Granting Intellectual Property Rights:

Invention Quality and the Scope of Protection

Lorenz Gschwent*

Job Market Paper

November 9, 2026

[Click here for the most recent version]

Abstract

A patent office judges the quality of an invention and the scope of protection the applicant requests. I measure both from the text of US patent applications and find that the outcome of examination depends far more on scope than on quality. Using the assignment of applications to examiners, I show that among granted patents, the average alternative examiner would have refused about 8.6% of the broadest requests but only 5.9% of the narrowest. Studies that use the examiner assignment to estimate what a patent does therefore describe broad requests far more than low-quality inventions. In an estimated model in which applicants choose how broadly to claim, I recover the value of scope conditional on quality and characterize granted patents under alternative examination rules. A stricter office narrows patents mostly because applicants request less scope and changes which inventions receive patents far less than an equally strict test of invention quality would.

Keywords: patent office, patent examiners, patent quality, natural language processing
JEL Classification: O31, O34, L24, L32, K12, D73, D83

*University of Duisburg-Essen, e-mail: *lorenz.gschwent@uni-due.de*.

I am grateful to Tobias Seidel, Philip Jung, and Martin Karlsson for their advice and encouragement. I thank Lukas Buchheim, Bernhard Ganglmaier, Daniel Gross, Hanna Hottenrott, Paul Jensen, Eugen Kovac, Matt Marx, William Matcham, Gaétan de Rassenfosse, Bhaven Sampat, Beth Webster, Galina Zudenkova, and seminar participants at the University of Duisburg-Essen, RTG Regional Disparities and Economic Policy, ZEW Mannheim, Melbourne Institute of Applied Economic and Social Research, as well as participants and discussants at the 18th RGS Doctoral Conference, 2nd Duke Strategy PhD Conference, REGIS Summer School, BSE Summer Forum, 20th EPIP Annual Conference, 8th RISE Workshop, DRUID Academy, and ICSSI Boulder for their helpful comments. This paper was previously presented under the title "Quality, Scope, and Leniency: Strategic Application Behavior at the US Patent Office". I gratefully acknowledge the computing time granted by the Center for Computational Sciences and Simulation (CCSS) of the University of Duisburg-Essen and provided on the supercomputer amplitUDE (DFG project 459398823; grant ID INST 20876/423-1 FUGG) at the Zentrum für Informations- und Mediendienste (ZIM).

1 Introduction

Like many other gatekeepers, a patent office judges the quality of an application and the size of the request. The applicant discloses an invention and writes claims that define the scope of the requested protection. The office decides whether to grant a patent and with what scope. The office's decisions matter for growth and innovation: a first patent helps a startup grow (Farre-Mensa, Hegde, and Ljungqvist, 2020), and in some industries large firms' patents block follow-on innovation by small firms (Galasso and Schankerman, 2015). In models of optimal patent policy, both the grant decision and the decision on the scope of protection depend on the invention: the office refuses patents to inventions that do not need the patent incentive (Schankerman and Schuett, 2022), and bigger improvements receive broader protection (Hopenhayn, Llobet, and Mitchell, 2006). Empirical results in other settings, by contrast, show the gatekeeper responding to the request: on a hiring platform, salary offers depend much more on the candidate's request than on the résumé (Roussille, 2024), and mortgage lenders more often deny applications whose proposed housing expenses are high relative to income (Munnell, Tootell, Browne, and McEneaney, 1996).

In this paper, I show that the outcome of patent examination depends on the scope of the request far more than on the quality of the invention. Earlier work could not make this comparison, because every established proxy of invention quality exists only for granted patents. I proceed in three steps. First, I construct a new measure of invention quality from the text of each patent application, which allows me to separate invention quality from requested scope. Both the quality measure and the scope measure come from the application, before the office examines it. I show that the chance of a grant decreases with scope far more than it increases with quality, and better inventions do not receive broader patents.

Second, using quasi-experimental variation from the assignment of applications to examiners, I show that grants depend on the assigned examiner about 46% more often for broad requests than for narrow ones, but only slightly more often for low-quality inventions. Which grants depend on the assigned examiner determines the incentives that examination sets for applicants. My results imply that by choosing the requested scope, applicants choose how much the outcome of their application depends on the assigned examiner.

Third, because examiner assignment cannot reveal how applicants would respond to changes in examination, I build a model in which applicants choose how broadly to claim their invention and, after each rejection, decide whether to continue prosecuting the application or abandon it. The model identifies how the value of a grant relative to the cost of a revision round varies with invention quality. In the model, a stricter office of the kind that reform proposals call for (U.S. Federal Trade Commission, 2003; Executive Office of the President, 2013; Productivity Commission, 2016)

narrows patents mostly because applicants request less scope rather than because it rejects the same requests more often. Because examination mostly tests the scope of legal protection, a stricter office changes which inventions receive patents far less than an equally strict test of invention quality would.

I measure the quality of the invention and the scope of the request separately, for refused applications as well as granted ones. The quality measure uses the title and abstract, which describe the invention and do not overlap with the text describing the claims, the requested legal protection. An invention is labelled high-quality when its text differs from earlier filings in its technological field and resembles later ones. To validate the measure, I show among other checks that refused applications in the top decile of quality are cited 42% more often than those at the median. The scope measure uses the claims, where more specific language defines a narrower scope (Kuhn and Thompson, 2019; Feng and Jaravel, 2020). For example, a claim to editing genes is broad, while a claim to editing genes in mouse liver cells using a specific enzyme is narrow.

The grant rate varies far more with the scope of the request than with the quality of the invention. Among applications that do not build on an earlier filing, the grant rate decreases from 77% for requests at the 10th percentile of scope to 59% at the 90th. From the bottom to the top decile of quality, it increases by about two percentage points on a base of 70%. Applicants with better inventions request only slightly broader claims and do not receive broader patents. Before a grant, they go through more rounds and narrow their claims more.

I use variation in the assignment of applications to examiners to identify which grants depend on the assigned examiner across scope and quality. Within a technological field, supervisors assign applications to examiners without reviewing their content (Sampat and Williams, 2019; Pairolero and de Grazia, 2026), and examiners differ in how strict they are. I compare the assigned examiner with the average examiner the application could have been assigned to in the same field and period. The main object is the share of granted patents that depend on the assigned examiner, measured by grants that the average alternative examiner would have refused. The kind of application receiving these grants determines the incentives that examination gives applicants. If they go to low-quality inventions, applicants can make a grant more certain only with a better invention. If they go to broad scope instead, applicants can make a grant more certain by requesting narrower scope.

The share of grants that depend on the assigned examiner increases with the scope of the request and decreases slightly with the quality of the invention. Overall, the average alternative examiner would have refused 7.1% of granted patents. The share increases from about 5.9% for the narrowest scope to about 8.6% for the broadest. Across the quality distribution, it decreases from about 7.2% to about 6.7%. At every level of quality, the share is 2.8 to 3.8 percentage points higher for the broadest scope than for the narrowest. The assignment is not exactly random with respect to quality and scope. The imbalance is smaller than in existing work with the same

design (Goldsmith-Pinkham, Hull, and Kolesár, 2026, reanalyzing Farre-Mensa et al., 2020), and adjusting for it changes the results little. As one of several robustness checks, I show that the results hold in the fields where the office is known to assign applications randomly.

Grants of broad requested scope depend more often on the assigned examiner because broader scope is closer to the grant threshold, whatever the quality of the invention. A higher-quality invention can support broader scope, but examination weighs the scope far more than the quality. The assigned examiner is pivotal for applications whose assessment lies close to that threshold. Applicants do not choose their examiner, but the scope they request therefore determines how much the outcome of their application depends on the assigned examiner. I also rule out the alternative explanation that broad requests drive the observed pattern because they are harder to assess.

Because applicants choose their claims, changes in examination affect what applicants request as well as how the office decides on a given request. To measure both effects, I build and estimate a model of patent prosecution. Examiners in the model do not optimize but apply the office's grant rule, shifted by their own leniency. Applicants choose how broadly to claim. A grant is worth the scope of the request times a value per unit of scope, which depends on the quality of the invention. A broader request is therefore worth more if granted, but it lowers the chance of a grant in each round. Because the law sets no limit on the number of rounds, applicants decide after each rejection whether to continue prosecuting the application or abandon it. In each new round, the scope can be no broader than before.

The model measures the value of scope conditional on quality, relative to the cost of a revision round. Because applications are assigned to examiners as good as randomly, the grant rate across leniency shows where applications of each quality and scope lie relative to the grant threshold. How that position changes with requested scope at the same quality identifies the price of scope. The cost of a revision round is identified from how long applicants keep negotiating. Given both, how the scope that applicants request and the number of revision rounds they go through vary with quality reveals what a unit of scope is worth to them relative to a round. I find that a unit of scope is worth about 15% more at the 90th percentile of quality than at the 10th. The model estimation does not target the share of grants that depend on the assigned examiner, but the model also makes grants of broader requests depend more often on the assigned examiner.

In the model, a stricter office narrows patents mostly because applicants request less scope. Given the estimated parameters, I simulate an office in which the median examiner is as strict as today's 45th percentile, applicants know about the change, and the set of filed inventions stays fixed. Grants decrease by 4.0%, and granted patents become 4.2% to 4.4% narrower. About 80% of the decrease in grants comes from stricter decisions on the requests applicants make today, but only 10.3% to 15.4% of the narrowing does. The rest of the narrowing comes almost entirely from

applicants requesting less scope. An evaluation that holds applications fixed therefore misses most of the narrowing.

The stricter office changes which inventions receive patents far less than an equally strict test of invention quality would. The comparison office tests the quality of the invention instead of the scope of the request and becomes stricter by the same amount. Under today's rules, the best and the worst quintile of inventions both lose about 4% of their grants. Under the test of quality, the worst quintile would lose 6.9% of its grants and the best quintile 2.4%.

This paper contributes to three literatures. First, I show which patent applications the examiner assignment decides. A growing number of studies use the assignment to estimate what a patent does for firms, for inventors, and for follow-on innovation (Gaulé, 2018; Sampat and Williams, 2019; Farre-Mensa et al., 2020; Melero, Palomeras, and Wehrheim, 2020; Hegde, Ljungqvist, and Raj, 2022). Their estimates apply to the applications whose grant depends on the assigned examiner. Frakes and Wasserman (2017) find that the additional grants of examiners with less time to review are of below-average quality. Avivi (2025) find that examiners differ in how the gender of inventors affects their decisions on applications with similar text. Unlike Goldsmith-Pinkham et al. (2026), who compare applications decided by the examiner assignment one characteristic at a time, I characterize these applications jointly by the quality of the invention and the scope of the request. I find that the grants that the assignment decides differ from other grants far more in scope than in quality. Feng and Jaravel (2020) propose that examiners matter mainly through how they craft the patents they grant. I show that whether an application is granted at all also depends on the scope of the request.

