of workers' objectives and the public agency mission.

In our general setup, the examiner would need to update their assessment of every claim's true distance and value on the basis of the applicant's actions in every negotiation round. Characterizing the equilibrium in this model is intractable, stymying any empirical implementation of the model. To make progress, we derive necessary and sufficient conditions on the functional forms of model components under which the examiner does not need to form beliefs over the underlying true distances and values of claims. This key 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 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 on patent claims over negotiation rounds. The decision dataset we use covers around 55 million decisions on 20 million patent claims between 2010–2015. The claim text data contains about 18 million claims granted in 1976–2014.

We apply modern natural language processing methods to the text data to develop a novel measure of distance between patent claims, which is a key ingredient in the model. The new distance

measure can be used 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." We derive formal conditions under which the computed threshold is a consistent estimator of the unobserved true threshold. Using the estimated patentability threshold, the model can quantify the extent to which invalid claims are granted (false grants, or "type 1" errors) and valid claims are not granted (false rejections, or "type 2" errors). This information is at the heart of evaluating screening effectiveness.

We conduct a series of external validation tests to confirm that our claim distance measure provides an informative signal. Nonetheless, since these "data" arise from the output of a neural network, we allow for the possibility of algorithmic error in distance observations. To address the potential for mismeasurement, we develop a methodology based on a multiple-indicators latent variable model to estimate the extent of measurement error in claim distances generated by the neural network (relative to the unobserved, properly measured distance). The methodology further allows us to purge the measurement error from our distance measure in preparation for its use in estimating our structural model. Since studies increasingly use natural language processing and more general AI methods to create input variables, our methodology should be useful more broadly.

The parameter estimates of the patent screening model imply four primary findings. First, patent screening is relatively effective, *given* the judicial standards of patentability 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 8% of granted claims are below the threshold. Still, nearly one in five granted patents contain at least one claim that does not meet the threshold, implying that type 1 errors do occur.

Second, inventors do exaggerate the scope of their invention in the initial patent applications. On average, this raises the value of claims by about 6%, but there is a lot of heterogeneity across applications. Moreover, since the decision on how much to exaggerate is endogenous in the model, it will change in response to reforms to the patent prosecution process. Our third finding is that the abandonment of valid claims is more common than the grant of invalid claims, with 20% of abandoned claims passing the threshold for patentability. This manifestation of imperfect screening has been largely ignored in the policy discourse. Finally, we estimate substantial, but heterogeneous, examiner intrinsic motivation. These estimates offer the first quantification of intrinsic motivation in a public agency, where one would expect worker motivation to be especially relevant, due to labor market sorting (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. The first is examination speed, as measured by the equilibrium number of rounds. The second and third capture screening errors, as measured by the frequency of granting claims that do not meet the patentability threshold and not granting claims that do pass the threshold. 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. Failures to grant valid claims dilute innovation incentives and discourage the development of new inventions that contribute positive social value.

A leading feature of our counterfactuals is that reforms typically generate a tradeoff between these type 1 and 2 errors: policies that make prosecution stricter lead to fewer grants of invalid claims but increased abandonments of valid claims. The policy conclusions from reforms are thus ambiguous and depend on the costs associated with both kinds of errors. To address this, we develop a methodology to quantify the social costs in the current environment and various counterfactual reforms. We estimate the total social costs of patent screening at $24.71bn per annual cohort of applications. This is equivalent to 6.3% of total R&D performed by business enterprises in the United States.

We find that restrictions on the number of negotiation rounds (absent in the current U.S. patent system) significantly reduce the social costs of screening—as much as 57% in the case of allowing only one round. Turning off intrinsic motivation increases the frequency of examiners granting invalid patents approximately four-fold, 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, in our quantification of social costs, we find that extrinsic incentives provided by examiner credits are largely ineffective. We interpret this result as an indication that the high levels of intrinsic motivation which we estimate leave little scope for extrinsic incentives.

Related Literature

We contribute to the literature analyzing how intrinsic motivation affects incentive design in mission-oriented agencies. Theoretical papers, notably Benabou and Tirole (2003; 2006), study how extrinsic rewards can crowd out intrinsic motivation. Besley and Ghatak (2005) emphasizes how intrinsic motivation—defined as the alignment between worker and agency objectives—induces welfare-improving sorting of workers across entities with different goals, and affects the

optimal design of incentives and authority. Empirical studies use field experiments to analyze intrinsic motivation and public agency performance, using various proxies for motivation. A leading example is Ashraf, Bandiera, Davenport, and Lee (2020), which finds that extrinsic rewards and intrinsic motivation are complementary. Our paper is the first *structural model* of a public agency that incorporates 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, including but not limited to the award of research grants. The paper examines how informational noise affects the optimal design of allocation rules. Li and Agha (2015) analyzes the allocation of research grants at the National Institutes of Health (NIH) and show 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.

We also contribute to the empirical literature on patent screening. In an early paper, Cockburn, Kortum, and Stern (2003) shows that patent examiner characteristics affect the quality of issued patents, as measured by subsequent citations and litigation. Frakes and Wasserman (2017) shows that when patent examiners are promoted, and resultantly obtain fewer credits for the same decisions, their grant rates rise sharply. This result suggests that examiner extrinsic incentives may affect the quality of screening. We analyze this hypothesis in a structural context.

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. The 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 this as an exogenous parameter—they do not model the prosecution process itself. By developing a dynamic equilibrium model of the patent examination process, in this paper we can study how reforms to the negotiation process and agents' incentives affect screening quality.

2 Model of the Patent Screening Process

We model the patent screening process as a dynamic game between an applicant, a, and an examiner, e, in technology area T. Both applicant and examiner are risk-neutral expected utility maximizers that discount future payoffs at the rate of β per period. The model features four stages: (1) Application Decision, (2) Examiner Search, (3) Negotiation, and (4) Patent Renewal. Figure 1 depicts the model extensive form. Section 2.1 describes the model's structure, agents'

payoffs, and the information structure.

Solving the general model requires a specification of the examiner's beliefs during negotiation, and finding a tractable way to handle the information subtleties of the unrestricted model is infeasible. In Section 2.2 we derive necessary and sufficient conditions on functional forms that remove the need for belief formation and thereby collapse the model to one that can be solved via backward induction. A brief analysis of the model under these restrictions follows in Section 2.3. We discuss and justify our functional form choices for the empirical implementation in Section 4.

Before the model description, we highlight two modeling choices. First, in a departure from most existing literature, we model patents as a collection of claims that are heterogeneous in private value and distance from previous inventions. The patent document is composed of independent claims that delineate the scope of the property right, and the examiner assesses the patentability of each claim separately. Second, we analyze patent screening conditional on invention development. The validity of the structural model and subsequent counterfactual analysis does not require a model of the potential applicant's decision to invest in developing their idea into an invention. Quantifying the social costs associated with the screening system, as we do in Section 7, will require us to add a model of the development decision.

2.1 Model Description and Information Structure

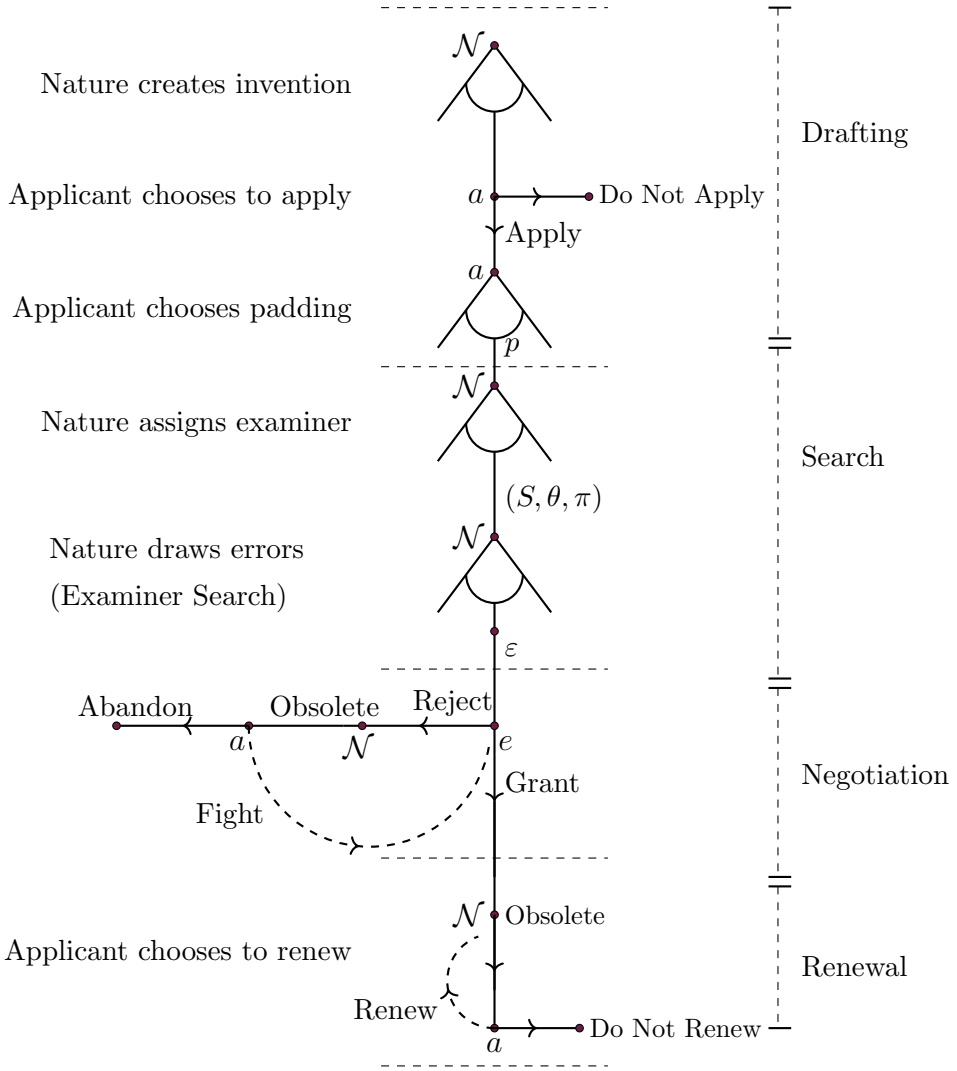
An applicant is endowed with an invention, comprising M_0 components, each of which can function as an independent claim on a patent application. We characterize each claim by the pair (D_j^*, v_j^*) where D_j^* is the distance of the true version of claim j to the nearest claim in any existing invention in the public domain ("prior art"), and v_j^* denotes the initial flow returns (or "value") generated by the true version of claim j once it is commercialized. We define the returns v_j^* as relative to the applicant's outside option, e.g., protecting the invention by trade secrecy.

Applicant Patenting Decision and Padding

The applicant first decides whether to apply for a patent. If they do not, the game ends, and their payoff is normalized to zero. Applying involves submitting 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 the true scope of the claims in the patent application. We call this choice the level of *padding*, denoted by p.[1] Padding obfuscates the true scope of the

[1]The applicant could choose to understate the true scope of the invention and thus earn lower returns, as it

FIGURE 1. EXTENSIVE FORM OF THE MODEL

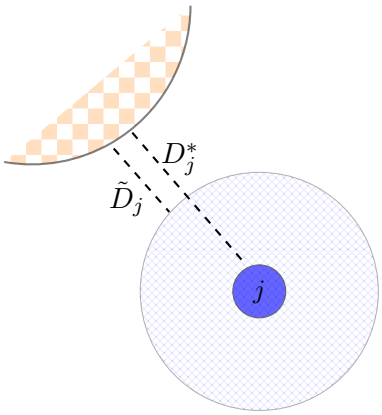


invention by concealing the true inventive step and thereby expands the property right. Figure 2 illustrates the concepts of claims and padding. We assume that the applicant pads all claims by the same proportion.