Second, I contribute to models of patent examination. In most of them, the grant tests the invention, and applicants choose neither their claims nor whether to continue prosecution after a rejection (e.g., Caillaud and Duchêne, 2011; Schuett, 2013; Llobet, Parra, and Suarez, 2026). I show that both of the applicant's choices shape who receives patents. Consistent with my findings, Aneja, Reshef, and Subramani (2026) find that women are less likely than men to continue prosecution after an early rejection. This difference explains about half of the gender gap in issued patents among filed applications. At the European Patent Office, applicants also withdraw their most valuable applications later (Harhoff and Wagner, 2009). Theories of optimal patent scope treat scope as a policy choice (Gilbert and Shapiro, 1990; Klemperer, 1990), but I show that in examination the applicant's request sets the scope at stake.

Matcham and Schankerman (2026) model the claims and repeated rounds of negotiation in an office that screens claims for validity. Their applicants decide after each rejection whether to continue, given the value of their claims, but that value is unobserved. Its distribution comes from patent renewals, and it is assumed to be unrelated to the claim's distance from prior art. I identify how the value of a grant relative to the cost of a revision round varies with measured invention

quality. The identifying variation is how requested scope and the number of revision rounds vary with quality, given the price of scope and the cost of a round. Without this gradient, higher-quality inventions would receive patents less often than low-quality ones. Matcham and Schankerman's (2026) counterfactuals change fees, caps on rounds, and examiners' credits and motivation, while holding the patent acceptance threshold fixed. Instead, I raise the patent acceptance threshold itself. Because I measure invention quality, I can also show that a higher threshold changes the quality composition of patented inventions far less than an equally strict test of quality would.

Third, I contribute to work on whether stricter screening acts through stricter decisions on what applicants submit or through changes in what they submit. Stricter screening changes what people submit in disability insurance (de Jong, Lindeboom, and van der Klaauw, 2011; Godard, Koning, and Lindeboom, 2024) and in tax enforcement (Kleven, Knudsen, Kreiner, Pedersen, and Saez, 2011). Ninety percent of the savings from Medicare audits come from deterring future spending rather than from recovering past spending (Shi, 2024). Studies that use the assignment of decision makers typically capture the first channel, and their counterfactuals hold the number and composition of applicants fixed (Autor, Kostøl, Mogstad, and Setzler, 2019). Studies of a stricter patent standard that follow applications filed before the change also hold applicants fixed (Toole and Pairolero, 2020; Frumkin, Pairolero, Tesfayesus, and Toole, 2024). Instead, applicants in my counterfactuals change their requested scope and decide after each rejection whether to continue prosecuting the application or abandon it, while the set of filed inventions stays fixed. At the patent office, most of the narrowing comes from applicants requesting less scope. Stricter decisions on today's requested scope matter far more for the number of grants than for granted scope.

The remainder of the paper is organized as follows. Section 2 describes examination at the USPTO, and Section 3 introduces the data and the two measures. Section 4 shows how grant rates and prosecution vary with the scope of the request and the quality of the invention. Section 5 describes the design based on the examiner assignment and tests its assumptions, and Section 6 presents its results and the mechanism behind them. Section 7 builds and estimates the model and uses it to simulate a stricter office. Section 8 concludes.

2 Institutional background

This section describes four features of patent examination and prosecution at the USPTO that my analysis rests on: what an application contains, how it reaches an examiner, on what grounds the examiner decides, and how examination unfolds over rounds. A patent application contains a title, a short abstract, a longer description, and the legal claims sought by the applicant. The description cites "prior art": earlier patents, applications, or publications, such as scientific articles, relevant to the claimed invention.

The USPTO assigns each incoming application a class and subclass code under its patent classification (USPC, later CPC) and, based on the subclass, routes it to an "art unit", a group of 15–30 patent examiners with expert knowledge in certain technology fields. While there is evidence of some specialization of examiners within art units (Righi and Simcoe, 2019), the leniency literature treats the assignment of applications to examiners within an art unit as as-good-as-random (Gaulé, 2018; Sampat and Williams, 2019; Farre-Mensa et al., 2020). I test the randomness of assignment formally in Section 5.

A legal presumption of patentability, together with time constraints that can be binding, shapes what examination can accomplish. Under the legal presumption, it is the examiner's task to show that an invention is not patentable. Although the expiry of the time allotment does not by itself lead to approval, in reality an application is likely to be granted if the examiner has not established a case for rejection within the time budget. Rejections for obviousness or lack of novelty require potentially time-consuming searches for prior-art references. Rejections for eligibility, written-description, enablement, and definiteness grounds do not, though the prior-art grounds dominate in practice. Searching for prior art beyond what the application already cites is time-intensive, the per-application time budgets bind, and truncated searches produce low-quality patents (Frakes and Wasserman, 2017, 2023). Examiners are also held to a biweekly quota of credits ("counts"), but its deadline pressure shows up in rushed first-round rejections rather than in rushed grants: an examiner short of time at the deadline issues a rejection to buy time, and when a 2011 reform smoothed work over the quota period those rejections disappeared while review accuracy did not change (Frakes and Wasserman, 2020, 2024). The time budget is the constraint that affects what is granted.[1] Examiner experience is a further systematic driver of grant behavior. Using examiner pay scale data, Frakes and Wasserman (2017) show that experience and grant rates follow an inverse U-shape relation, with experience having the largest effect after 4-5 years as patent examiner.[2] Additionally, examiners undergo on-the-job training in their first year.[3] These patterns motivate the experience controls in my estimation strategy and imply that the term leniency bundles examiner-specific standards, time pressure, and experience effects. I use the existing terminology throughout, but the underlying source of the leniency bundle matters for the counterfactual policies of Section 7.7.

The application process at the USPTO is often iterative. Carley, Hegde, and Marco (2015) show that first-round grants are rare, consistently below 20% between 1996 and 2005. A first-round rejection is not final, however: cumulatively over multiple rounds, 71% of the applications

[1] Credits are awarded for a first office action, a patent grant, or a final rejection, with at most two credits per application. The quota sets the USPTO's application throughput expectations.

[2] The authors distinguish the effects of experience and merit-based promotions (leading to a higher workload) using examiner pay scale data received via a Freedom of Information Act request. Because this information is not available for my full sample, I rely on experience proxies instead.

[3] See information provided by the U.S. Department of Commerce.

in my sample end in a grant. Applicants can revise their initial application, after which examiners can either allow, non-finally reject, or finally reject the application. A non-final rejection invites a further amendment as of right. A rejection is typically made final when the applicant's amendment does not overcome the examiner's stated grounds and necessitates no new ground of rejection, which closes the regular exchange. After a final rejection, applicants can appeal within the USPTO, file a request for continued examination, which reopens prosecution and appears in my data as further revision rounds, or file a continuation application. Marco, Sarnoff, and deGrazia (2019) find that the number of claims and their broadness decrease during the review process, matching the common understanding that applicants respond to examiners' rejections by narrowing their claims. Once such concessions have been made, they become costly to reverse. On the basis of the prosecution history estoppel principle, if an applicant narrows a claim in order to obtain a patent, they are thereby presumed to have given up the right to later recover the subject matter that was surrendered by means of the doctrine of equivalents. This presumption can be challenged and is specific to the amendment in question. Together with the empirical regularity that narrowing amendments are in practice not reversed (Marco et al., 2019), the presumption means that a revision costs more than the labor of drafting it: the concession itself is a cost in foregone future profits. The important take-away is that granted scope is an endogenous outcome of a multi-round negotiation, while filed scope is set before the examiner is assigned.

3 Data and measurement

3.1 USPTO data

The raw data cover the universe of *published* patent applications received by the USPTO from 2001, the first year for which the USPTO publishes data including rejected applications, through 2023 inclusive. Applications abandoned before the eighteen-month publication date and the small share filed with a nonpublication request never enter the public record and so cannot appear in any dataset built on it. I combine PatEx – a dataset including patent applications and relevant examiner information – with the application-text dataset to obtain a total of 6,593,022 patent applications.[4] I restrict the estimation sample to applications published between 2005 and 2020, reducing the number of observations to 5,367,642 applications. Of these, 2,671 receive no quality score because their comparison windows contain no other application, and the examiner-dummy estimators of Section 5 drop a further two to stay within the largest leave-out-connected component of the examiner-cell design,[5] so estimation totals read 5,364,969. The lower bound discards a break

[4] Both PatEx and patent application texts can be directly downloaded in bulk via the USPTO website.
[5] I discuss this in more detail in Appendix E.

in the claim-text parsing regime before 2005, which would otherwise contaminate the scope measure described in Section 3.2. The upper bound addresses the problem of right-censoring of prosecution outcomes at the end of the time period. The years outside this window still contribute to the construction of examiner decision histories and to the five-year text windows behind the quality measure, but the quality percentiles are re-ranked within the estimation sample. Appendix A reports summary statistics for the estimation sample.

The variation I use comes from the examiner assigned to an application. Section 5 measures the leniency of each examiner in each publication year from examiner × publication-year indicators. The pass-through estimator of Appendix E and several checks in the appendix use a simpler measure, which I construct in a realized and an expected form. Following Farre-Mensa et al. (2020), I calculate realized examiner leniency as a leave-one-out average of the grant outcomes among all patent applications previously examined by the same examiner. The measure is undefined for an examiner's first observed decision, and I exclude degenerate leniency values of exactly zero or one, which arise for examiners with short histories. Additionally, based on the art unit structure of the USPTO, I calculate an expected leniency. That is, for each patent application, I calculate the average leniency of examiners over all other applications decided in the same art unit within a three-year window centered on its decision date, i.e. eighteen months on either side.[6] Any examiner who made a decision in that art unit during this window could also have been assigned the application in question, so the expected leniency averages over the examiners the application could have been assigned.[7] The deviation between realized and expected leniency, Δl, isolates the random component of the examiner assignment in this measure.

From the PatEx data, I additionally construct the control variables and auxiliary outcomes used throughout the analysis. These comprise the small and micro entity status of applicants as recorded in the PatEx snapshot, which determines the fee schedule and therefore proxies for applicant size.[8] They further comprise the number of citations made to the publication and, in case of a grant, to the patent, and the number of revisions, counted as non-final rejections in the transaction codes. An identifier marks applications that are a (partial) continuation of a previous application. To proxy for the experience effects discussed in the previous section, I use the years since an examiner's first observed decision, top-coded at five, with examiners already active in the first year of the data flagged separately. The descriptive figures use the equivalent set of first-

[6] The decision date must be defined for granted and failed applications alike. I calculate it from the patent issue date for grants and from the date of the first office action for all other applications. Both dates are part of PatEx.

[7] Since a more lenient examiner also grants sooner, the issue date on which the window of a granted application centers is not strictly fixed. In Appendix F, I reconstruct the benchmark using a pre-treatment anchor which the examiner assignment cannot move. Under this assumption, the first stage is stronger, so the conventional construction remains the baseline as the conservative choice.

[8] A small entity is an individual, a small business under the Small Business Act, or a non-profit organization. A micro entity is a small entity that is either a US institution of higher education or falls below limits on both yearly gross income and previous patent applications.

five-year experience dummies. Maintenance fees to keep a utility patent in force are due 3.5, 7.5 and 11.5 years after grant, so I use the grant date and the expiry status of a patent to calculate renewals. Because many granted patents in my sample have not passed all three payment dates, I calculate realized renewals as the difference between actual and potential renewals.[9] Renewals additionally serve as a validation outcome for the quality measure.

For around 2.1 million patent applications in my sample, I can also use the Office Action Research Dataset (Lu, Myers, and Beliveau, 2017), including the reasons for rejections. For validation, I add the measure of Kogan, Papanikolaou, Seru, and Stoffman (2017) for around 1.3 million applications that end up granted and the measure of Kelly, Papanikolaou, Seru, and Taddy (2021) for around 2.5 million.

3.2 Scope measurement

I measure the scope of patent claims using the average number of words per claim, following Feng and Jaravel (2020) and similar to Kuhn and Thompson (2019). Narrower claims use more words, because each limitation on what is protected must be stated. I therefore treat the word count as the inverse of scope. Claim length is the usual measure of scope in this literature, and two descriptives documented below show that it is related to scope rather than to the drafting style alone. On the one hand, I show in Section 4 that the grant rate decreases sharply across the percentiles of scope, falling from 77% at the 10th percentile to 59% at the 90th among applications that do not build on an earlier filing. On the other hand, examiners reduce scope during prosecution, enforcing concessions that increase with prosecution length, as shown in Figure 3d. Wordiness alone would neither lead to rejections nor be the object applicants have to give up under estoppel costs.[10]

I construct filed scope, the scope the applicant requests at filing, from the application texts and granted scope from the granted patent texts separately, enabling a direct comparison of scope at the beginning and, if granted, the end of the prosecution process. For both filed and granted scope, I turn the raw word counts into percentile ranks by using the empirical distribution pooled over both measures within each publication year. I reverse the sign so that a higher value corresponds to wider scope. By pooling filed and granted scope, I put both measures on the same scale, which makes it possible to measure scope concessions directly. I rank the measures within publication year to make them robust to changes in claim-text data over time. Appendix J tests both dimensions of this construction: the results survive measuring scope from the first claim alone or from the total length of the claims, and the rank-based results are exactly invariant to the cardinal scale

[9] This implies that the resulting variable is equal to zero for a patent that has been renewed at every possibility, while it is negative for patents that at some point were not renewed.

[10] Claim length may still carry residual variation from technological complexity or house drafting style. The two facts establish that the component examiners screen and applicants concede is scope.

behind the ordering.

3.3 Quality measurement

A terminological distinction matters for everything in this subsection. In the patent literature (for example, Lemley and Shapiro, 2005; Galasso and Schankerman, 2015; de Rassenfosse, Griffiths, Jaffe, and Webster, 2020; Higham, de Rassenfosse, and Jaffe, 2021), *patent quality* means the robustness of a granted patent as a legal instrument, that is, whether its claims would survive an invalidity challenge. However, legal robustness is jointly determined by the quality of the invention and the granted scope of the patent. This paper studies the first part, the quality of the *underlying invention*. Therefore, I distinguish between the technological quality of the idea seeking protection and the granted scope, which I treat as a separate, endogenous outcome. This distinction is important, as on the one hand, a high-quality invention forced into a narrow scope yields a legally defensible patent with limited economic value. On the other hand, a low-quality invention receiving a broad granted scope yields a legally fragile patent.

I measure the quality of the invention in each application from its text, using transformer-based sentence embeddings (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, 2017). An embedding maps a piece of text to an n-dimensional vector in a space that represents semantic meaning, so that similar text is closely grouped and vector distances measure the similarity of language, here between two patent applications. I use `sentence-t5-xl` (Ni, Hernández Ábrego, Constant, Ma, Hall, Cer, and Yang, 2022), which outperforms both co-occurrence-based embeddings such as `GloVe` (Pennington, Socher, and Manning, 2014) and `SimCSE` (Gao, Yao, and Chen, 2021) on semantic-similarity benchmarks. I argue that for patents, where the signal of interest is an application's proximity to prior and subsequent work in its field, semantic similarity in the embedding space is the correct object for measurement. The word frequencies based on term frequency-inverse document frequency (TF-IDF) used by Kelly et al. (2021) can conflate term frequency with conceptual similarity, whereas transformer-based sentence embeddings can capture the latter directly.

For each patent application in my sample, I first concatenate the title and the abstract. I run this combined text through sentence-t5-xl (Ni et al., 2022) to get a 768-dimensional vector representation for each application. I use this vector v_p for each patent application p, dated by its filing year t, to calculate my quality measure. Adapting the forward-to-backward similarity ratio of Kelly et al. (2021), I compute within every art unit:

$$FS_p = \text{sim}\left(v_p, \bar{v}_p^{+}\right) = i_p \qquad i_p \in [0,1] \tag{1}$$

$$BS_p = \text{sim}\left(v_p, \bar{v}_p^{-}\right) = 1 - n_p \qquad n_p \in [0,1] \tag{2}$$

with the cosine similarity $\text{sim}(v, w) = \frac{vw}{\|v\|\|w\|}$, the average vector of the applications filed in the five years preceding p's filing date,[11] $\bar{v}_p^- = \frac{1}{N_p^-} \sum_{\tau=t-5}^{t-1} \sum_{i \in p_\tau} v_{i,\tau}$, and its counterpart over the five following years, $\bar{v}_p^+ = \frac{1}{N_p^+} \sum_{\tau=t+1}^{t+5} \sum_{i \in p_\tau} v_{i,\tau}$, where p_τ denotes the set of applications filed in year τ within p's art unit and N_p^-, N_p^+ count the applications in the respective window. i_p and n_p represent impact and novelty.[12] Next, I take the ratio of forward similarity over backward similarity. The ratio implies that the measure rewards being simultaneously novel and impactful, so an application that is unlike its predecessors but also unlike its successors is not scored higher than an application that is unremarkable in both directions.[13] Finally, I take percentiles of the ratio to arrive at my quality measurement:

$$Q_p = P\left(\frac{FS_p}{BS_p}\right) \qquad Q_p \in [1, 100] \tag{3}$$

Three design choices require discussion. Appendix I discusses them in further detail. First, I embed titles and abstracts rather than patent claims. Claims are the object of the examination. They are narrowed during prosecution in response to examiner objections, whereas titles and abstracts are fixed in the filing text. Therefore, the text Q_p is built from cannot have been rewritten in response to the decisions it is used to study. Second, the percentile transformation is calculated separately for each art unit $\times$ publication year cell – the same partition used as the fixed effects in the first stage in Section 5 – so Q_p ranks each application relative to the cohort it was compared to. Q_p therefore measure quality relative to a technology field rather than on an absolute scale. Third, for my identification strategy, Q_p must be orthogonal to the examiner assignment. I report two tests of this assumption. In Table I.7, I show that leniency does not move an art unit's forward technological trajectory. In Section 5.5, I formally test the balance of quality across leniency.

3.4 Validation of the quality measurement

I validate the measure in two steps. Every existing proxy for quality, whether forward citations, renewals, the stock market reaction of Kogan et al. (2017) or the text-based measure of Kelly et al. (2021), exists only for granted patents, while Q_p ranks every application, including those that are

[11] Anchoring the comparison windows on the filing date rather than the publication date is a deliberate design choice. Filing marks the application's actual entry into the USPTO. Publication comes eighteen months after the earliest priority date, so it trails filing by an amount that varies with the filing route, so a publication anchor would assign otherwise identical inventions systematically different comparison sets.

[12] In general, cosine similarities range over $[-1, 1]$. In practice, cosine similarities between these embeddings are strictly positive, so both components lie in the unit interval as stated.

[13] The ratio does not, however, make the two near-perfect complements: the iso-score curves of FS/BS are rays through the origin, so novelty and impact substitute along them at a constant rate. Appendix I decomposes the composite quality measure into its components.

refused. The first step therefore uses an outcome that refused applications have too. A published application can be cited as prior art whether or not it is ever granted. Table I.2 shows that among applications that were never granted, those in the top decile of Q_p receive 41.8% more citations from patents examined in other art units than applications between the 40th and 60th percentiles. This test counts only citations from other art units because Q_p is built from similarity to later filings in the application's own art unit, so citations from the same art unit could reflect textual proximity rather than quality.[14]

The second step compares Q_p with the existing proxies among granted patents, which I show in Figure 1. The net conditional correlation is positive for all four. Moving from the field-year mean of quality to one standard deviation above it raises log citations by 0.071 standard deviations, renewals by 0.015, the stock market reaction by 0.006 and the Kelly et al. (2021) measure by 0.036. Citations count the citations a granted patent receives from later patents. Renewals incorporate the decision to pay maintenance fees, which are due 3.5, 7.5 and 11.5 years after grant. The fees are paid by the patent holder, which a holder presumably only does when the patent is worth more to them than the fee. The Kogan et al. (2017) measure is built from abnormal stock returns around the grant announcement. The Kelly et al. (2021) measure is also built from text, but from word frequencies summed over closely related patents rather than from embedding centroids.

The associations grow toward the top of the ranking. Relative to granted patents between the 40th and 60th percentiles, those in the top decile of Q_p receive 32.2% more citations, those in the top 5% receive 39.8% more and those in the top 1% 54.9% more, as Table I.3 shows. Across the same bins, the stock market reaction is 7.6%, 10.5% and 12.8% higher. Q_p is built without reference to grant outcomes, market values or the grant margin, so nothing in its construction guarantees or rules out these associations.