The basic tradeoff for the applicant in their padding choice is increased value versus increased risk of rejection. Padding raises the applicant's revenue but moves the application closer to the prior art and thus increases the likelihood of examiner rejections during the examination process on the grounds of non-obviousness (closeness to the prior art) and indefiniteness (clarity of description). To capture the tradeoff, we define initial padded (flow) returns to claim j, $\tilde{v}_j^1 = \mathcal{V}(p, v_j^*)$, which

reduces the likelihood of rejection by the examiner.

FIGURE 2. CLAIMS AND PADDING



Notes: 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 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 nearest point is $\tilde{D}_j$.

is an increasing function of v_j^* and p, and initial padded distance $\tilde{D}_j^1$ as a function increasing in D_j^* but decreasing in p. Finally, padding involves additional drafting work for the attorney and thus creates a direct cost to the applicant at the point of application, denoted $F_{app}(p)$. Applying also involves paying a Patent Office application fee.

Examiner Assignment

The Patent Office assigns each application randomly to an examiner within the relevant technology area (Sampat and Williams, 2019). We characterize an examiner by the tuple (S, θ, π). The first term S represents examiner seniority, which affects their payoffs for different decisions. The second term θ corresponds to the level of intrinsic motivation. Intrinsically motivated workers incur a disutility from awarding patent rights that do not meet the patentability standard based on the information available to them. The third term π corresponds to the examiner's "delay cost" of continuing to another negotiation round. Delaying creates pressure on examiners since they are evaluated in part based on effective and timely management of their portfolio of applications. As such, the delay cost also reflects the examiner's productivity.

Examiner Search and Grounds for Rejection

Once assigned to an application, the examiner searches the existing prior art to build an evidence

base to use throughout the application process. Their search allows them to assess whether the applicant's claims meet the requisite eligibility criteria. There are three main grounds for rejection: *novelty*, *non-obviousness*, and *indefiniteness*. Novelty requires that the claim has not been in use for one year before filing. Non-obviousness requires that the claim makes an inventive step beyond the closest existing invention 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 all claims in prior patents, and we characterize non-obviousness as satisfied if the distance is greater than a specified threshold (which will be estimated in Section 4.2.1). Indefiniteness requires that the claim is precise and clear on the exact boundaries of claimed property rights. We focus on novelty/non-obviousness (hereafter referred to as non-obviousness).[2]

After searching the prior art, the examiner assesses the obviousness of each claim. Their initial assessment of claim j's padded distance is given by $\hat{D}_j^1 = \mathcal{D}(p, D_j^*, \varepsilon)$ where ε denotes the drawn examiner error in assessing non-obviousness. More significant error means that the examiner over-estimates distance. The function $\mathcal{D}$ is strictly increasing in D_j^* and ε and decreasing in p. The error is independent of the true claim distance D_j^* and p, uniform across claims in the patent, and constant during the examination.

The examiner thinks they have *grounds* for rejecting claim j if their assessment of its distance is less than an obviousness threshold, τ. We note here briefly that examiners may end up granting claims they think they should reject. This is because, as we will explain in detail later, the examiner will make decisions to maximize their utility, where their utility depends both on their intrinsic motivation but also their extrinsic incentives. This is a crucial point as it implies that examiners' decisions in the data may not align with decisions made solely on legal grounds.

Finally, it is important to note that, as an institutional fact, the examiner's mandate is to reject a claim if it does not reflect a sufficient inventive step relative to prior art. We interpret this condition for rejection as equivalent to padded distance being below the patentability threshold. The examiner's job is not to reject claims on the basis of padding *per se*.

Structure of Negotiation

Once the examiner has made their initial assessment of claim distances, the game moves to the

[2] Using the Office Action Research Dataset described in Section 3.1, which provides the grounds for examiner rejections, we analyzed the overlap between novelty/non-obviousness (35 U.S.C. §102/103) and indefiniteness (§112) rejections. We find that 73% of office actions containing a 112 rejection also contain a 102/103 rejection. Thus, novelty/non-obviousness rejections cover most of the observed indefiniteness rejections, so omitting indefiniteness from the baseline model is a reasonable simplification.

negotiation stage. The Negotiation Stage is a finitely repeated version of the stage game shown in the "Negotiation" section of Figure 1. At round r (assuming it is reached), the examiner chooses whether to grant or reject the patent application as a whole. If granted, all constituent claims are awarded, and the game moves to the Renewal Stage. If the examiner rejects the application, they reject all claims for which the assessed distance is below the threshold.[3] In response to the rejection of the patent, the applicant then chooses whether to abandon the application or fight. Abandoning ends the game; continuing to round $r+1$ entails narrowing the rejected claims and incurring fighting costs along with any Patent Office fees. If the applicant and examiner reach the final round of the negotiation and the examiner rejects, the applicant must abandon.

Examiner's Decisions and Payoffs

At the start of round r, armed with new assessments of narrowed distances $\hat{D}_j^r$, the examiner decides whether to grant or reject the patent. The stage game payoff for granting in round r comprises two components: an extrinsic incentive and an intrinsic cost. The examiner's extrinsic incentive comes from *credits*. The Patent Office utilizes a point system that gives the examiner a specified number of credits for various decisions at each negotiation round. 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).[4] Bonuses are based on the extent to which the examiner exceeds their production target (falling short triggers reviews and potential penalties). This incentivizes examiners to maximize the credits they obtain.

Credits for decision y (e.g. granting, GR, rejecting, REJ) in round r, which we denote by $g_y^r(S,T)$, 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).[5] Credits modestly decline as the applicant enters subsequent rounds ($g_y^{r+1}(S,T) \leq g_y^r(S,T)$ for all y, S, and T), intended to provide an incentive for examiners to grant early. We provide the full schedule of credits in Appendix D.

[3] An alternative way of modeling the examiner's decision-making would be to assume that they decide which subset of claims to grant, rather than granting the patent, which implies granting all claims. Our approach dramatically simplifies what would otherwise involve a very large set of potential decisions.

[4] We abstract from examiners' inter-application incentives, 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 added complications.

[5] For example, the most junior examiner gets 2.5 times as many credits 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.

The second component to the examiner's granting payoff moderates the incentive to purely maximize credits and grant patents prematurely. Examiners face a potential intrinsic cost of granting claims they believe to be invalid based on their assessment of claim distances. Let M_r denote the number of claims in round r that the examiner thinks are invalid, i.e., the number of claims with $\hat{D}_j^r < \tau$. We denote by $\mathcal{R}(M_r, \theta)$ the examiner intrinsic cost from granting. This function satisfies $\mathcal{R}(0, \theta) = \mathcal{R}(M_r, 0) = 0$ and is weakly increasing in both arguments: for an examiner with any intrinsic motivation, granting a patent with claims they believe to be invalid goes against the organization's mission statement, which reduces the utility from granting.[6]

Put together, the stage game payoff to the examiner from granting a patent in round r is

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

It is apparent that in the face of declining credits over rounds, examiners with insufficient intrinsic motivation may grant patents containing claims they believe to be invalid. If the examiner rejects the application in round r, they obtain credits $g_{REJ}^r(S, T)$, and the negotiation continues. In this case, when reporting back to the applicant, the examiner rejects all claims on which they think there are grounds to reject. To clarify, the examiner rejects the application, they reject all claims j such that $\hat{D}_j^r < \tau$.

Applicant's Decisions and Payoffs

If the patent is granted in round r, the stage game payoff to the applicant is $V^r - \phi$, where ϕ is the issuance fee (part legal, part Patent Office) and V^r is the expected net returns from owning the patent, which is a function of the flow returns of each narrowed claim at the point of round r and the renewal decisions by the applicant after grant.

If rejected, the applicant decides whether to continue the negotiation or abandon the application. At this point the invention becomes obsolete with probability P_ω^{pre}, in which case the flow returns become zero permanently.[7] We denote the obsolescence state variable by ω_r, equal to one if obsolescence occurs in or before round r and zero otherwise. Formally, obsolescence is a Markov

[6] 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.

[7] The returns from patenting shrink to zero because the Patent Office publishes all applications, undermining appropriation of innovation rents by trade secrecy as an alternative.

process, where, for all r, if $\omega_r = 1$, then $\omega_{r+1} = 1$ (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.

If the applicant abandons, their stage game payoff is zero, and the examiner obtains credits $g^r_{ABN}(S,T)$. If the applicant chooses to fight, both the applicant and examiner incur fighting costs. Fighting costs for the applicant consist of two elements. The first is attorney fees for amending the application, F_{amend}. The second cost is a Patent Office fee, denoted F^r_{round}. These fees start in round three since entry into the third (and the fifth) round of negotiation requires that the applicant submit and pay for a "Request for Continued Examination" (RCE). If the applicant fights, the examiner pays delay cost π in the following period. Upon entering each RCE, the examiner receives credits $g^r_{FIGHT} > 0$.

If the applicant decides to fight, they must narrow their application. Narrowing involves reducing the padded distance (and, as a result, padded value). The extent of narrowing for each rejected claim in round r is exogenous and denoted by η_r.[8]

It is worth clarifying how narrowing works in practice and in the model. Examiners ask applicants to remove content that is too close to existing prior art, that is, to reduce *padded distance*. This is different from examiners asking applicants to remove padding explicitly. Crucially, examiners do not know (or care) about padding *per se*. The content they ask to be removed might be padding, but it might not be. The key point is that it does not matter.

For a general vector of narrowing terms collected across all potential rounds, denoted $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_R) \in [0,1]^R$, where R is the final round, we write the narrowed padded value of claim j as $\tilde{v}_L(p, v_j^*, \boldsymbol{\eta})$. This notation covers the case of a claim being granted in any round h, in which case $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_{h-1}, 0, ..., 0)$. The notation also allows us to write the padded value of any claim j at any point r during the negotiation phase (which we denote $\tilde{v}_j^r$) by setting all terms η_s to zero for $s \geq r$.

With this notation, the initial padded value before any narrowing is $\tilde{v}_j^1 = \tilde{v}_L(p, v_j^*, \mathbf{0})$. If claim j is then granted in round 1, $\tilde{v}_j^s = \tilde{v}_j^1$ for all $s \geq 1$. If claim j is not granted in round 1, its

[8] We could extend the model to allow the applicant to choose whether to narrow padded distance by proportion η with some probability or respond by arguing that the examiner is in error and not narrow at all. However, data on patent word counts imply that this extension is empirically unimportant. To see this, we look at word counts on 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 7% of the time, so we view the choice to ignore the possibility of no narrowing as a simplifying assumption in the baseline.

narrowed padded value in round 2 is $\tilde{v}_j^2 = \tilde{v}_L(p, v_j^*, \boldsymbol{\eta}_2)$ with $\boldsymbol{\eta}_2 = (\eta_1, 0, 0, \ldots, 0)$. A similar notation applies for the examiner's assessment of claim j's narrowed distance in round r. For assessments of distance given narrowing vector $\boldsymbol{\eta}$, we use the notation $\hat{D}_L(p, D_j^*, \varepsilon, \boldsymbol{\eta})$.