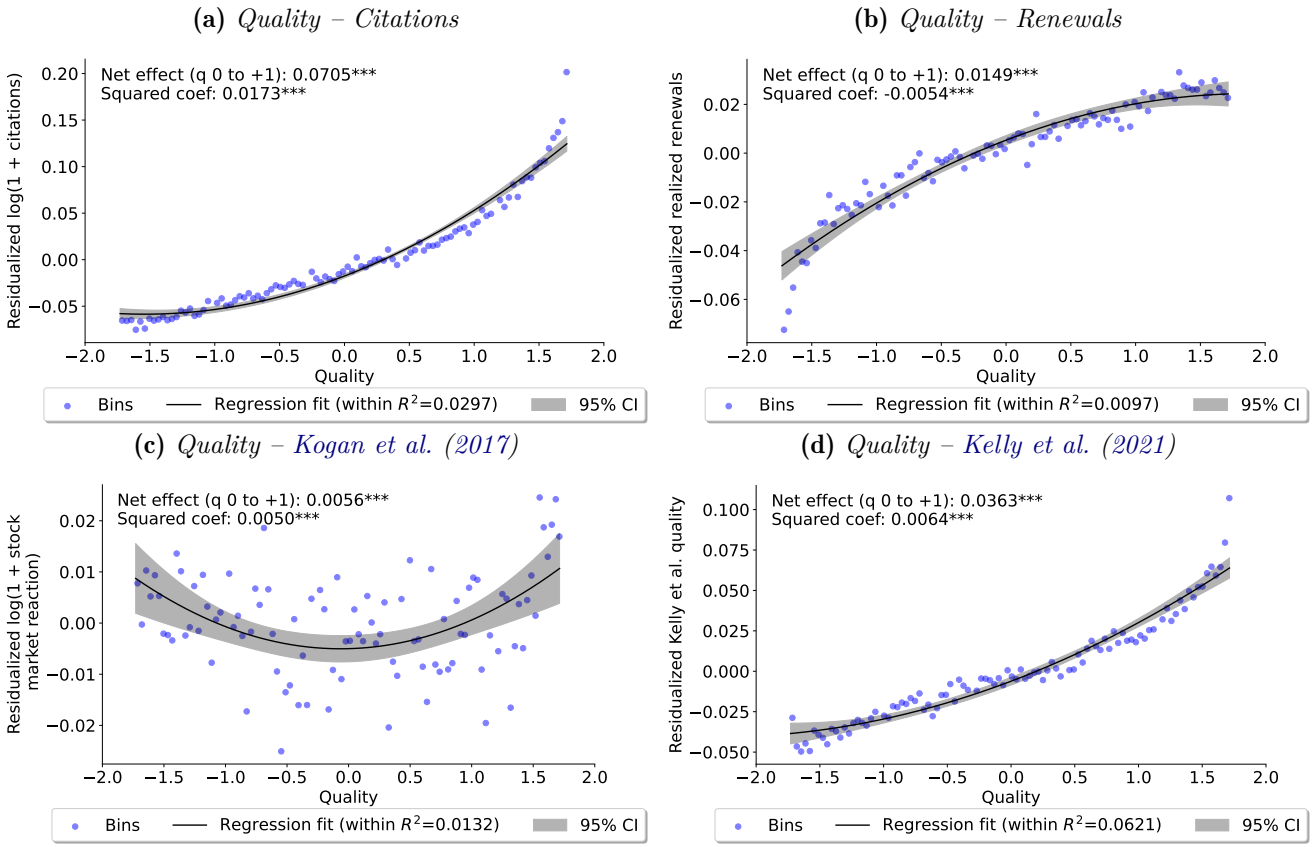
Appendix I adds a comparison that uses neither citations nor market values. Patents on the active ingredients of drugs that the FDA approved as new molecular entities, patents that independent evaluators found essential to technical standards, and the US patents of finalists for the European Inventor Award rank 2.4 percentile points higher on Q_p than random granted applications from the same art unit $\times$ publication year cell.

4 Grants and prosecution by filed scope and quality

The grant rate varies with the filed scope of the application by an order of magnitude more than with the invention's quality. Figure 2 shows conditional correlations for free-standing applications, which claim no earlier US application and do not enter as the national stage of an international

[14] Counting citations from the same art unit, the difference is 64.3%. Distinguishing by who added the citation, it is 35.7% for citations added by examiners and 46.8% for citations added by applicants.

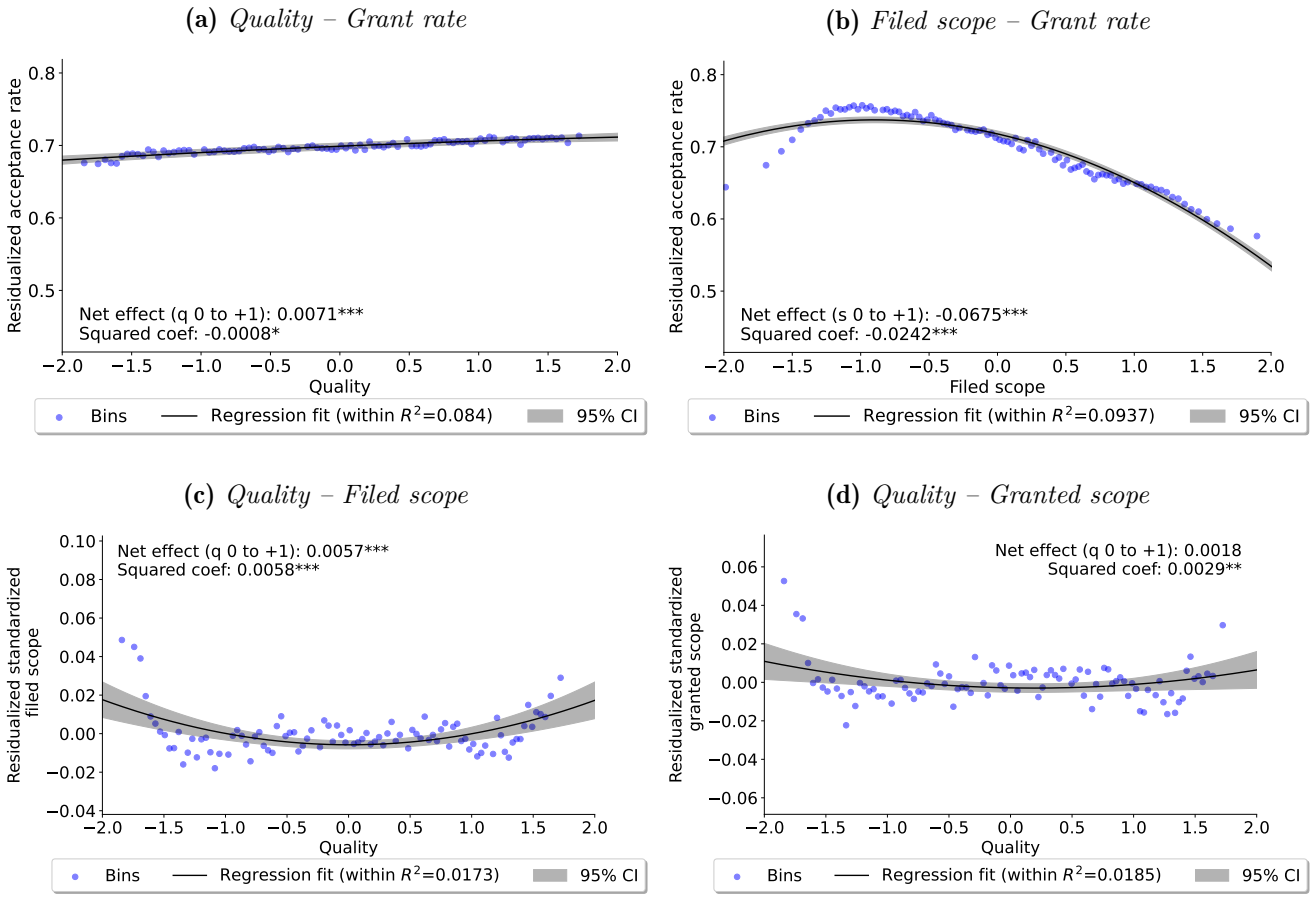
Figure 1: *Conditional correlations – Validation*



Notes: The figure shows the conditional correlations between my quality measure and four existing quality proxies, all of which exist only for granted patents. For comparability across panels, all variables are standardized to mean zero and unit standard deviation and demeaned within art unit × publication year cells. Quality enters as its percentile rank, so the horizontal axis is a linear rescaling of that rank: zero is the median, and one unit corresponds to 29 percentiles. Panel 1a counts the citations a granted patent receives from later US patents. Panel 1b measures renewals as the maintenance fees paid minus the fee decisions observable by June 2023, so a patent renewed at every observable decision takes the value zero. The regression samples contain 3,812,246 grants in Panel 1a, 3,805,365 in Panel 1b, 1,216,203 in Panel 1c and 1,740,472 in Panel 1d. The control variables include small and micro entity indicators, categorical dummies for the type of continuation of a patent application, and dummies for the first five years of examiner experience. The binned scatter plots are produced using `binsreg` (Cattaneo, Crump, Farrell, and Feng, 2025) and covariates are included via auxiliary regressions and residualizing the outcome variable of interest. 95% confidence intervals based on standard errors clustered at the art unit × publication year level are represented by the shaded areas.

application, and the regressions behind them are in Appendix L.1. The claims of these applications are the applicant's first request, while a continuation inherits a claim set that has already been examined and a national-stage entry arrives after an international search. The top row shows how the grant rate varies with quality and with filed scope, and the bottom row shows the associations between quality and scope. Panel 2a shows that the grant rate increases only mildly in quality, by about two percentage points between the bottom and the top decile, on a base of 70%. The binned scatter shows no threshold of quality above which applications are materially more likely

Figure 2: *Conditional correlations – Free-standing applications*



Notes: The figure plots conditional correlations among the grant outcome, invention quality, filed scope and granted scope for free-standing applications, which claim no earlier US application and are not national-stage entries. The regressions are estimated on all applications with every term interacted with the filing group, and the panels show the free-standing coefficients. Appendix Figure D.2 shows the other groups. Quality, filed scope and granted scope are standardized to mean zero and unit standard deviation and demeaned within art unit × publication year × filing group cells. The grant rate is kept in levels. Filed scope is the within-publication-year percentile rank of the average number of words per claim among filed claims, and granted scope its percentile among the filed and granted claims of the same year, both reversed so that a higher value means broader claims. Quality and filed scope enter the horizontal axes as percentile ranks, so each axis is a linear rescaling of the rank: zero is the median, and one unit corresponds to 29 percentiles. The free-standing samples contain 2,900,123 applications in Panels 2a and 2c, 2,901,417 in Panel 2b and 2,024,630 grants in Panel 2d. The controls are a new-examiner indicator, small and micro entity indicators, categorical dummies for the type of continuation, and dummies for the first five years of examiner experience. The binned scatter plots are produced using `binsreg` (Cattaneo et al., 2025) and covariates are included via auxiliary regressions and residualizing the outcome variable of interest. 95% confidence intervals based on standard errors clustered at the art unit × publication year level are represented by the shaded areas.

to be granted. Panel 2b puts the grant rate against filed scope on the same vertical scale. The grant rate increases over roughly the narrowest fifth of filings and falls steadily thereafter, from 77% at the 10th percentile of filed scope to 59% at the 90th. Higher quality goes with almost no additional chance of a grant, but narrower claims go with a much higher one.

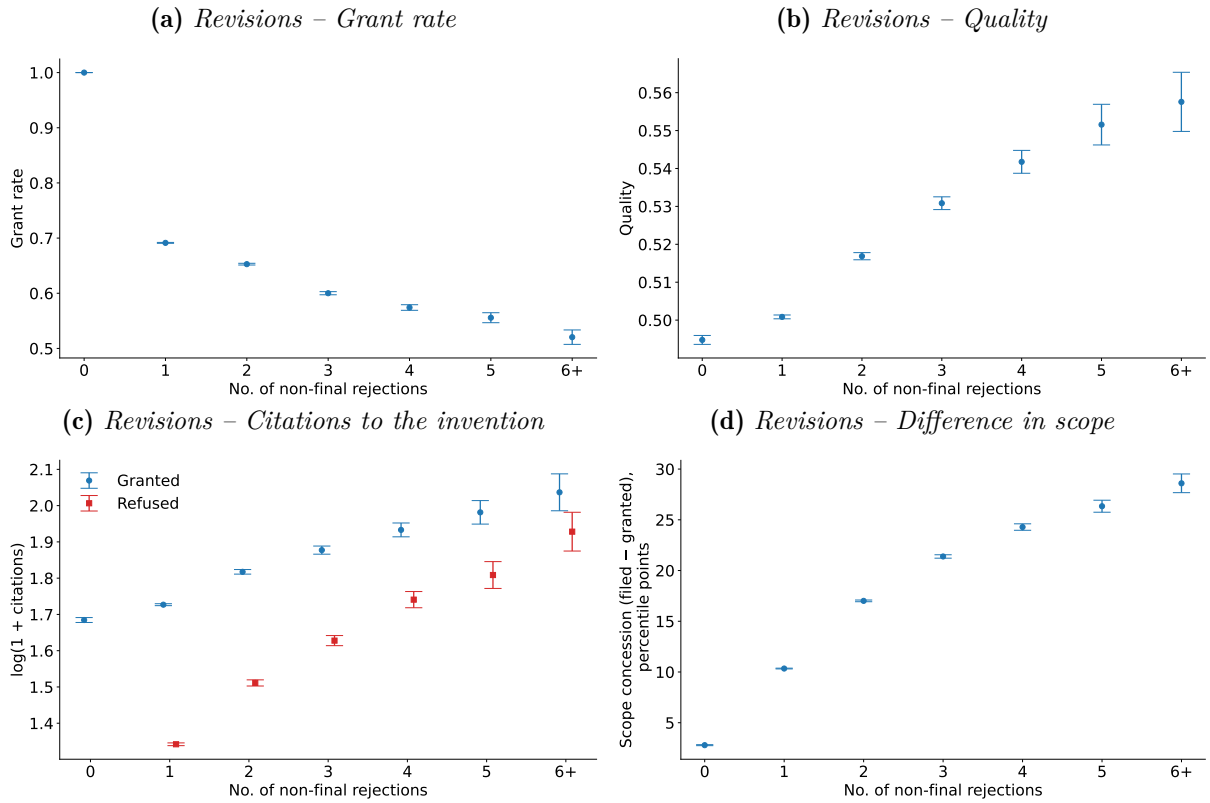
Panels 2c and 2d relate scope to quality. Among free-standing applications, filed scope varies

little with quality. Its profile has a shallow U-shape, with the highest filed scope for the least promising applications and rising again through the upper half. From the median to one standard deviation above it, filed scope increases by 0.17 percentile points. Granted scope bears no statistically distinguishable association with quality. The concession made during prosecution, the narrowing from filed to granted scope, instead increases in quality, as Figure D.1 shows. In Japan, examination narrows the claims of more important applications more often (Okada, Naito, and Nagaoka, 2018). Applications whose claims are not a first ask behave differently, as Appendix Figure D.2 shows. In continuations and in national-stage entries, filed and granted scope both fall with quality, while the concession increases with quality in every group.

Applications differ in how long they stay in prosecution, and those that stay longest are on average the better inventions. Figure 3 plots application attributes by the number of non-final rejections an application received before it was granted, abandoned or finally rejected. The variable on the x-axis, which I use to group observations, is the realized number of rounds. The number of rounds are measured as an outcome of the prosecution, so these profiles compare different populations rather than tracing what an additional round does to a given application. In Panel 3a, I show that the grant rate falls with the number of rounds, from 69% after one rejection to 52% after six or more. The applications that keep going are nonetheless better inventions. In Panel 3b, I highlight that quality increases with the number of rounds, from the 49th percentile for applications allowed without a rejection to the 56th after six or more. Since my measure is built from text fixed at first filing, before any interaction with the examiner, the pattern cannot come from examiners changing the text. At the European Patent Office, applicants withdraw their most valuable applications later (Harhoff and Wagner, 2009). That comparison is in calendar time, while mine counts rounds of examination. Panel 3c shows the same for citations to the invention. Within the same art unit and publication year, citations increase with every additional revision round. This pattern is even stronger among refused applications than among granted ones. Since the increase appears among applications that were never granted, this pattern is not selected on the grant. Citations added by examiners and those added by applicants both increase, as Panels D.3c and D.3d of Appendix Figure D.3 show.

Panel 3d replicates the finding of Marco et al. (2019) that the claims in the application are broader than the claims in the granted patent, and adds that the difference grows with the number of rounds, from 3 percentile points for applications allowed without a rejection to 29 points after six or more. Filed and granted scope behave differently over prosecution length, as Appendix Figure D.3 shows. Filed scope increases with every additional round, from the 40th to the 62nd percentile, so applications that ended after more rounds were filed broader from the start. Granted scope instead peaks after one round and declines thereafter. Within the same issue year, the measures of Kogan et al. (2017) and Kelly et al. (2021) also increase with the number

Figure 3: *Patent and patent application attributes by the number of revisions*



Notes: The figure shows patent application attributes by the number of non-final rejections an application received before it was granted, abandoned or finally rejected, top-coded at six. Panels 3a, 3b and 3d show means. Panel 3c shows log citations to the invention, the later US patents citing the application's publication or, once granted, the patent, separately for granted and refused applications, each at its mean after one rejection plus the difference from it within art unit × publication year. The error bars span 3.2 standard errors on either side, a Bonferroni-adjusted 95% level for 35 simultaneous intervals. Appendix Figure D.3 reports the remaining attributes.

of rounds, while renewals barely change. Applications assigned to a more lenient examiner go through fewer rounds, as Panel D.3i shows. Taken together, longer prosecution goes with a lower grant probability and larger scope concessions, and it is concentrated among the better inventions, whether or not they end in a grant.

These patterns imply an asymmetry in the USPTO's screening process. Applicants pay costs in the form of lower grant probability and more revision rounds for requesting more scope, while higher invention quality yields little compensation. This sharpens the observation of Bryan and Williams (2021) that any application could be granted if its claims were narrowed sufficiently. Patent law does not imply this asymmetry. Each claim describes the invention and sets the

protection requested, and the office tests each claim against the prior art.[15] Because no claim may cover prior art, an invention further from earlier work leaves more room for broad claims. So the asymmetry is a finding about how examination works, not a consequence of the law. On the applicants' side, wider claims are worth more if granted, and a broad first request keeps them within reach. A rejection invites an amendment, so an applicant who files broadly can still narrow the claims in a later round, while a concession, once made, is costly to reverse, as Section 2 describes. This holds for strong and weak inventions alike, so the breadth of a request says little about the invention behind it. Filed scope varies little with quality, as Panel 2c shows. Examination operates on how much is claimed, while the greater value of wider scope pushes applicants to file broadly against it.

A noisy quality measure would also flatten the grant profile. Two facts limit this reading. First, the noise from the documents an application is compared with is small. Appendix I scores each application on two halves of these documents and finds a reliability of 0.94. Second, the same measure separates applications by how long they stay in examination and by how much scope they concede, so it is not blind to what examination does. The measure cannot record whether an invention works, since it reads only the text. The office does not observe this either. It judges the invention from the words of the application and does not ordinarily require proof that the invention works.[16] In Section 5, I establish which applications end up at the resulting acceptance margin by first formalizing the underlying incentive structure and then setting out the identification strategy.

5 Empirical framework and strategy

This section first sets out the minimal structure the empirical strategy needs by describing what determines whether an application clears the examiner's threshold, and therefore which applications are close to the margin, where the assigned examiner decides the outcome. I then clarify the

[15] After the America Invents Act, 35 U.S.C. § 100(j) defines the claimed invention as "the subject matter defined by a claim", and §§ 102(a)(1) and 103 test it against the prior art. Before it, courts compared the prior art with "the claims at issue" (*Graham v. John Deere Co.*, 383 U.S. 1, 17 (1966)). Claim scope is bounded by the prior art, since "that which infringes, if later, would anticipate, if earlier" (*Peters v. Active Mfg. Co.*, 129 U.S. 530, 537 (1889)). A claim that reaches prior art therefore fails however good the disclosed invention is. In the words of the Federal Circuit, an invention "may be outstanding in its field, but the name of the game is the claim" (*In re Hiniker Co.*, 150 F.3d 1362, 1369 (Fed. Cir. 1998)).

[16] Except in cases of perpetual motion, "a model is not ordinarily required by the Office to demonstrate the operability of a device" (MPEP § 608.03), and "it is well settled that an invention may be patented before it is reduced to practice" (*Pfaff v. Wells Electronics, Inc.*, 525 U.S. 55, 61 (1998)). Examiners "have access to little evidence beyond the words of the application" (Freilich and Kim, 2026, p. 543). In applications that contain material from retracted scientific papers, examiners fail to notice the retraction in 97% of cases and do not reject these applications more often than comparable ones (Freilich and Kim, 2026, p. 542). In 2018, after the failure of Theranos' blood-testing technology was widely known, the office granted Theranos U.S. Patent 10,156,579 on detecting an analyte in a blood sample of less than about 500 microliters, and the examiner did not question the claim (Freilich and Kim, 2026, p. 542).

estimation. The framework part is deliberately short. The full dynamic game that I solve and estimate is developed in Section 7. Neither the empirical strategy of this section nor the results of Section 6 depend on it.

5.1 The acceptance threshold

The framework is a sequential game between a patent applicant and an examiner, in four stages. (1) Nature draws invention quality q, which neither side observes. The applicant holds a subjective prior over their own quality, the examiner a prior updated from noisy signals. (2) The applicant, holding beliefs over their own quality and over the leniency of the examiner they will be assigned, chooses filed scope s. (3) An examiner with leniency type l is assigned, receives a noisy quality signal, and observes filed scope directly. The examiner grants a patent application if a latent acceptance index consisting of the quality signal, the filed scope, and their own leniency clears a threshold, otherwise they reject. (4) Both players update by Bayes' rule, and the applicant either continues or abandons for an outside option normalized to zero. If the examiner grants, the game ends with a patent of granted scope $\hat{s} \leq s$.

Following Scotchmer (2004), a potential applicant A is endowed with an invention of quality q and R&D costs, treated as sunk costs at the time of filing. The applicant's strategic choice is the filed scope s of the patent claims submitted to the USPTO. Broader scope increases the value of a granted patent but decreases the grant probability. Additionally, broader scope also increases expected prosecution costs through longer expected revisions.