Patent Renewal

We enter the renewals stage of the model if the examiner grants the patent. Our renewal model adapts Schankerman and Pakes (1986) to the context of the U.S., adding a probability of post-grant obsolescence in addition to deterministic depreciation. Suppose the patent is granted in round r. The returns for each granted claim j start at $\tilde{v}_j^r$ and depreciate at rate δ each period after the grant. The value of the patent is the sum of the value across claims at any point in time. With probability P_ω^{post}, the invention becomes obsolete, at which point all returns shrink to zero permanently. To keep the patent rights, the applicant must pay renewal costs F_{renew}^4, F_{renew}^8, and F_{renew}^{12} at ages four, eight, and twelve after grant. These costs include an attorney fee and an Office fee. The renewal decision occurs at the patent level, not the individual claim level. If renewed for the full length, the patent ends 20 years after submission of the application, at which point the invention enters the public domain.

2.1.1 Information Structure

Applicant: At all stages, the applicant knows (i) the true distance of each claim D_j^*, (ii) the true private value of each claim v_j^*, (iii) the complete set of attorney and Patent Office fees, (iv) the full set of examiner credits across all rounds, (v) their choice of padding, (vi) the values of narrowing η_r the examiner will require if their claim is rejected; and (vii) the characteristics (S, θ, π) for all examiners in the roster and the distributions of examiner errors as a function of these characteristics.

Before applying for a patent, the applicant does not know which examiner will be assigned to their application and the search error that examiner will draw once assigned. After the examiner is assigned, the applicant can calculate their search error exactly through the examiner's report of their distance assessment. Finally, after applying, the applicant knows the current and all prior realizations of obsolescence but doesn't know future realizations.

Examiner: The assigned examiner does not observe true claim distances, true claim values, or the extent of padding 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, and (v) the *initial* padded value of all claims. Finally, all other parameters of the model are common knowledge to the applicant the examiner.

2.2 Simplifying the General Form

Under the given information structure, the applicant need not form beliefs about examiner characteristics or actions. The applicant knows all current payoff-relevant variables and can calculate all future payoff-relevant variables.[9] However, given that the examiner does not observe padding, true value, and 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. This feature results from the general functional forms of narrowed padded values and distance assessments. For example, for a general specification of $\tilde{v}_L$, without knowledge of p and v_j^* the examiner cannot forecast $\tilde{v}_j^2 = \tilde{v}_L(p, v_j^*, \boldsymbol{\eta})$ in the first round, even with knowledge of $\tilde{v}_j^1$ and $\boldsymbol{\eta}$. As a result, the examiner cannot predict the applicant's future actions that follow from their own decisions.

In the absence of restrictions on the functional forms of padded value and assessed distance, the examiner must formulate beliefs over padding, as well as true distances and values of all claims, and these beliefs would have to be updated at each round of negotiation based on the applicant's willingness to continue fighting. The general model described above thus presents serious informational challenges, and an empirical implementation of the general model is unlikely to be feasible.

Our key contribution to solve this problem is to derive conditions on functional forms $\tilde{v}_L$ and $\hat{D}_L$ under which the examiner does not need to form beliefs. Under such conditions, the examiner can use the observed values of $\tilde{v}_j^1$ and $\hat{D}_j^1$ to calculate future narrowed padded values and distance assessments *for any* narrowing vector $\boldsymbol{\eta}$, without knowledge of v_j^*, D_j^*, and p. Loosely speaking, observations of $\tilde{v}_j^1$ and $\hat{D}_j^1$ become "sufficient statistics" for the unknown terms. Upon imposing the restrictions, we can find the subgame perfect equilibrium by backward induction and formulate an empirical implementation of the model.

[9] We note two additional points here. 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 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.

To start, we state formally what it means for the examiner to be able to calculate all future narrowed padded values and distance assessments using initial observations only.

Condition 1. *For any vector of narrowing $\boldsymbol{\eta}$, the examiner can:*

- *Calculate all future narrowed padded values, using only the initial padded value, if there is a function W_v such that for all j and all $(p, v_j^*, \tilde{v}_j^1, \boldsymbol{\eta})$, we have $\tilde{v}_L(p, v_j^*, \boldsymbol{\eta}) = W_v(\tilde{v}_j^1, \boldsymbol{\eta})$.*
- *Calculate all possible future assessments of distance, using only their initial distance assessment, if there is a function W_D such that for all j and all $(p, D_j^*, \hat{D}_j^1, \varepsilon, \boldsymbol{\eta})$, we have $\hat{D}_L(p, D_j^*, \varepsilon, \boldsymbol{\eta}) = W_D(\hat{D}_j^1, \boldsymbol{\eta})$.*

The examiner does not require beliefs on applicant types if and only if Condition 1 holds. Our main proposition now follows. We assume that all functions in the proposition are continuously differentiable.

Proposition 1. *The examiner can calculate all future narrowed padded values and narrowed distance assessments, using only (1) narrowing, (2) the initial padded values, and (3) the initial distance assessments (i.e., Condition 1 holds) if and only if*

1.1 There exist functions $\tilde{v}_C(\cdot, \cdot)$ and $V(\cdot, \cdot)$ such that $\tilde{v}_C(V, \boldsymbol{\eta})$ is strictly increasing in its first argument at $\boldsymbol{\eta} = \mathbf{0}$, and for all (p, v_j^*)

$$\tilde{v}_L(p, v_j^*, \boldsymbol{\eta}) = \tilde{v}_C(V(p, v_j^*), \boldsymbol{\eta}) \tag{2}$$

1.2 There exist functions $\hat{D}_C(\cdot, \cdot)$ and $D(\cdot, \cdot, \cdot)$ such that $\hat{D}_C(D, \boldsymbol{\eta})$ is strictly increasing in its first argument at $\boldsymbol{\eta} = \mathbf{0}$ and for all (p, D_j^*, ε)

$$\hat{D}_L(p, D_j^*, \varepsilon, \boldsymbol{\eta}) = \hat{D}_C(D(p, D_j^*, \varepsilon), \boldsymbol{\eta}) \tag{3}$$

The necessity of the functional form restrictions follows from the definitions of initial padded value and distance assessment: $\tilde{v}_j^1 = \mathcal{V}(p, v_j^*)$ and $\hat{D}_j^1 = \mathcal{D}(p, D_j^*, \varepsilon)$. The sufficiency of the restrictions follows from the fact that the initial padded values (and distance assessments) contain no narrowing, so that $\tilde{v}_j^1 = \tilde{v}_C(V(p, v_j^*), \mathbf{0})$ and $\hat{D}_j^1 = \hat{D}_C(D(p, D_j^*, \varepsilon), \mathbf{0})$. Upon inversion, these deliver $V(p, v_j^*)$ and $D(p, D_j^*, \varepsilon)$ as functions of only initial padded value and initial padded distance assessment, respectively. This key step provides the intuition for why, among the class of functional forms satisfying Equation (2), the initial padded value serves as a "sufficient statistic" for the examiner to account for any such way that the applicant has combined padding with true value (similarly for Equation (3) and initial assessment of distance). The proof of Proposition 2 is in Appendix B.

The key assumption of Proposition 1 is the separability of the padded value function in (p, v_j^*) and $\boldsymbol{\eta}$, and similarly of the distance assessment function in (p, D_j^*, ε) and $\boldsymbol{\eta}$. The separability condition implies that the marginal rate of substitution (MRS) between true initial claim value and padding, for a given value of ε, in generating claim value, is independent of subsequent narrowing. Similarly, the condition implies that the MRS between true claim distance, padding, and examiner error, in generating assessed distance, is independent of narrowing.

While not needed for Proposition 1, it is reasonable to assume that the $\mathcal{V}$ function exhibits supermodularity (complementarity) in padding and true value. Our functional form choices for the empirical specification, which satisfy the proposition conditions and exhibit supermodularity, write the padded value of a claim narrowed r times as $pv_j^* \prod_{s=1}^{r}(1-\eta_s)$ and the assessed distance at that point as $D_j^* \varepsilon \left(p \prod_{s=1}^{r}(1-\eta_s)\right)^{-1}$. In the empirical implementation in Section 4, we elaborate on the theoretical and empirical advantages of the multiplicative formulation and offer extensive robustness analysis to alternative choices.

2.3 Analysis of the Model

Renewal Decisions

We characterize equilibrium path actions, starting at the end of the model with the renewal decisions. The applicant decides whether to renew based on the expected returns from retaining patent rights. The expected returns of holding the patent from age (years after the patent is granted) t_1 to t_2 is[10]

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

Suppose the application is granted in round r. Then, conditional on not becoming obsolete, the applicant renews at age 12 if $V_{12}^r \equiv \mathbb{E}_{\boldsymbol{\omega}} V_{12,20-r} - F_{\text{renew}}^{12} > 0$.[11] The applicant renews at age eight if $V_8^r \equiv \mathbb{E}_{\boldsymbol{\omega}} V_{8,11} - F_{\text{renew}}^{8} + I_{12}\beta^4 V_{12}^r > 0$, where I_t is equal to one if the applicant will renew at year t and zero otherwise. An analogous decision rule holds for renewal at age four. Finally, we define the ex post net expected benefits from patent rights, when granted in round r, as

$$V^r = \mathbb{E}_{\omega} V_{1,3} + I_4 \beta^4 V_4^{8,r}. \tag{4}$$

[10] We use the notation $\mathbb{E}_{\boldsymbol{\omega}}$ to denote expectations taken over the vector of obsolescence shocks that are not yet realized. With a slight abuse of notation, whenever we use $\mathbb{E}_{\boldsymbol{\omega}}$ with an emboldened $\boldsymbol{\omega}$, it refers to the sub-vector of $\boldsymbol{\omega}$ that have not yet occurred. The notation $\mathbb{E}_{\omega_r}$ refers to an expectation over ω only in round r.

[11] These decision rules differ from those in patent renewal studies in Europe (Schankerman and Pakes, 1986), which require annual payments. The fact that the decision intervals are not symmetric creates a more complicated, non-stationary decision rule.

Negotiation Decisions

Let x_a^r and x_e^r denote the actions by 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, $\mathcal{G}$ 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, with expectation taken over obsolescence outcomes.

The value function for the applicant *upon being rejected in round* r, W_a^r is defined as follows. If the invention becomes obsolete, so that $\omega_r = 1$, $W_a^r = 0$. Otherwise, we have $W_a^r = \max\{0, \mathcal{U}_{\text{Fight}}^{r+1}\}$, where

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

$1(A)$ is the indicator function, equal to one if statement A is true and zero otherwise, and V^{r+1} defines the ex post, net expected 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 $\mathcal{U}_{\text{Fight}}^{r+1} > 0$).

Finally, we can define the expected utility for the applicant, *before applying*, for a given examiner, error, and choice of padding as

$$\mathcal{Z}_a^0(e, \varepsilon, p) = 1(x_e^1 = \text{GR})[V^1 - \phi] + 1(x_e^1 = \text{REJ})\mathbb{E}_{\omega_1} W_a^1, \tag{5}$$

where terms on the right-hand side are (implicitly) functions of the level of padding, error and examiner.

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, where the expectation is taken over the roster of potential examiners $e = 1, \ldots, \underline{E}$ (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, $\Gamma(p)$, defined as

$$\Gamma(p) = \frac{1}{\underline{E}} \sum_{e=1}^{\underline{E}} \int \mathcal{Z}_a^0(e, \varepsilon, p)\, dG_{e,\varepsilon}(\varepsilon) - F_{\text{app}}(p).$$

Finally, the applicant applies if the expected utility of the subsequent negotiation game is positive, that is, if

$$\Gamma^* \equiv \Gamma(p^*) \geq 0. \tag{6}$$

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 U.S. patent claims granted between 1976 and 2014. This dataset covers about 18 million patent claims. In Section 3.2, we describe how we use these data to train an algorithm to construct a measure of claim distance.

Prosecution Rounds Data: Since we estimate a model of the patent prosecution process over multiple rounds, we require comprehensive round-level data on the patent process. We use the *Transactions History* data in the *Patent Examination (PatEx) Research Dataset* to create a dataset on the round-by-round evolution of patent applications between 2007 and 2014. For every patent application, these data contain examiner and applicant decisions at each round of the examination process.

We match the round-level data to three 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 seven technology classifications used by the USPTO, called "technology centers." Second, we match the data to renewal decisions by patent holders using the *USPTO Maintenance Fee Events Dataset.* Third, since we focus on novelty/obviousness rejections, we require data on the types of rejections of each claim. This is available from the *USPTO Office Action Research Dataset for Patents.* We use applications in these data between 2011-2013. Once we merge datasets, we obtain a sample covering around 55 million claim-round decisions on 20 million independent claims.

Legal Fees: We use data from the *2017 American Intellectual Property Law Association (AIPLA) Report of the Economic Survey.* The survey reports statistics of the distribution of hourly attorney fees for different tasks, such as preparing and filing an application, paying renewal fees, and amending applications. The statistics are split by three broad technology areas (biotechnol-

ogy/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 (2017), which provides a panel of General Schedule (GS) grades for examiners over time. Using this, we calculated the seniority of the examiner for each application. Finally, we obtained (unpublished) information on examiner credits from the Patent Office at the highly disaggregated US patent classification level. We then aggregated them to the technology centers in our data.

Patent Challenges and Outcomes: The Patent Trial and Appeal Board (PTAB), established in 2012, is an administrative mechanism within the USPTO that serves as a second layer of post-grant screening. Third parties initiate challenges (called inter-partes reviews), which are adjudicated by a panel of senior examiners. A subset of PTAB challenges make it to final trial, with many settled beforehand. We extracted data on trial outcomes from the unstructured text using the OpenAI GPT-4 API. The dataset covers about 12,000 challenges.

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. We summarize our approach to creating a distance measure here, with more detail in Appendix E.

The approach calculates distances between claims by representing the text of each patent claim as a numerical vector and calculating a metric on that vector space. Previous studies use variations of the standard *bag-of-words* method for representing the 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 the *ordering* and the *semantics* of words. Word overlap is particularly troublesome in the context of patent applications since attorneys strategically attempt to describe the invention differently from existing 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 the corpus of patent claim texts between 1976 and 2014) 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. As is common in the natural language processing literature, we measure distances between word vectors using the angular distance metric. To reflect the distance to the *prior art*, we compute the distance from each claim to every previously granted claim. The

Table 1. Summary Statistics

Variable	Mean	Median	S.D.	5%	95%
Application Granted	0.70	1.00	0.46	0.00	1.00
Years of Prosecution	2.96	2.67	1.57	1.01	5.95
Negotiation Rounds	2.39	2.00	1.45	1.00	5.00
Independent Claims	2.99	3.00	2.93	1.00	7.00
Small Entity	0.24	0.00	0.43	0.00	1.00
Not Renewed at Age 4	0.13	0.00	0.33	0.00	1.00
Full Renewal	0.46	0.00	0.50	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.

relevant distance measure for each claim, as in the model, is between the focal claim and the nearest previously granted claim.

3.3 Descriptive Statistics

Table 1 presents summary statistics of the data. First, 70% of applications result in the issuance of a patent. Second, the mean duration of patent prosecution is 2.96 years, and the mean number of rounds is 2.39, but there is substantial variation, with the coefficients of variation of both exceeding 50%. This fact implies that some applications involves lengthy negotiation between applicant and examiner. Third, 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 are rejections. In first-round rejections in our data, approximately 78% of an application's independent claims are rejected on the grounds of obviousness, novelty, or indefiniteness. The most common outcome in the first round is for the examiner to reject *all* claims, but 12% of applications are granted in the first round, in which case the examiner rejects no claims. This bimodality does not vary across technology centers, with 96% of the variation in *round-one* rejection rates found 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, where claim distances are broadly similar. Further, the distribution of claim distances and rejection rates by round are similar across technology centers. About 98% of the variation in distances is across patent applications within a technology center.

While most applications have three independent claims at publication, the number of claims varies significantly, with a 95th percentile of 7.00. Again, technology centers do not differ much in this dimension: 98% of the variation in the number of claims is within technology centers. In the sample, 24% of applications were filed by firms with fewer than 500 employees (a so-called "small entity"). Lastly, 46% of granted patents were renewed to the statutory limit, and only 13% were not renewed at the first renewal date (age four).

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, corroborating that senior examiners grant more frequently (Frakes and Wasserman, 2017). Also, we find that the frequency of multi-round negotiation is much lower for senior examiners and varies across technology centers. In addition, small entities are less likely to negotiate. Finally, we decompose the variation in examiner-specific outcomes (such as their grant rate) within and between technology center-seniority pairs. The results imply that 80% of the variation in examiner grant rates and 81% of the variation in each examiner's average number of rounds is within technology center-seniority pairs.

Taken together, these descriptive findings highlight the heterogeneity in the sample (in distances to prior art, number of claims, examiners' rejection rates, etc.) and confirm that the predominant variation in outcomes is within technology area, rather than between. Our empirical implementation of the model shall incorporate sources of patent- and claim-level heterogeneity to explain these variations by technology area that we find in the data.

3.4 External Validation of the Distance Measure

Since there are no underlying true observations for patent distance, we cannot check the performance of our distance algorithm on an out-of-sample test set. Instead, in this section, we illustrate our distance measure's ability to produce intuitive results in contexts that are entirely external to its construction.

National Inventors Hall of Fame (NIHF): The USPTO and National Inventors Hall of Fame jointly maintain a record of over 600 inventors whose patented inventions represent substantial technological achievements. Since NIHF patents are supposed to represent significant technological advances, their distance to prior patents should be greater. To test for significant differences between NIHF and non-NIHF claim distances, we collected distance data for all claims in NIHF patents between 1976 and 2014 (the range of our primary sample of distance data). A regression of the log of claim distance against a dummy for NIHF, along with grant-year and technology-center dummies, shows that distance is 14% larger for claims in NIHF patents ($p < 0.001$). Empirical

CDFs, unfiltered and filtered by grant-year and technology-center dummies, show that NIHF claim distance stochastically dominates non-NIHF claim distance.

Selection Into PTAB Challenges: The second validation test compares the distance to prior art for patents challenged in the PTAB to those not. On average, claims in PTAB cases should have smaller distances to prior art than those not involved in challenges. To test this, we consider a sample of granted patents applied for between 2011 and 2013, for which we can compute a distance measure and see whether there was a PTAB challenge. Regressing the log of claim distance against a dummy for a PTAB challenge and fixed effects for grant-year and technology-center, we find that claims in PTAB challenges are 6% closer to prior art, relative to those not challenged ($p < 0.001$). Results are also robust to aggregation up to the patent level: the minimum and mean (over claim) distance is 5–6% smaller for patents in PTAB challenges relative to those not challenged.

In our presentation of the latent variable model in the next section, we provide a third external validation test. We show that patent claims with a larger distance to prior art are also more likely to be upheld in PTAB challenges that reach a final decision.

3.5 Purging Measurement Error in the Distance Measure

The findings of the previous section show that our distance measure contains some signal on the actual (padded) distance between claims and the prior art. Nevertheless, we acknowledge that our distance measure could still contain measurement error. This feature is not specific to our context; it applies to any study that uses AI-based outputs as data, though the literature is yet to address this issue constructively.[12] In this section, we develop a general methodology to estimate the magnitude of the error and to purge the distance measure of such error. Our approach adopts a latent variable model (LVM) that relates multiple observed indicators to an underlying latent variable that represents perfectly measured padded distance.[13] This approach can be applied to quantify and purge measurement error in other studies that use AI-based outputs used as data.

We consider the specification

$$y_{kj} = \xi_k \tilde{D}_j + \Theta_k' X_{kj} + e_{kj}$$

where y_{kj} is the value of the k^{th} indicator for claim j, $\tilde{D}_j$ is the correctly-measured padded

[12] In a recent paper, Battaglia, Christensen, Hansen, and Sacher (2024) shows that using estimates from upstream "information retrieval models" as data in downstream econometric models can lead to incorrect inference. The paper also derives a one-step strategy that can reduce bias, but it is not adaptable to our context.

[13] For early applications of latent variable models in the innovation literature, see Pakes and Schankerman (1984) and Lanjouw and Schankerman (2004).

distance to the closest claim, which is unobservable, ξ_k is the factor loading for indicator k, and X_{kj} is a vector of controls, where all observable variables are de-meaned. Since $\tilde{D}_j$ is unobservable, we normalize $\xi_1 = 1$ (the choice of normalization is inconsequential).

Our LVM includes three indicators. The first two are continuous variables: (i) the *closest* claim distance, $y_{1j} = D_j^1$, as generated by the algorithm, and (ii) the *fifth-closest* claim distance, $y_{2j} = D_j^5$. Both indicators will contain some signal on the true closest distance because the algorithm may incorrectly *rank* which previously granted claims are actually closer. This explains why we would want to use multiple distance indicators. We assume that e_{1j} and e_{2j} are not correlated.[14]

The third indicator, y_{3j} is a dummy variable denoting whether the PTAB upholds a claim; this binary indicator is modeled using a Probit specification so that y_{3j}^* is the utility of upholding a claim in the PTAB, and $y_{3j} = 1(y_{3j}^* > 0)$. The sample size is limited by the smaller PTAB outcome subsample: around 1,500 patents, or 6,800 claims, make it to a final decision. We estimate the model by maximum likelihood, assuming joint normality of errors.[15] The three-indicator model is exactly identified under the assumption of uncorrelated errors. We estimate a constrained version of the LVM with equal error variances for D^1 and D^5 (we cannot reject this restriction, p-value = 0.45). The estimated error variances allow us to measure the noise ratio for the distance indicators.

Our estimates imply a noise ratio in the distance measure of 9%, confirming the presence of modest measurement error. The estimated factor loading for $\tilde{D}$ in the PTAB outcome equation is $\hat{\xi}_3 = 0.47$ (s.e. = 0.22). This estimate confirms that patent claims more distant from prior art are significantly more likely to be upheld. Indeed, the average marginal effect of distance is substantial—claims with 10% higher distance have a 5% higher probability of being upheld. This result is the third external validation of our distance measure.