Scope is chosen before the examiner is assigned, so the applicant cannot condition on l. Writing w for the applicant's belief over (q, l) at the time of filing (p is reserved for the application subscript), the objective is the ex-ante value

$$V(s \mid w) = \underbrace{\mathbb{E}_w\Big[P(q,s,l)\cdot \nu\hat{s}\Big]}_{\text{expected granted value}} - \underbrace{C(s)}_{\text{filing and revision costs}}, \tag{4}$$

where $P(q, s, l)$ is the grant probability, $\hat{s}$ is the granted scope conditional on a grant, ν is the per-unit value of IP rights, and $C(s)$ captures expected prosecution costs. The term inside the expected value is an interim object with some abuse of notation. It describes the value of a granted patent to an applicant who already knew both their own quality and the examiner's leniency. Only the expected value over the applicant's beliefs w is the decision criterion, which becomes a distinction that matters once rejections start signaling l across rounds. I describe the complete problem of the applicant in Section 7.3.

An examiner E has a type of leniency $l \in [0, 1]$. The examiner directly observes filed scope s and receives a noisy signal $\xi = q + \eta$ of the application's quality, where $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$ is independent

of both the filed scope and of the examiner's own type. Importantly, this implies that the examiner forms beliefs about quality from the quality signal alone, so they do not draw inferences about quality from the applicant's choice of scope. The examiner holds a Gaussian prior $q \sim \mathcal{N}(\bar{q}, \sigma_q^2)$ and grants when the latent acceptance index clears a threshold,

$$D \;=\; \mathbf{1}\{\alpha_q\, \mathbb{E}[q \mid \xi] \;-\; \Pi(s) \;-\; \mu \;+\; l \;+\; \varepsilon \;\geq\; 0\}\,, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2_{\text{grant}}), \tag{5}$$

where ε is an idiosyncratic assessment shock independent of η, μ the threshold, and $\Pi(s)$ a scope penalty. α_q describes the dependence of the grant decision on quality and $\Pi(\cdot)$ captures the dependence on filed scope.[17] The posterior mean is $\mathbb{E}[q \mid \xi] = \bar{q} + f\,(\xi - \bar{q})$ with precision weight $f = \sigma_q^2/(\sigma_q^2 + \sigma_\eta^2)$. Substituting $\xi = q + \eta$ separates the index into a part fixed by the application and the signal noise $\alpha_q f\,\eta$. The grant probability of an application with quality q and filed scope s before an examiner with leniency l is therefore

$$P(q, s, l) \;\equiv\; \Pr(D = 1 \mid q, s, l) \;=\; \Phi\!\left(\frac{a(q,s) + l}{\sigma}\right), \quad \underbrace{a(q,s) \;=\; \alpha_q\big[\bar{q} + f\,(q - \bar{q})\big] \;-\; \Pi(s) \;-\; \mu}_{\text{acceptance index}}, \tag{6}$$

with $\sigma^2 = \sigma^2_{\text{grant}} + \alpha_q^2 f^2 \sigma_\eta^2$ the variance of the total assessment noise. σ collects both sources of noise in the examiner's assessment, the shock ε and the noise of the quality signal as it passes through the posterior mean. Leniency enters additively and separably from quality and scope. A lenient examiner tolerates a lower index, instead of reading the same application as better.[18]

5.2 Implications for estimation

Three features of the framework directly shape the empirical strategy. First, filed scope is chosen before the examiner is assigned, so it can serve as a placebo outcome in a randomization test. If assignment is as good as random within art unit × year, the leniency of the assigned examiner must be unrelated to filed scope. I formally test this assumption in Section 5.5. Since filed scope is determined before examiner assignment, it can also serve as a dimension along which to characterize the applications whose grant status depends on the examiner assigned to them.[19]

Second, the additive separability of leniency informs the regression specification for the estimation of the effect of examiner leniency on grant outcomes – the typical first-stage regression

[17] Nothing in the empirical strategy below or in Section 6 uses the functional form of Π, which is laid out in Appendix M.1. The empirical strategy uses the grant rate by filed scope of Section 4 as estimated, which increases over roughly the narrowest fifth of filings and declines thereafter, so no global sign of $\partial a/\partial s$ is assumed.

[18] This separability makes the composition analytically tractable in Section 6.3, and it is the form the estimated model of Section 7 uses.

[19] Essentially, filed scope is a valid control as it is pre-determined. However, I do not use filed scope as a control, but the same way as quality, as one of the dimensions along which I split the data for the nonparametric estimation.

of the examiner fixed effects design. By assuming additive separability, an examiner shifts the index of every application by the same amount and the regression needs no interaction between leniency and the characteristics of the application. Applications still differ in how strongly examiners affect their grant probability, through the first-stage sensitivity $\psi(q,s) = \partial P(q,s,l)/\partial l$, but this is application-specific variation. Additive separability also directly implies the necessary monotonicity assumption. With one threshold and a common leniency effect, a more lenient examiner grants every application a stricter one grants. In Section 5.5, I test the testable parts of both assumptions.

Third, the framework clarifies the population of interest, the compliers. Applications are assigned to examiners within art unit $\times$ publication-year cells c, so the examiners who decide applications in a cell form the pool of examiners an application can be assigned. I measure leniency in grant-probability units, as the examiner's grant rate on the applications of the cell minus the average over the pool, ℓ_e. Each examiner enters this average with their share of the cell's caseload, the probability with which an application in the cell is assigned to them, so ℓ_e has mean zero within the cell. I write ℓ to distinguish this measure from the index leniency l of Equation (6). The reference examiner is an examiner at the pool's average grant rate, $\ell = 0$. Since the potential grant outcome $D_p(\ell)$ is non-decreasing in leniency, each application p has a critical leniency ℓ_p^*, the lowest leniency at which it is granted. An application is a complier if its assigned examiner and the reference examiner fall on opposite sides of ℓ_p^*, and a treated complier if its assigned examiner grants it and the reference examiner would not. Comparisons between the most and the least lenient judge, the convention of the judge-design literature (Dahl, Kostøl, and Mogstad, 2014; Dobbie, Goldin, and Yang, 2018; Agan, Doleac, and Harvey, 2023), count every application that some examiner in the office would reclassify. My definition is a strictly smaller, baseline-relative set, as it counts applications that the examiner an applicant could expect would have decided differently. Throughout, "complier" refers to this reclassification event. My empirical strategy is silent on always-takers and never-takers, whose grant status no examiner assignment can shift.

Together, these features make treatment-effect heterogeneity a single object, the distribution of compliers over (q,s). This distribution is jointly determined by quality and scope choice, which motivates the two-dimensional characterization. The following subsections describe the estimation.

5.3 The target and its estimator

Target. For a set B of values of a predetermined characteristic $X \in \{q,s\}$, or a cell of the quality $\times$ filed-scope plane, the mass of treated compliers in B is

$$T(B) \;=\; \Pr\Big(D_p(\ell_{e(p)}) = 1,\; D_p(0) = 0,\; X_p \in B\Big), \tag{7}$$

where $e(p)$ is the examiner assigned to the application and $D_p(0)$ its grant decision before the reference examiner. The share of the patents granted in B that the reference examiner would have refused is $T(B)/\Pr(D_p = 1, X_p \in B)$. Over all applications the mass is T and the share is $T/\Pr(D_p = 1)$. The reverse event, refused by the assigned examiner and granted by the reference examiner, yields the share among refused applications, and both events together yield the flip rate.

Assumptions. Four assumptions identify these objects. (i) Within a cell and given the controls W, the examiner an application is allocated to is independent of its potential grant outcomes among the applications in the same set B. Examiners may be sorted on the characteristic that defines B. The controls are small and micro entity status and the type of continuation. The reference examiner is defined over the whole pool, so, for example, being assigned a junior or a senior examiner is part of the variation the design uses. (ii) Monotonicity implies that within a cell, an examiner who grants more of the cell's applications grants every application that a stricter examiner also grants. (iii) Within each set B, the grant probability is linear in leniency over the full pool of examiners. (iv) Within a cell, the dispersion of leniency among the applications in B is the same fraction of its dispersion among all applications of the cell for every set B.

The aggregate share of patents accepted due to a lenient examiner needs only assumptions (i) and (ii). An examiner with $\ell_e > 0$ grants a share ℓ_e more of the cell's applications than the reference examiner, and under monotonicity every one of them is a treated complier (Imbens and Angrist, 1994; Goldsmith-Pinkham et al., 2026), so

$$T \;=\; \mathbb{E}\left[(\ell_{e(p)})_+\right]. \tag{8}$$

The equation is exact when the examiner assignment is independent of the application within the whole cell. If examiners sort on a characteristic, it holds up to a term that Section 5.5 bounds. For a set B, the reference examiner's grant rate in B is not observed. Under (iii) it equals the pool's average grant rate in B, and

$$\frac{T(B)}{\Pr(X_p \in B)} \;=\; \psi(B)\, d(B), \qquad d(B) \;=\; \mathbb{E}\left[(\ell_{e(p)})_+ \,\middle|\, X_p \in B\right], \tag{9}$$

where $\psi(B)$ is the sensitivity of the grant to leniency in B. By the definition of ℓ, ψ averages to one over the population. Assumption (iii) is what links the instrumental-variables comparisons below to the reference examiner. An instrumental-variables estimator weights each critical leniency by the deviation of the examiners beyond it, on both sides of the pool mean. $T(B)$ instead counts the applications between the reference examiner and the assigned examiner. The two agree when the grant probability is linear in leniency, and Section 5.4 tests this.

Instrument. The instrument is defined via the unbiased jackknife instrumental-variables estimator (UJIVE; Kolesár, 2013) built from examiner × publication-year indicators, $U_p = \hat{D}^{[Z,W]}_{-p} - \hat{D}^{W}_{-p}$. It leaves application p out of both the examiner projection and the control projection. As the indicators are year-specific, U_p estimates the assigned examiner's leniency in the application's publication year, $\mathbb{E}[U_p \mid e(p)] = \ell_{e(p)}$, and its estimation noise does not involve D_p. The estimator is consistent under the many-instrument and many-control asymptotics that apply here.

Sensitivity. I demean U within cell × bin, so that examiners are compared only within the bin, and take its covariance with the grant among the applications in B:

$$\hat{\psi}(B) \;=\; \frac{1}{\hat{V}(B)} \, \frac{\sum_p \tilde{U}_p \, \mathbf{1}[X_p \in B] \, D_p}{\sum_p \mathbf{1}[X_p \in B]}. \tag{10}$$

Here $\tilde{U}_p$ is the demeaned instrument and $\hat{V}(B)$ the variance of leniency in the art units of the applications in B. The numerator is the instrumental-variables moment of the modified outcome $\mathbf{1}[X_p \in B] \cdot D_p$ of Goldsmith-Pinkham et al. (2026). Without the demeaning, a cell whose examiners are assigned different mixes of X would have that difference credited to the compliers. The variance of leniency comes from the same instrument: the covariance of U with the grant within a cell estimates it without any linearity, because the leave-out noise is independent of D_p. I estimate it for each art unit, pooling art units with fewer than 2,000 applications within their technology center.