Under joint normality of the latent variable and two distance indicators, we can use the estimated factor loadings to compute the posterior mean of the latent variable, conditional on the two

[14] In the LVM, *both* D_j^1 and D_j^5 are projected onto the true distance to the *closest* claim $\tilde{D}$ (the object of interest for the model), not onto their respective true distances as would be the case in a pure measurement error model. Hence, in addition to algorithmic error, the terms e_{1j} and e_{2j} capture the projection error coming from our specification of the LVM. This fact has two implications. First, the variance of the composite error is an upper bound of the pure algorithmic error that we want to measure and then purge. Second, even if algorithmic errors for both equations were positively correlated, as one might expect, this does not necessarily translate into a particular correlation in the LVM errors e_{1j} and e_{2j}. Appendix Figure A.2 illustrates the latter point.

[15] For the PTAB equation, we include fixed effects for challenge year and technology center interactions with a dummy for small patentees (fewer than 500 employees).

distance indicators. Specifically, the assumption

$$\begin{pmatrix} \tilde{D} \\ D^1 \\ D^5 \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

implies $E(\tilde{D} \mid D^1, D^5) = \Sigma_{12}\Sigma_{22}^{-1}(D^1, D^5)'$. This conditional expectation provides an estimate of (de-meaned) padded distance purged of measurement error. While we use the third indicator to estimate the factor loadings, its binary form means that we cannot condition on it in computing the posterior mean of $\tilde{D}_i$. We use $E(\tilde{D}_i \mid D^1, D^5)$ as our measure of distance that is purged of algorithmic error in all empirical analysis that follows. However, all results that follow are robust to using the raw distance measure.

4 Empirical Implementation of the Model

4.1 Functional Forms and Distribution Choices

This section describes our distributional and functional form choices. Given the complexity of our setup, we do not consider nonparametric identification, but discuss the sources of identification of parameters in detail in Section 4.3. Appendix Table A.2 summarizes all parameters and their associated distributional assumptions.

Functional Forms

Before we provide our functional form choices, it is important to note that some key objects of interest—true distance, true value, and padded value—are not observable. This means that we cannot ground our choice of functional forms on the basis of external data on these objects. Instead, we conduct extensive robustness checks to our functional form choices, detailed after estimation in Section 5.3.

Padded distance: By Proposition 1, we require a function mapping true distance, padding, and examiner error into assessed padded distance, $\hat{D}(p, D_j^*, \varepsilon)$, and a function mapping the examiner's distance assessment and narrowing into a narrowed padded distance assessment. For the former, we first specify padded distance as $D_j^* p^{-1}$ in the baseline. This choice defines padding as a proportional reduction in distance to prior art. We then take examiner error as multiplicative, so that $\hat{D}(p, D_j^*, \varepsilon) = D_j^* \varepsilon p^{-1}$. Thus, we interpret the error as a proportion of initial padded distance (e.g., 1.1 represents a 10% positive error).

We use the form $D_j^* \varepsilon \left(p \prod_{s=1}^{r}(1-\eta_s)\right)^{-1}$ for the distance assessment after r degrees of narrowing. This form is weakly separable in assessed distance and narrowing, as required by Proposition 1. In the baseline model, we use constant narrowing over rounds, $\eta_s = \eta$ for all s, simplifying the

expression to $D_j^* \varepsilon \left(p(1-\eta)^r\right)^{-1}$. Our results are robust to more demanding specifications that allow narrowing to differ across rounds (see Section 5.3).

Padded value: For function $\mathcal{V}$ that links padding and true claim value to create padded claim value, we start with a Cobb-Douglas form: $\tilde{v}_j = \mathcal{V}(v_j^*, p) = p^{\zeta}(v_j^*)^{\Upsilon}$. [16] First, because padded value is unobservable, we can rescale it and use $pv_j^{*\chi}$ where $\chi = \frac{\Upsilon}{\zeta}$. Second, 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.[17] 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.

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 they believe to be invalid. 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. We also use an alternative $\mathcal{R}(M_r, \theta) = \theta M_r$. There are no theoretical grounds for preferring one over the other specification, but using the proportion fits the data better.

Cost of padding: The cost of padding in terms of attorney fees is increasing and symmetric in the absolute value of padding. We specify $F_{app}(p) = f_{app}(1+ \mid p-1 \mid)$ where f_{app} is the application drafting fees per unit of padding. The idea is that it takes more time for the attorney to draft an application meeting the full set of patentability standards when there is either positive or negative padding.[18]

Distributions

Applicant: We start by discussing true distances D_j^* and true unpadded flow returns v_j^*. We assume that D_j^* and v_j^* are independent. Given that both D_j^* and v_j^* are unobservable, we cannot estimate a correlation between them and test this simplifying assumption. It is worth noting though, that the theoretical literature on imperfect competition with differentiated products does not predict a clear relationship between these two characteristics of the applicant's claims.[19]

[16] We estimated a more general CES function and the estimates imply an elasticity of substitution of 0.94, which is very close to the Cobb-Douglas value of 1.

[17] 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 distribution contain χ.

[18] 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 indefiniteness requirement. Understating the patent scope involves deciding what elements to drop while still adequately revealing the invention to allow replication (as mandated by the "enablement requirement").

[19] Greater distance to rivals in product and technology market space is likely to soften price competition and increase private value, which suggests a positive relationship. However, the distribution of demand will typically

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 Appendix Table A.2). Motivated by Schankerman and Pakes (1986), the log of initial claim flow returns is normally distributed with mean μ_v and variance σ_v^2. Finally, application drafting legal fees per unit padding, f_{app}, are log-normally distributed with parameters $\mu_{f,\text{app}}$ and $\sigma_{f,\text{app}}$, with different parameters for simple and complex applications in chemical, electrical, and mechanical fields.

Examiner: Intrinsic motivation is log-normally distributed, and we allow for different μ parameters for junior pre-GS-14 grade examiners (μ_θ^J) and senior examiners (μ_θ^S). We constrain the σ_θ parameter to be the same for juniors and seniors, but this does not imply equal variances. In the baseline, we treat examiner delay costs, π, as constant but allow them to vary by round and seniority in robustness analysis.

Examiner errors are normally distributed with mean and variance that depend inversely on the degree of intrinsic motivation.[20] We microfound the relationship between moments of examiner search errors and intrinsic motivation in Appendix C; the basic idea is as follows. The examiner chooses how intensely to search 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 optimal search intensity rises, and errors decline, with intrinsic motivation. As a result, intrinsic motivation affects endogenous outcomes through two channels: (i) making strategic errors more costly in the examiner payoff function, and (ii) intensifying optimal search. Consistent with the microfoundations, we specify the error mean as $1 + (\varrho\theta)^{-1}$ and the variance as $\sigma_\varepsilon^2(\theta)^{-1}$. Because θ is log-normal, we set $\varrho = 1$ without loss of generality.[21]

4.2 External Parameters

Discount Rate (β): We follow standard practice in the literature and set $\beta = 0.95$. Estimates are robust to modest variations in this parameter.

Applicant Fighting Costs (f): We have data on the quantiles of the distributions of *amendment*,

vary with location with firms endogenously locating in areas of high demand. This would suggest a negative correlation between distance to rivals and value.

[20] In cases where the draw of the error leads to assessments outside the $[0, 1]$ interval, we truncate the distances assessments at the endpoints.

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

maintenance, and *issuance* hourly fees charged by lawyers. We assume these three costs are log-normally distributed. 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 mean and variances of the log of fighting costs using two-step generalized method of moments estimation procedure for each of these three types of attorney costs.[22]

Depreciation of patent flow returns (δ): Bessen (2008) estimates the combined effect of depreciation and the probability of obsolescence at 0.14, using U.S. renewal data. In our context, this corresponds to $(1-P_{\omega}^{\text{post}})\cdot\delta+P_{\omega}^{\text{post}}\cdot 1$. Hence, for each iteration of P_{ω}^{post} in our model estimation, we extract the implied pure depreciation rate δ from this relationship.

4.2.1 Estimating the Distance Threshold

We first describe our estimator of the distance threshold $\hat{\tau}$ and then, through Proposition 2, provide the formal conditions under which the estimator is consistent for the true threshold τ. Estimation is external to the model, using observations on claim distances and examiner grant decisions only.

For every examiner e, we calculate the minimum value of the distances among claims they grant in all years in our sample. This is denoted 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. We estimate the distance threshold as the maximum of τ_e over examiners, i.e., $\hat{\tau} = \max_e \tau_e$. The intuition for this estimator is as follows. If an examiner is perfectly intrinsically motivated and does not make errors, their "revealed threshold" τ_e will, for a large number of applications for each examiner, converge to the true threshold τ. However, for all other examiners, their "revealed threshold" will eventually be below τ. Hence, under specific assumptions, the maximum of "revealed thresholds" will converge to τ as the number of applications converges to infinity.

In Proposition 2, we state formal conditions under which the maximum "revealed threshold" across examiners converges to the true threshold. The key condition is that we can find an examiner whose intrinsic motivation is unboundedly large. As a result of the inverse relationship between intrinsic motivation and search errors, this examiner will not make errors and will not grant below the true threshold.[23]

[22] We cannot estimate the distribution of *application* fighting costs externally because these are proportional to padding in the model so our measure of it is contaminated by the endogenous choice of padding. Thus, we estimate the parameters of application fighting costs as part of the simulated method of moments procedure described in the next subsection.

[23] Though in a different context, the key condition is reminiscent of the "one smart agent" assumption in Sutton

For the proof of threshold validity, let A_e denote the number of examinations conducted by examiner e and, to simplify exposition, suppose that $A_e = A$ for all e.

Proposition 2. *Suppose that the following conditions hold:*

2.1 The number of claims on any application is bounded above (i.e., there exists $\bar{M} > 0$ such that $M_{0a} \leq \bar{M}$ for all applications a).

2.2 For every application a, $\varepsilon_a \sim \mathcal{N}(1 + \mu(\theta), \sigma(\theta)^2)$ with $\mu(\theta)$ and $\sigma(\theta)$ converging to 0 as θ converges to infinity.

2.3 There exists an examiner with unboundedly large intrinsic motivation, i.e., for all positive constants A and H, and arbitrarily small δ, there exists an examiner with θ such that

$$A\bar{M}\left[1 - \Phi\left(\frac{H - \mu(\theta)}{\sigma(\theta)}\right)\right] < \delta \tag{7}$$

Then $\hat{\tau} \xrightarrow{P} \tau$ as $A \to \infty$.

The full proof of this result is in Appendix B. Condition 2.1 is standard in proofs of consistency. Condition 2.2 requires normality of examiner errors, a functional form assumption we imposed already in Section 4.1. Furthermore, the requirements on the mean and variance of examiner error are consistent with the microfoundations and our empirical implementation.

We calculate the 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 in different technology centers are similar, ranging from 0.47 to 0.50.

4.3 Model Estimation and Identification

We start by summarizing the model variables that are observable in the data. Relating to 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 threshold). We do not observe examiners' intrinsic motivation, delay costs, or distance assessment errors.

(1997). An alternative formulation would be to conduct the asymptotics in the number of examiners. Then, the continuity of intrinsic motivation will ensure that there exists such an examiner.

With our observed data, the model parameters left to estimate are given by the vectors $\boldsymbol{\psi}_a = \left(\eta, P_\omega^{\text{pre}}, P_\omega^{\text{post}}, \alpha_D, \gamma_D, \mu_v, \sigma_v, \boldsymbol{\mu}_{f_{\text{app}}}, \boldsymbol{\sigma}_{f_{\text{app}}}\right)$ and $\boldsymbol{\psi}_e = \left(\mu_\theta^J, \mu_\theta^S, \sigma_\theta, \pi, \sigma_\varepsilon\right)$, for the applicant and the examiner, respectively. We estimate using simulated method of moments (SMM), solving

$$\min_{\psi} \left(\boldsymbol{m}(\boldsymbol{\psi}) - \boldsymbol{m}_{\mathcal{S}}\right)' \Omega \left(\boldsymbol{m}(\boldsymbol{\psi}) - \boldsymbol{m}_{\mathcal{S}}\right),$$

where $\boldsymbol{m}(\boldsymbol{\psi})$ is the vector of simulated moments computed from the model when the parameter vector is $\boldsymbol{\psi} = (\boldsymbol{\psi}_a, \boldsymbol{\psi}_e)$, $\boldsymbol{m}_{\mathcal{S}}$ is the vector of corresponding sample moments, and Ω is a weighting matrix.[24]

The number of available moments in the model far exceeds the number of parameters. To select a subset of moments for estimation, we followed a data-driven methodology based on the sensitivity of parameter estimates to the inclusion of particular moments (described below), along with plots of how the estimated model moments and the value of the SMM objective vary with parameter values. We describe this procedure in detail in Appendix F. The procedure pruned the set of moments down to 44 that assist in estimating the parameters. An advantage of pruning is that after estimation we can assess model fit using the moments not used in estimation.

The selected moments corresponding to outcomes for examiners are the proportion of applications granted by seniority and round, the standard deviation of examiner rejection rates by seniority, the means and standard deviations of type 1 errors by seniority, and the proportion of patents granted containing an invalid claim by seniority and round. The moments corresponding to outcomes for applicants are the proportion of abandonments by seniority and round, renewal rates, means and standard deviations of granted claim distances by grant round, and means and medians of legal application fees by technology class.

Sensitivity Analysis

As is typical for complex nonlinear models, we cannot prove the point identification of our parameters. Instead, we adopt the empirical approach developed by Andrews, Gentzkow, and Shapiro (2017). They propose a sensitivity matrix that quantifies how a change in the value of a moment affects the estimate of a parameter. In our case, 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 moment in column j would impact the estimate of the parameter in row i. Because our moments and parameters

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

are not on the same scale, we express the elements as elasticities to make them comparable. For each parameter, we normalize the sensitivity elasticities by the sum of their absolute values across all moments. After the normalization, matrix elements provide the relative importance of each moment as a source of variation for estimating a given parameter.

We highlight the key moments for the primary parameters of interest.

1. The two main moments inducing changes in narrowing, η, are grant/abandonment rates, which contribute 43%, and the error moments, which contribute 28%. The intuition for the influence of grant/abandonment moments is that examiners are more willing to grant (and applicants more willing to abandon) in the model when required narrowing is higher.

2. The dominant moments for intrinsic motivation parameters are the error moments, which contribute 48%, and the grant/abandonment rates, which contribute between 15%–18% across the three parameters. Intrinsic motivation affects errors in two ways in our model: through its impact on the intensity of prior art search and through its impact on strategic decisions to grant/reject in the presence of invalid claims.

3. For parameters of the distance distribution, error moments account for 46% for α_D and 32% for γ_D. For a given level of intrinsic motivation, increased rates and sizes of errors require reductions in distances. Further, grant and abandonment rates by round and seniority provide 27% for α_D and 49% for γ_D. Higher grant rates in early rounds necessitate larger claim distances. Finally, the post-grant mean and standard deviation of observed distances provide around 10% for both parameters.

4. For each technology category (simple, electrical, mechanical, and chemical), application attorney cost parameters are moved almost entirely by moments of the corresponding fighting cost. For example, for simple applications, the moment on the median fighting cost contributes 84% for $\mu_{f_{app}}$, and the mean and median moments provide 98% for $\sigma_{f_{app}}$.

5. The four renewal moments contribute 97% of the sensitivity for post-grant obsolescence probability P_ω^{post}. Recall from our descriptive statistics that nearly 50% patents are renewed to the full length, the decision for which happens 12 years after application. To match this moment, the post-grant obsolescence probability must adjust to ensure most patents are renewed the entire length.

6. For pre-grant obsolescence probability P_ω^{pre}, abandonment moments contribute 33% of the sensitivity. Higher abandonments require a higher likelihood of pre-grant obsolescence. Third and fourth-round grants contribute an additional 20%. The central moments for patent returns are the grant and abandonment rate moments, which together contribute

43% for both μ_v and σ_v.

7. The main source for the error variance parameter σ_ε is the error moments. The proportions of patents granted that should not have been contribute 26% of the variation. Higher variance in examiner search errors leads to more decision errors. Variation in examiner rejection rates (fixed effects) also contributes 20%.

8. For delay costs, the grant and abandonment rates are the main moments, contributing 19% and 24% respectively. Higher grant rates require higher costs, so that examiners are not willing to delay by rejecting patent applications.

Alongside the sensitivity analysis, we follow Jalali, Rahmandad, and Ghoddusi (2015) and plot the value of the SMM objective for different values of each parameter, fixing all other parameters at their estimates. Ideally, the curve will be convex in each parameter to ensure a well-defined global minimum exists. We show these plots for our main parameters of interest in Appendix Figure A.3; most curves are U-shaped, as desired.

5 Empirical Results and Robustness

5.1 Parameter Estimates

Table 2 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.

Applicants: We estimate the per-round proportion of narrowing by the examiner as 27% per round. Thus, for a patent application that lasts the mean number of rounds (2.51), screening reduces the granted property rights for rejected claims by 38%. Thus, the grant rate sharply overstates the extent of property rights obtained.

The distribution of initial returns from an unpadded claim is highly skewed, consistent with previous literature: the mean claim value is \$58,390, while the median is \$40,222. The median initial unpadded returns from a patent *application* are \$114,508. Our estimate is broadly in line with previous estimates of U.S. patent values (Bessen, 2008), though the comparison is not perfect.[25]

Pre-grant obsolescence is high, estimated at 13% per negotiation round (typically a year long).

[25]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.

The post-grant obsolescence rate is 4% per year, 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 to be obsolete earlier in its life cycle (e.g., discovering that commercialization costs make the project unviable), and abandonments during prosecution are driven by obsolescence, making those granted a selected sample.

Our estimates of claim distance imply that 86% of application claims have distances below the estimated threshold. Nonetheless, many applications are eventually granted due to substantial narrowing or instances where examiners grant invalid claims.

Applicants bear legal costs for drafting an application. At the mean, the legal application cost varies from \$8,019 to \$11,115, depending on technological complexity, rising to between \$20,808 and \$27,182 at the 90th percentile of padding and fighting cost. Applicants also incur legal fighting costs for each round of amendment. These costs vary at the mean from \$2,143 to \$3,737 per round but with considerable variation (the 90th percentile varies from \$3,201 to \$6,002). Appendix Tables A.3 and A.4 present the application and amendment cost parameters for separate technology fields. Both drafting and amendment costs constitute part of the social costs of patent screening.

Examiners: Our estimates indicate a high degree of intrinsic motivation and substantial variation across examiners: the estimated σ_θ implies a coefficient of variation of 1.27. Junior examiners are more intrinsically motivated than seniors—the median junior examiner is 2.82 times more motivated than the median senior—but there is also large variation within these two categories. 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 imply that the jading effect is stronger than the selection effect. As the first structural estimates of intrinsic motivation in a public agency, there is no direct comparison in the literature.

While the estimated impact depends on seniority and technology area, intrinsic motivation costs are large relative to extrinsic rewards for all categories. 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 1.62 credits. At the mean level, the utility cost is 2.63, which exceeds the two credits received for a final rejection.

Our estimated examiner delay costs are small. The maximal delay cost across all examiner seniorities and technology centers is 0.07 credits per round. This finding suggests that pressure to resolve applications promptly is ineffective (or unnecessary) despite docket management being

Table 2. Parameter Estimates

Parameter	**Symbol**	**Estimate**	**S.E.**
Panel A: Applicant			
Per-round narrowing	η	0.27	0.000
Initial returns log-mean	μ_v	10.60	0.074
Initial returns log-sigma	σ_v	0.86	0.007
Pre-grant obsolescence	P_ω^{pre}	0.13	0.004
Post-grant obsolescence	P_ω^{post}	0.04	0.000
Initial distance alpha	α_D	3.96	0.000
Initial distance beta	γ_D	7.71	0.001
Simple application fighting cost log-mean	μ_f^{simple}	8.55	0.004
Simple application fighting cost log-sigma	σ_f^{simple}	0.88	0.078
Panel B: Examiner			
Junior intrinsic motivation log-mean	μ_θ^J	3.81	0.001
Senior intrinsic motivation log-mean	μ_θ^S	2.77	0.001
Intrinsic motivation log-sigma	σ_θ	0.98	0.000
Delay cost	π	1.01	0.022
Error standard deviation constant	σ_ε	0.18	0.000

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

one of the stated grounds for examiner evaluation.

Recall that examiner assessment errors, ε, are normally distributed with mean $1 + \theta^{-1}$ and variance $\sigma_\varepsilon^2 \theta^{-1}$. Therefore, our findings on intrinsic motivation imply that junior examiners have lower bias and variance in their search errors, relative to seniors. Evaluated at the median level of intrinsic motivation, the junior examiner mean assessment error is 2.2% and standard deviation is 2.7%; for senior examiners, the respective estimates are 6.2% and 4.5%. In summary, examiner errors in assessing distance are modest, remaining within 5.3% of their mean for junior examiners and 8.8% for senior examiners.

5.2 Simulated padding and examiner errors

Padding is not observable in the data, so we simulate our estimated model to calculate the distribution of optimal initial padding for those who apply. At the mean, padding increases claim value by 6%, rising to 15% at the 70th percentile and 31% at the 90th percentile.

We compute two additional key performance metrics with the simulated model: type 1 and type 2 error. Type 1 errors occur when an examiner grants a patent with invalid claims. At the extensive margin, nearly one in five grants contain at least one claim that should not have been approved. However, at the intensive margin, only 8% of all granted claims are invalid. Further, at the extensive and intensive margins, most 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 (equivalently, 10%) below the threshold. Given our earlier finding that 86% of claims have unpadded distance below the threshold, it is clear 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, over 35% of abandonments contain at least one valid claim. Among all abandoned claims, 20% are valid. As with type 1 errors, most type 2 errors also occur in cases of marginal validity: 18% of abandoned applications contain a claim with a distance over one standard deviation above the threshold, and 8% of abandoned claims are at least one standard deviation above the threshold.

5.3 Model fit and robustness analysis

We compare simulated model moments, calculated at the estimated parameters, with moments in the data. We successfully match most of the moments used for estimation (see Appendix Figure A.4). The appropriate test of model fit is the ability to match data moments external to the estimation procedure. We calculate (i) percentiles on granted distances in each round; (ii) mean distance for fourth, fifth, and sixth rounds; (iii) means and percentiles of round one rejection rates across seniority categories; and (iv) percentiles on the number of rounds during prosecution. Appendix Figure A.5 displays the value of these moments. We match the entire 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 $\tilde{D}_j = (D_j^*)^{\vartheta} p^{-1}$. The estimated ϑ is 1.79. Other model parameters are similar to the baseline, 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, we examine a change to the estimator of the distance threshold. In the baseline, we compute the threshold as the maximum, across examiners, of the closest distance among examiners' granted claims. For robustness, we experiment with the first percentile rather than the closest distance for each examiner. We do this robustness check because the way we construct the threshold could be sensitive to outliers in finite samples. The parameter estimates from using the alternative construction of the distance threshold are very similar.