Level. Like other nonlinear functions of decision makers' effects, the aggregate dose $\mathbb{E}[(\ell)_+]$ cannot be computed directly from noisy estimates of each examiner's leniency (Arnold, Dobbie, and Hull, 2022; Angelova, Dobbie, and Yang, 2026), and it has no estimator free of an assumption on the distribution of leniency. Two bounds on the dose, however, need no such assumption, and when leniency and the noise are jointly normal, as in parametric empirical Bayes (Morris, 1983), their geometric mean equals the dose. I report this geometric mean as the point estimate. The lower end of the bracket is the lower bound computed with each examiner's career-long leniency instead of the yearly one, which is less noisy but ignores how an examiner's leniency changes from year to year (Frakes and Wasserman, 2017). Within a bin, $d(B)$ is computed from the mean and variance of leniency among the applications in B. The profiles report how the share changes across bins, and their level comes from the aggregate: each profile is multiplied by one factor so that its grant-weighted average equals the aggregate share. The level of every profile therefore rests on assumptions (i) and (ii) alone, and its shape on (iii) and (iv). Appendix E derives the two bounds and validates the point, the bracket and the bin profiles on simulated data with a known target.

5.4 Flip rates and complier shares

The share of the patents granted in B that the reference examiner would have refused combines the components of Section 5.3,

$$\phi_c^{D=1}(B) \;=\; \frac{\psi(B)\; d(B)}{P[D=1 \mid X \in B]}. \tag{11}$$

The flip rate, the share of all applications in B whose grant status depends on the examiner assignment in either direction, replaces the one-sided dose by the two-sided one,

$$\phi_c(B) \;=\; \psi(B) \cdot \mathbb{E}\Big[\, |\ell_{e(p)}| \,\Big|\, X_p \in B \Big]. \tag{12}$$

Weighted by grants, Equation (11) averages to the aggregate share $T/\Pr(D_p = 1)$.

Resolution and inference. I report each profile in twenty equal-mass bins and as a curve, with a Gaussian kernel of bandwidth 0.07 on the percentile axis in place of the bin indicator. The joint profile uses a quality × filed-scope plane of eight equal-mass bins per axis, with the instrument demeaned within cell × four-by-four blocks of the plane so that every block holds enough applications for the comparison. Standard errors come from the joint influence function of all components, the dose and the grant rate included, clustered on the cell. The curves carry simultaneous bands from a Gaussian multiplier bootstrap.

Checks. The first check tests assumption (iii). Within each cell, I sort applications into five arms by the leniency of their examiner. In each quintile of the characteristic, I estimate the sensitivity twice, once from the three most lenient arms and once from the three strictest. If the two profiles differ jointly at the five percent level, and the lenient profile differs from the linear one by more than 1.5% in some quintile, I also report the profile estimated from the lenient arms alone, which counts the treated compliers the way $T(B)$ does. The second check reruns the estimator with two other instruments: examiner indicators that pool each examiner's career, and the instrument of Appendix E, which conditions on examiner experience. Appendix E reports a third check, which tests assumption (iv).

5.5 Identifying assumptions and their tests

Reading the objects of Sections 5.3 and 5.4 as complier shares requires assumptions (i) and (ii) of Section 5.3: that applications are assigned quasi-randomly to examiners within art unit × year, and monotonicity, that no application is granted by a stricter examiner but rejected by a more lenient one. The identifying variation is which examiner an application is assigned to within its

cell. Conditioning on art unit × year fixed effects absorbs the technology field, the competitive environment, and the annual composition of the examiner pool. Within these cells, every complier object uses the examiner × year indicators through the leave-out instrument U. I discuss both assumptions here, and the first check in Section 5.4 tests assumption (iii).

Quasi-random assignment. Judge fixed-effects designs for patent-related outcomes at the USPTO rest on two premises: examiners differ in leniency, and applications are assigned to them quasi-randomly within an art unit. Under quasi-random assignment, examiners in the same unit receive, in expectation, the same mix of applications, so their differing grant rates reflect different standards rather than different caseloads. However, Righi and Simcoe (2019) raise doubts about the validity of these instruments. They find evidence for examiner specialization within technology groups, and specialized examiners are also less lenient, which could violate the exclusion restriction. I test balance with the instrument of Section 5.3 and the test of Goldsmith-Pinkham et al. (2026): under quasi-random assignment, the mean characteristic of the compliers is the same whether it is computed from the granted or from the refused applications. The two means differ by 1.9 percentiles of quality and by 1.5 percentiles of filed scope, with z-statistics of -17.9 and -14.3, and the difference is significant in 9 and in 10 of twenty equal-mass bins. Assignment is therefore not exactly random with respect to either characteristic, but the imbalance is small. The gaps are 0.07 and 0.05 standard deviations of the characteristic, smaller than the gap of 0.11 standard deviations in a patent-class indicator that Goldsmith-Pinkham et al. (2026) report in their reanalysis of Farre-Mensa et al. (2020). The test detects them because the sample is large. With the 32,514 applications of that reanalysis, it would detect them with probability 0.29 and 0.20. Appendix F compares the two designs in detail. Examiner experience accounts for most of the quality imbalance. Within cells, examiners with less than five years of experience decide applications one to three percentiles higher in quality and grant 9 to 21 percentage points less often than more experienced examiners, and an instrument that compares examiners only within the same experience level removes 86% of the quality imbalance. For the pass-through estimator of Appendix E, Appendix F traces the scope imbalance to the composition of caseloads: an examiner whose caseload holds many broad filings grants less often without being stricter.

The estimator does not require balance on quality or filed scope. The profiles compare examiners within cell × bin, so sorting on a characteristic does not enter the sensitivity in that characteristic's profile, and residualizing the instrument also on the other characteristic moves no profile by more than 0.05%. Sorting enters the dose instead, as the mean leniency of the examiners assigned to the applications of a bin. The reference examiner is the same for every application of a cell, so an application of a kind that tends to be assigned stricter examiners is less often a treated complier. This location term accounts for 40% of the decline of the quality profile between the

10th and the 90th percentile. Without it, the share falls by 4.6% instead of 7.6% across quality and rises by 50% instead of 46% across filed scope. For the aggregate, letting examiners sort on quality changes the share by 0.02% of its value, and letting them sort on filed scope changes it by 0.06%, because the mean leniency of the assigned examiners differs little across bins.

Assumption (i) still requires that, among applications of the same cell and bin, the examiner assignment is unrelated to the application's potential grant outcomes. I test its observable part on three characteristics fixed before the examiner is assigned: the number of claims and the length of the first claim as filed, and the month of publication. Within cell $\times$ bin, the correlation of the assigned examiner's leniency with each of them is at most 0.012 in absolute value, although most correlations are statistically significant at this sample size. Residualizing the instrument on all three moves no bin of either profile by more than 0.3% and the aggregate share by 0.03%. Specialization by technology within an art unit, the concern of Righi and Simcoe (2019), enters these tests only through quality and filed scope.

Monotonicity. The LATE framework requires that any application granted by a less lenient examiner would also have been granted by a more lenient examiner. The joint test of the monotonicity assumption and the exclusion restriction proposed by Frandsen, Lefgren, and Leslie (2023) requires a second-stage outcome. For my object of interest, I estimate only a first stage, so I test a weaker implication. Under monotonicity, the mass of compliers in every bin of a predetermined characteristic is non-negative (Goldsmith-Pinkham et al., 2026). With the instrument of Section 5.3, the estimated complier mass is positive in each of ten bins of quality and of filed scope, with t-statistics of at least 38, and the joint one-sided test of non-negativity does not reject, with $p = 1.00$ on both axes. Appendix F reports the average monotonicity tests of Frandsen et al. (2023) across publication years, art units and quality percentiles for the pass-through estimator.

Another related concern is that leniency might be two-dimensional, such that an examiner could be lenient on quality but strict on scope. This would generate defiers specifically along filed scope. In Appendix F Table F.2, I tackle this concern and show that examiners' leniency ranks are robust across scope. I separately compute the leniency of an examiner on the narrow and the broad halves of their own caseloads and find that the two rankings agree to 96% of the ceiling that sampling noise allows. Scope-specific defiers are therefore rare.

6 Results from the examiner assignment

6.1 Overall complier share

In my sample, 7.1% of granted patents owe their grant to an examiner more lenient than the average examiner the application could have been assigned to. These are grants that the examiner the application could have expected would have refused. de Rassenfosse et al. (2020) estimate that 1.2% of granted US patents would be refused under the office's own standard, using twin applications filed at several national patent offices. Their figure comes from disagreement between offices rather than from variation in examiners' decisions within one office, and Appendix E compares the two constructions.

The 7.1% lies between a lower bound of 6.1% and an upper bound of 7.9%, neither of which needs an assumption on the distribution of leniency, and its 95% confidence interval runs from 7.0% to 7.2%. With examiner indicators that pool each examiner's career, the share is 6.2%, and with an instrument that conditions on examiner experience, it is 5.7%.

The natural external comparison of the overall complier share is with the judge fixed-effects designs in the economics of crime, which report that 8 to 13% of their samples are compliers. However, those shares measure a different object.[20] Estimating a more similar object, I find that examiner dispersion at the USPTO is wide: across the examiners an application can be assigned, leniency has a standard deviation of 13 percentage points of grant probability.[21]

I focus on the share of compliers among granted patents, but the mirror image also exists: the share of applications rejected only because a strict examiner was assigned. I discuss this share, the share of compliers among rejected applications, in Appendix E.

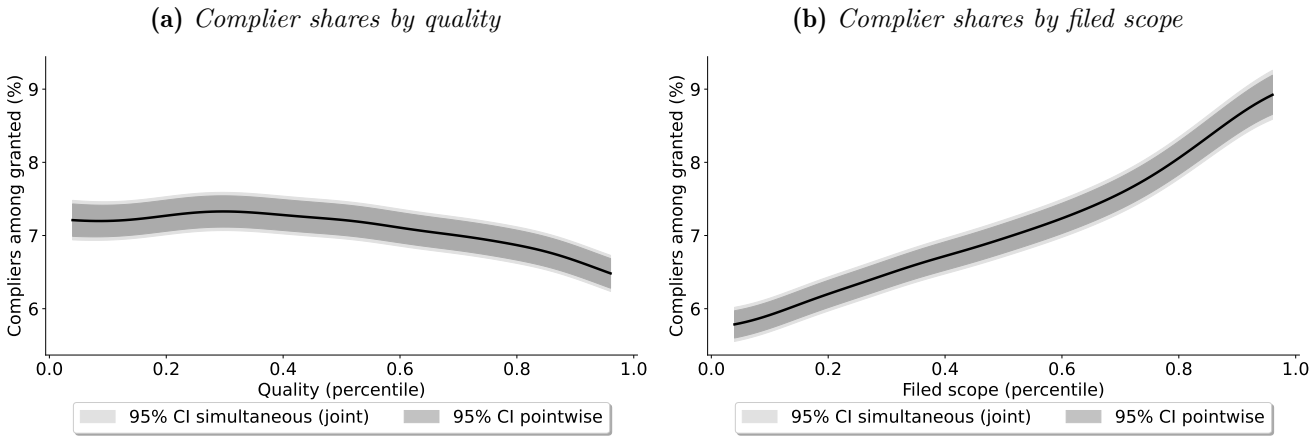
6.2 Rising in scope

The share of compliers among granted patents increases with filed scope. Figure 4 plots it across quality and across filed scope. In Panel 4b, the share increases from approximately 5.9% among the

[20] The conventions differ in both the contrast and the denominator. In the economics of crime literature, the contrast is a full-sample difference in treatment probability between the most and least lenient decision-maker, conventionally the ninety-ninth and first percentiles of the leniency distribution. My contrast is the treated-complier mass, measured against the expected leniency as defined in Section 5.2 and reported as a share of granted patents. Dobbie et al. (2018) identify 13% of their sample as compliers under a local linear first stage and 11% under a linear one, and Agan et al. (2023) approximately 10%, and $8-10\%$ across the cutoffs and specifications they report. An extreme-versus-extreme contrast counts more marginal cases by construction than a contrast against the expected leniency. It is therefore not surprising that I identify a smaller number here.