Third, we re-estimate the baseline model using unpurged distance moments. While we prefer the purged distance measure, all results and conclusions go through with the unpurged distance measure.

Fourth, 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: we estimate its value at 0.29 for the first round and 0.22 for later rounds. The other parameters remain robust. Second, we allow narrowing to vary by examiner seniority. We find that senior examiners ask for more narrowing per round, consistent with senior examiners' ability to resolve applications in fewer rounds. Other parameter estimates are similar.

Fifth, 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 one ($\hat{\varsigma}$ =0.89), and 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.

Sixth, we allow examiner delay costs to differ after the second round (i.e., in RCEs). Since delay costs are supposed to capture pressure on examiners for timely resolution of cases, we expect it to rise with later rounds. The estimated delay cost is twice as high in RCEs than in the first two rounds, as expected, but remains small.

Lastly, we allow the variance of examiner distance assessment errors to differ by seniority directly. Our estimates, $\hat{\sigma}_{\varepsilon}^{J}$ =0.23 and $\hat{\sigma}_{\varepsilon}^{S}$ =0.32, show that, even for the same level of intrinsic motivation, junior examiners have a smaller variance in search errors than seniors. Accounting for differences in intrinsic motivation as well, the estimated standard deviation of examiner error is 3.8% for

Table 3. Counterfactual Experiments

Counterfactual	**Not Apply** (%)	**Pad** (%)	**R1 Gr** (%)	$\tilde{v}_j$	**T1** (%)	**T1 Egr** (%)	**T2** (%)	**T2 Egr** (%)
Baseline	5.11	5.73	11.01	50.85	17.64	4.25	38.01	17.76
50K Round Fee	15.33	1.18	16.16	53.33	16.83	4.28	43.21	22.06
Three Rounds	23.91	0.56	14.05	54.61	16.96	4.11	47.45	23.18
Two Rounds	45.79	-3.15	23.58	59.00	14.37	4.29	51.19	30.80
One Round	74.21	-8.29	88.81	64.09	5.50	1.55	75.25	56.06
15% IM	2.39	10.90	31.74	60.42	80.28	52.81	22.16	10.71
Credit$\searrow$	5.05	5.53	11.09	50.83	17.19	3.90	38.21	18.12
Credit$\searrow$ + 15% IM	1.72	32.30	33.52	70.54	81.47	53.35	15.00	6.16

Notes: "Not Apply' is the percent of inventors who do not apply for a patent. "Pad" is the mean level of padding. "R1 Gr" is the percent of applications granted in Round 1. $\tilde{v}_j$ denotes the average padded claim value at grant, in thousands of 2023 U.S. dollars. "T1" represents the proportion of granted patents with some invalid claims, and "T1 Egr" the proportion of granted patents with at least one claim with 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 distance more than one standard deviation above the threshold.

the median intrinsically motivated junior examiner and 7.2% for the senior equivalent.

6 Counterfactual Analysis

We conduct a series of counterfactual experiments to study the impacts of various reforms on the speed and quality of the screening process. These reforms include removing intrinsic motivation, changing Patent Office fees, removing examiner extrinsic incentives, and restricting the number of allowable negotiation rounds. Table 3 presents the results. We report bootstrapped confidence intervals of counterfactual outcomes in Appendix Table A.7. The confidence intervals are tight across all outcomes, and the differences we describe here are statistically significant.

Fees: In the baseline, applicant fees are set at the actual Patent Office levels, which are relatively low and do not include any per-round fees until Requests for Continued Examination. In the first counterfactual, we introduce a $50,000 fee that the applicant must pay for every round of negotiation.[26] Including per-round fees gives applicants greater incentive to exit the patent

[26]We also consider substantially increasing the *application* fees to as much as $50,000. However, because this is

process swiftly and less incentive to apply in the first place. Imposing this fee reduces padding by 79%, and the fraction of inventions for which patents are not sought triples from 5.11% to 15.33%. The mean (padded) value of claims rises when the fee is imposed, 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, one-quarter of type 1 errors and one-half 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.[27]

Rounds Restrictions: Rather than fees, we consider a cap on the number of negotiation rounds. We analyze three alternative caps: three rounds (allowing one RCE), two rounds (removing all RCEs), and one round (removing all negotiation between applicant and examiner).[28] Round restrictions materially affect screening quality and speed. Even a three-round cap increases the proportion not applying nearly five-fold, and virtually eliminates padding at the mean. All three caps increase the mean claim value through the selection effect, with an increase of 26% in the limitation to one round.

Across all rounds restrictions, the proportion of patents granted with invalid claims falls. In the case of only one round, type 1 error falls sharply—at the extensive margin from 17.64% to 5.50%. The downside of rounds restrictions is that they increase the proportion of abandoned applications with valid claims. For example, with no RCEs, type 2 error at the extensive margin

---

a fixed fee paid upon application, provided it is still profitable to apply, applicants will not change their padding decision. Even at this level, the fee does not materially alter average padding and, since there is no substantial change to the proportion of inventors who choose to apply, introducing an application fee acts mainly as a transfer from applicants to the Patent Office, with minimal changes to quality or speed of prosecution.

[27]It may be surprising that such a large per-round fee does not have a larger impact. The reason is that the private value of patent rights is sufficient to make applying for a patent on many of these inventions worthwhile. Fees would have to be much higher to impact outcomes materially. Of course, these fees would be significant for small firms or single inventors who may be cash-constrained. However, round fees for small entities could be reduced, as the Patent Office already does for other types of fees.

[28]These counterfactuals are motivated by a 2007 Patent Office proposal to restrict the number of RCEs. The proposed rulemaking was successfully challenged in federal court (SmithKline Beecham Corp. v. Dudas, 541 F. Supp. 2d 805, 2008). The court decided that the proposed new rules were substantive and that the Patent Office did not have the rulemaking authority to make substantive changes, though the Court noted that the Patent Office could make procedural changes, such as increasing fees. Since one can achieve the same equilibrium number of rounds with an "equivalent" fee, the distinction is problematic from an economic point of view.

rises from 38.01% to 51.19%. Moreover, round restrictions exacerbate the proportion of egregious errors, particularly for type 2 errors.

Finally, we compute the equivalent per-round fee—in the sense of equalizing the mean number of rounds—to restrictions on the number of RCEs. The fee equivalent to removing all RCEs is approximately $275,000 per round, which is politically unpalatable.

Removing Intrinsic Motivation: Next, we evaluate two changes to intrinsic motivation. First, we reduce intrinsic motivation for every examiner by 85% of its original value. Reducing the intrinsic motivation value lowers the utility cost to the examiner of granting invalid claims and increases their errors in assessing distance. Knowing that examiners are more willing to grant invalid claims, only 2.39% of inventors decide not to apply, and the proportion of applications granted in round one almost triples. Though decreased intrinsic motivation makes applicants with lower claim values now apply, the 90% increase in mean padding (which increases padded value both for those who already applied in the baseline and the new entrants in the counterfactual) means that padded values increase on average. Not surprisingly, type 1 error jumps sharply– at the extensive margin, nearly five-fold, with about 65% of these errors being egregious. Type 2 error declines by 40%.

In the second counterfactual related to intrinsic motivation, we keep the values of θ fixed but remove the intrinsic motivation cost from the examiners' grant payoffs. Doing this holds the error distributions fixed, while still removing the utility cost of intrinsic motivation. This quantifies the importance of the direct impact of intrinsic motivation on examiner payoffs. The results of this experiment are similar to the counterfactual that reduces the intrinsic motivation parameter, implying 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 for economic analyses of other public agencies.[29]

Removing Examiner Credits: We remove all credits for the examiner after the first round.[30] This could be justified on efficiency grounds of "marginal cost" pricing since estimated examiner delay costs for an additional round are small. Removing credits has little impact on any outcome

[29]Interestingly, increasing intrinsic motivation (not reported) does not have much impact in reducing padding or type-1 error. The explanation for this feature is that examiners are already sufficiently intrinsically motivated to get most of the benefits, so further increases do not have much bite.

[30]This counterfactual does not incorporate interactions across different applications that the examiner faces. The counterfactual is best thought of as informing an examiner that, for *one* of the new applications in their docket, they will only receive credits for the first round.

variable. This result suggests that intrinsic motivation is sufficiently large for examiners to want to avoid granting invalid patents, even in a context where they will receive no further extrinsic reward if they do so. This finding reflects the dominant effect of intrinsic motivation, and it is not consistent 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 alongside reducing intrinsic motivation to 15% of its value. In this case, we find non-trivial impacts of credits consistent with economic intuition. Padding triples, up to 33.52% (relative to 10.90% when only intrinsic motivation is changed). Type 2 error declines because the increased padding means that abandonments are less likely to include valid claims. These results indicate that extrinsic incentives and intrinsic motivation are *substitutes*, not complements, as sometimes found in the experimental literature. Credits only work as an effective device to incentivize examiners when examiners are not intrinsically motivated.

Final Remarks: As is standard in counterfactual analysis, the maintained assumption in these experiments is that all other model parameters remain unchanged. However, one can envision scenarios where this might not be the case. For example, if a cap on negotiation rounds were introduced, examiners may be more aggressive about reducing padding, and thus, the narrowing parameter, η, could increase. Alternatively, a reform that makes patent screening more rigorous, resulting in stronger patents, might reduce the probability of post-grant obsolescence. Finally, since we find that removing RCEs leads to a 43% decrease in the number of patent applications, the decline in demand for patent attorneys would presumably reduce applicants' fighting costs. These general equilibrium effects would need to be accounted for in a more complete assessment of major reforms.

To conclude, we highlight that none of our reforms unambiguously improves both prosecution speed and quality. For example, policies that make prosecution stricter speed up the process and lead to fewer grants of invalid patents, but 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 quantification methodology here, with further details and formulas in Appendix G.

Type 1 costs

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

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 (an extension to Cournot competition produces similar results—see Appendix G.1). In this case, the change in price equals the unit cost saving. Therefore, the deadweight loss is $DWL = \frac{1}{2}\Delta\wp\Delta q = \frac{\lambda}{2}\frac{\Delta\wp}{\wp}\tilde{V}$, where $\wp$ is the initial price (without the royalty associated with the patent), $\tilde{V} = q\Delta\wp$ denotes total royalty payments, and λ is the absolute value of the elasticity of product demand.[31] We calculate deadweight loss for demand elasticity values $\lambda \in \{1, 2, 3\}$ and report $\lambda = 2$ in the main analysis, though qualitative conclusions are the same for any value in this range.

Litigation costs on invalid patents: Not all invalid patents are "exposed" to litigation because their private value is not large enough to justify the litigation expense. Letting $\tilde{V}$ denote the value at stake in litigation, patents are exposed to litigation if $\tilde{V} \geq \check{V}$, 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), which is $\tilde{v} = 89.6\%$. Hence, $\check{V} = G_{\tilde{V}}^{-1}(\check{v})$, where $G_{\tilde{V}}(\cdot)$ denotes our distribution of the value at stake in litigation.

Type 1 social cost computation: The social cost for invalid patents not exposed to litigation is just the deadweight loss from royalties. The social cost for exposed patents depends on whether they are challenged in court. For exposed patents that are challenged, we assume that the courts are perfect, so always invalidate wrongly granted claims. In this case, the social cost is the sum of litigation costs for both the patentee and challenger, each denoted $\mathcal{C}(\tilde{V})$.[32] Exposed patents

[31] To calibrate, we follow Schankerman and Schuett (2022) and take the ratio of corporate licensing revenue from intangible industrial property to R&D at 39.3%. Multiplying this ratio by the ratio of R&D to sales in manufacturing in 2002 (4.1%) yields $\frac{\Delta\wp}{\wp} = 1.61\%$. See Appendix Section G.1 for further explanation.

[32] We take $\mathcal{C}(\tilde{V})$ as linear in $\tilde{V}$ and calibrate the coefficients using data from the American Intellectual Property Law Association—see Appendix Section G.1 for details.

that are not challenged only impose deadweight loss.[33] Finally, to determine whether an exposed patent is challenged, we use the estimate in Schankerman and Schuett (2022) that, conditional on exposure, the challenge probability is 16.3%.

Type 2 costs

From the ex post perspective, there is *no social cost* from type 2 error since the innovation is already produced and publicized through the patent document, and the R&D cost is sunk. As such, we analyze the social cost of type 2 errors from the ex ante (incentive) perspective. Type 2 error reduces the expected value of patent protection for the inventor, which stops some inventors from developing welfare-enhancing inventions. We calculate 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 quantities: the ex ante value of patent rights (Γ^*), the development cost (κ), and the value of the invention without patent rights (Π). Our model delivers the ex ante value of patent rights, net of all costs (see Equation (6 in Section 2.3)). For the cost of developing an invention, we draw values from the distribution estimated by Schankerman and Schuett (2022).[34] Regarding the value of the invention without patent rights, we first define the patent premium, Ψ, as the proportional increase in private value due to patent protection. Hence, for positive Γ^*, by definition, $\Gamma^* = \Psi\Pi$, implying values of Π. We assume that the patent premium is constant across inventions and calibrate it based on the study of U.S. patents by Bessen (2008).

For the social costs of type 2 errors, we must calculate a development decision rule in the presence and absence of type 2 error, and the social benefit of development. Letting $\mathcal{B}_i = \Pi_i + \max\{\Gamma_i^*, 0\}$ denote the private benefit of development, an inventor does *not* invest to develop an idea i in the presence of type 2 error if the private net benefit from development is negative, that is, if $PNB_i = \mathcal{B}_i - \kappa_i < 0$. An idea is socially valuable to develop if the net social benefit of development is non-negative, i.e. if $SNB_i \equiv \frac{\rho^s}{\rho^p}\mathcal{B}_i - \kappa_i \geq 0$, where ρ^p and ρ^s denote the private

[33] Patentees with invalid patents can pre-empt a challenge by charging a royalty payment (typically a lump sum) equal to the cost of litigation for the challenger (this is commonly referred to as "trolling" behavior). For these cases, the social cost is only the deadweight loss associated with the patent, since the payment is a pure transfer from the licensee to the patentee (Schankerman and Schuett, 2022).

[34] An alternative approach is to assume that inventors do not know their development cost and thus use the mean cost $\bar{\kappa}$. We experimented with this approach and the qualitative conclusions are robust.

and social rates of return. For the baseline, we use a conservative estimate of $\frac{\rho^s}{\rho^p} = 2$ 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 instead obtains the value of all valid claims in the patent. By definition, in this scenario, all remaining abandoned claims are invalid, so there is no type 2 error. Let Γ' denote the expected value of patent rights in this new scenario. The idea i would be developed in this scenario if $PNB_i' = \Pi_i + \max\{\Gamma_i', 0\} - \kappa_i \geq 0$. We can then compute type 2 costs as the sum of net social benefits SNB_i, across ideas with $PNB_i < 0$ but $PNB_i' > 0$.

Patent prosecution costs

The social cost of patent prosecution is the sum of per-application Patent Office administrative costs and applicant legal costs. For applicant legal costs, we include the application drafting cost (F_{app}) and the total amendment costs (equal to the per-negotiation amendment cost F_{amend} multiplied by the number of negotiations). For the administrative cost of an application, we multiply the number of claim-rounds by the average Patent Office cost per round and claim. To calculate the latter average, we take the official USPTO operations budget per application, which equals $4,117 (in 2018 dollars), and divide it by the average number of rounds and the average number of independent claims in our baseline model.[35]

Benefits of type 1 and type 2 errors

We also account for 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 is analogous to the cost of 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 "counterfactual" in this case is one where the inventor only obtains the value of the valid claims in the patent when granted. The benefit from type 2 errors is the deadweight loss avoided by not having granted the patent right. Note that the type 2 benefit is ex post, even though the type 2 cost is considered from an ex ante perspective. 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.

[35]The patent prosecution costs exclude Patent Office fees and the loss in patent value from pre-grant obsolescence, since those represent pure transfers from the applicant to either the Patent Office or the owner of the invention that superseded it.

Before turning to the results, one subtle qualification should be noted: our quantification of social costs assumes that the patentability threshold is set at the optimal level (refer to Schankerman and Schuett (2022) for a discussion of optimal patent eligibility). To see this, suppose the 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 and type 2 benefits. A similar argument applies to type 2 costs. Whether the patentability threshold currently applied is optimal remains an open research question that we are currently pursuing.

7.2 Estimated social costs of patent screening

Table 4 summarizes the social costs for the baseline model and counterfactual reforms. The baseline row presents the social costs associated with a yearly cohort of ideas. Appendix G explains how we calibrate the annual number of ideas. All values are presented in 2023 U.S. dollars. In Table 4, we report the 95% percentile bootstrapped confidence intervals for the total social cost; Appendix Table A.7 provides the confidence intervals for each component of social costs.

In the baseline, social costs from type 1 error are $4.75bn, from type 2 error are $1.31bn, and prosecution costs dominate at $18.65bn. The bulk of the prosecution cost is due to applicant legal costs rather than Patent Office administrative costs. The total social cost of patent screening is $24.71Bn, equivalent to approximately 6.3% of all R&D performed by business enterprises in the U.S. in 2011.

Introducing a $50,000 per-round fee reduces prosecution costs by discouraging applications and reducing padding. The fee has no material effect on type 1 error and hence little 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 about 10% with the per-round fee. If this fee raises extra revenue that can be reinvested in more intensive/faster examination, then social costs would decline further. The importance of reinvestment is highlighted in Schankerman and Schuett (2022).

The impact of caps on the number of negotiation rounds is much larger than that from the $50,000 per-round fee. Removing all RCEs (cap of two rounds) reduces total social costs by 23% compared to the baseline, and restricting the process to one round more than halves total social costs. The non-overlapping confidence intervals confirm statistical significance of these results. While the total declines, type 2 costs rise with rounds restrictions. This may induce political opposition to such a reform from the patent community in the absence of some compensation,

Table 4. Net Social Costs of Patent Prosecution

Counterfactual	T_1 Cost	T_2 Cost	T_3 Cost	Total	Total C.I.
Baseline ($Bn)	4.75	1.31	18.65	24.71	[23.13, 26.05]
50K Round Fee	4.30	2.14	15.87	22.31	[21.18, 24.20]
Three Rounds	3.81	3.98	13.65	21.44	[19.37, 25.29]
Two Rounds	2.92	7.28	8.79	18.99	[17.48, 22.03]
One Round	0.57	6.50	3.57	10.64	[8.92, 14.00]
15% IM	17.66	1.68	18.95	38.29	[32.51, 40.62]
Credit$\searrow$	4.42	1.32	18.65	24.39	[22.61, 25.73]
Credit$\searrow$ + 15% IM	3.38	2.98	20.63	26.99	[25.25, 30.47]

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. The "baseline" row provides the total social costs in billions of 2023 U.S. dollars. 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 $\Psi = 0.025$.

such as an adjustment to the R&D tax credit.

Reducing intrinsic motivation to 15% of its original level increases total social cost by 55% (the difference is statistically significant), and all components of social cost rise—there is no tradeoff. 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 countervailed 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 four-fold. This finding confirms the importance of intrinsic motivation in this public agency.

Finally, with the baseline level of intrinsic motivation, removing all examiner credits after the first round for one examination has a marginal effect on social costs, consistent with our results in Section 6. However, removing both intrinsic motivation and examiner credits beyond round one reduces social costs relative to only removing intrinsic motivation. This counter-intuitive result suggests that credits are counter-productive even when intrinsic motivation is low. The explanation for this finding, which is not an artifact of our model, is the large increase in type 1 *benefits*. In this counterfactual, examiners grant more readily, which further increases inventor development incentives. This result highlights the importance of accounting for the positive incentive effects of less stringent screening, as opposed to just the ex post social costs that arise

through deadweight losses and litigation. These interaction effects illustrate the role for structural modeling in public agency reform.[36]

8 Conclusion

In this paper, we study the allocation of property rights for innovation by estimating the first structural model of the patent screening process. The model incorporates incentives, intrinsic motivation, and multi-round negotiation between the examiner and applicant. We find that patent screening is moderately effective given the statutory and judicial standards for patentability within which the Patent Office is required to operate. This effectiveness is primarily driven by the substantial intrinsic motivation of examiners. Among counterfactual reforms we study, restrictions on the number of allowable rounds of negotiation significantly reduce the social costs of screening. We estimate the total social cost of patent screening at $24.71bn per year, which represents 6.3% of R&D in the United States performed by business enterprises.

There are two research direction of interest. The first is to model patent screening in other important jurisdictions, such as the European and Japanese Patent Offices, which would allow for comparative evaluation of the institutions that underpin innovation. A second direction is to estimate our model on individual, separate, technology fields, to assess whether a one-size-fits-all institutional design is appropriate. Finally, we believe there are promising opportunities to use structural models to study other innovation-supporting institutions, such as the NIH, NSF, and similar institutions in other countries.

Appendices

Supplemental appendices are available **here**.

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.

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.

[36] This point is also emphasized in Freilich, Meurer, Schankerman, and Schuett (2024).

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. Christensen, S. Hansen, and S. Sacher (2024): "Inference for Regression with Variables Generated from Unstructured Data," *arXiv preprint arXiv:2402.15585*.

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., S. Kortum, and S. Stern (2003): *Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigation Outcomes*, Washington, DC: The National Academies Press.

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

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

Frakes, M. D. and M. F. Wasserman (2017): "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.

Freilich, J., M. Meurer, M. Schankerman, and F. Schuett (2024): "A New Approach to Patent Reform," *UC Irvine Law Review*, 14, 351–403.

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.

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

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.

Jalali, M., H. Rahmandad, and H. Ghoddusi (2015): *Using the method of simulated moments for system identification*, MIT Press.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021): "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 3, 303–20.

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

Lanjouw, J. O. and M. Schankerman (2004): "Patent Quality and Research Productivity: Measuring Innovation with Multiple Indicators," *The Economic Journal*, 114, 441–465.

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.

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

Pakes, A. and M. Schankerman (1984): "An Exploration into the Determinants of Research Intensity," in *R&D, Patents, and Productivity*, ed. by Z. Griliches, Chicago: University of Chicago Press, 209–232.

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.

Sutton, J. (1997): "One Smart Agent," *The RAND Journal of Economics*, 28, 605–628.

The Economist (2015): "Time to fix patents," *The Economist Group*, August 8th-14th, 9.