[21] An institutional asymmetry distinguishes the two settings as well: a defendant cannot change the case a judge rules on after the crime has been committed. The treatment is assigned by the judge, and defendants cannot affect the "broadness" of their case at the time the judge makes their decision. USPTO applicants choose filed scope before the allocation of the examiner, and that choice concentrates them near the grant margin, a mechanism with no analog in pretrial detention.

Figure 4: *Share of compliers among granted patents*



Notes: The Figure plots the share of compliers among granted patents, $\phi_c^{D=1}$, across the percentiles of quality and of filed scope. Estimates come from the estimator of Section 5.3, Equation (11), with the instrument built from examiner × publication-year indicators and a Gaussian kernel of bandwidth 0.07 in place of the bin indicator. Each profile is multiplied by one factor so that its grant-weighted average equals the aggregate share of 7.1%. The shaded areas show 95% pointwise (dark gray) and 95% simultaneous (light gray) confidence bands, with the scaling factor held fixed. The outermost percentiles, below 0.03 and above 0.97, are kernel-edge regions and are trimmed. Both panels share one vertical axis.

narrowest filings to approximately 8.6% among the broadest, an increase of roughly 46%. Along filed scope, the arm check of Section 5.4 rejects linearity, and the profile estimated from the lenient arms alone rises more steeply, as Appendix Figure E.6 shows. Panel 4a shows the quality profile, which falls from approximately 7.2% among the lowest-quality applications to approximately 6.7% among the highest, by about 7.6%.[22] Compliers are broader applications at every level of quality.

The scope gradient holds at every level of quality. On the two-dimensional surface, which holds quality fixed, the share of compliers among granted patents is higher for the broadest than for the narrowest filings in every quality band, by 2.8 to 3.8 percentage points. Holding filed scope fixed instead, the share falls with quality: the highest quality band lies below the lowest in all eight scope bands. I plot the joint surface in Appendix Figure D.5 and discuss it in Appendix D.

Appendix K shows what the composition of compliers implies for existing work that uses the examiner assignment, re-examining the settings of Sampat and Williams (2019) and Farre-Mensa et al. (2020) with marginal treatment effect curves. The integrated average effect lies below the LATE the paper's own specification delivers in every Farre-Mensa et al. (2020) outcome, and is statistically indistinguishable from it in Sampat and Williams (2019).

Robustness. I test the quality profile across multiple decisions in the construction underlying the result. In Appendix I, particularly Table I.4, I show that the profile of the composite measure is the result of offsetting components rather than an absent signal. Decomposing the quality measure into

[22] A simultaneous test rejects flatness of the complier share along both characteristics, with $p < 0.001$.

its novelty (backward-looking) and impact (forward-looking) parts, shows that novelty decreases the chances of a grant and impact increases it. Estimated on novelty alone, the complier share rises by about 15% across the distribution, and on impact alone it falls by about 17%, against the decline of about 7.6% in the composite. In Appendix I, I show that rebuilding the measure under eight alternative construction choices (window lengths and gaps, comparison pools, and the choice of the decision date) leaves the profile declining under every choice, by between 3% and 10%, with flatness rejected under all eight. Two constructions that compare each application only with its closest documents, its ten nearest neighbours or its similarities above a cutoff as in Kelly et al. (2021), give declines of 8.1% and 6.1%. Appendix I also bounds the attenuation from the noise in the comparison documents.[23]

Apart from questions about the construction of my measure, I also test the robustness of the compliers profiles to sample composition. Continuation applications inherit prosecution history from a parent application, which could bias the estimates if their grant probability differs from that of original applications at the same quality, or inflate the effective leniency variation through the within-examiner correlation between parent and child filings. Without continuations, the share rises by 45% across filed scope and falls by 11% across quality. Restricting the sample to the art unit $\times$ year cells in which the USPTO allocated applications by random terminal docket digit (Feng and Jaravel, 2020), about a fifth of the sample, the share rises by 49% across filed scope and falls by 18% across quality. Appendix G reports both restrictions in detail.

Repeat filers raise a separate concern. One law firm or patent department prosecutes many applications, so the instrument could partly reflect the firm rather than the examiner. Leaving every application of the same customer number out of the instrument, or adding customer-number fixed effects to the controls, leaves the share rising in filed scope, with a quality profile that does not reject flatness. I report both checks in Appendix H.

Finally, an examiner who is handed an unusually broad caseload would have a lower grant rate without necessarily being stricter. Since grant probability falls steeply in filed scope, measured leniency carries a caseload component. In Appendix F, I show for the pass-through estimator of Appendix E that purging this component from the measure only steepens the observed gradient of the complier share in scope, while not affecting the share across quality. Appendix F states the argument formally and signs the bias.

[23] I score each application twice on independent halves of the documents it is compared with, which results in a full-corpus rank reliability of 0.94. This reliability bounds the attenuation from that noise, so that the true gradient can exceed the observed one by at most seven percent, against an observed decline of about 7.6%.

6.3 Mechanism: the complier composition is an inverse Mills ratio

The composition documented above follows from a single feature of the grant decision. Compliers are the applications close to the grant margin, the ones the examiner assignment can flip, and that margin lies at a threshold in an index that weights filed scope more than the examiner's assessment of invention quality. The asymmetry is institutional: the main rejection reasons require examiners to assess whether a claim covers a novel and non-obvious step, and whether a claim covers more than the application discloses. Filed scope therefore sets an application's distance to the margin, and quality enters only through the weight the acceptance index places on the examiner's assessment. The applicant chooses filed scope before the examiner is assigned, so the applicant's own choice of filed scope sets this distance. One object, the sensitivity of the grant probability to the assigned examiner's leniency, then delivers two findings: the increase in filed scope and the smaller movement of the flip rate.

Under the threshold rule (5) of Section 5.1, the acceptance index $a(q, s)$ of Equation (6) collects the effects of invention quality and filed scope on the grant outcome. Since leniency enters the index additively, as Section 5.2 describes, the first-stage sensitivity depends on (q, s) only through the index,

$$\psi(q,s) \;=\; \frac{\partial P}{\partial l}\bigg|_{l=\bar{l}} \;=\; \sigma^{-1}\varphi\left(\frac{a(q,s)+\bar{l}}{\sigma}\right), \tag{13}$$

which is the density of applications at the grant margin. Dividing the treated-complier mass $\psi \cdot \bar{d}$ by the grant rate P from Equation (11) then gives the share of compliers among granted patents, up to the leniency dose $\bar{d}$, as

$$\phi_c^{D=1}(q,s) \;\propto\; \frac{\varphi\Big((a+\bar{l})/\sigma\Big)}{\Phi\Big((a+\bar{l})/\sigma\Big)} \;=\; \lambda\Big((a(q,s)+\bar{l})/\sigma\Big), \tag{14}$$

the *inverse Mills ratio* of the acceptance index. Since $\lambda(\cdot)$ is strictly decreasing, the entire complier composition is a monotone transformation of an application's position in the grant distribution. Every profile is evaluated at the examiner's pool mean leniency $\bar{l}$, which enters as a constant added equally to every application's index. Such a common shift cannot reorder applications, so the ordering of the composition over (q, s), and with it the sign of every gradient, depends on $a(q, s)$ alone and not on the choice of reference. [24]

[24] The invariance stops at the ordering because λ is nonlinear: a shift common to every argument is not a shift common to every value, so a change in $\bar{l}$ moves the level and the steepness of a profile. Magnitudes therefore carry the reference $\bar{l}$, while orderings and signs do not.

In this model, the derivative of the grant probability with respect to leniency is the density of applications at the grant margin, which is the marginal treatment effect weight of Heckman and Vytlacil (2007). A threshold-crossing model alone adds no testable restriction to the assumptions behind the instrumental variables estimator (Vytlacil, 2002), while a threshold-crossing model with enough parameters to match every grant probability yields numerically identical estimates of the local average treatment effect (Kline and Walters, 2019). The following proposition adds a restriction. Since leniency enters additively and the noise has one common distribution for every application, the share of compliers among granted patents depends on filed scope and quality only through the acceptance index. The share of compliers among granted patents increases as the index decreases and vice versa. The grant rate increases with the same index, so the complier share must be higher wherever the grant rate is lower. Any finding that violates this pattern falsifies the single-index structure.

Proposition 1 (Complier composition)**.** *Under the threshold grant model with additively separable leniency and Gaussian noise, (i) the share of compliers among granted patents $\phi_c^{D=1}(q,s)$ depends on filed scope and quality only through the acceptance index $a(q,s)$ and strictly decreases in it, so that it is higher wherever the grant probability is lower, and (ii) the unconditional flip rate* $\Pr(\textit{complier} \mid q,s) \propto \varphi\Big((a+\bar{l})/\sigma\Big)$ *is single-peaked in the acceptance index, attaining its maximum where the grant probability equals one half.*

Proof in Appendix B.

Proposition 1 turns the two grant-rate patterns of Section 4 directly into the two complier facts of Section 6.2. Along filed scope, the conditional grant rate across all applications decreases strongly, dropping from 77% at the 10th percentile of filed scope to 63% at the 90th. Therefore, the acceptance index decreases, and by part (i) the complier share increases from approximately 5.9% to 8.6% between the same percentiles.

Part (i) applies at every level of invention quality, because broader filed scope lowers the acceptance index at any fixed quality. Whatever the invention, an applicant who chooses to file more scope moves the application closer to the threshold, so the complier share increases. The two-dimensional surface of Section 6.2 shows this increase within every quality band. Applicants do not choose their examiner, but through the scope they file they choose how likely their grant depends on the assigned examiner. Thereby, applicants select into how strongly the examiner assignment can affect their patent application.

Along quality, part (i) predicts at most a mild decline. Holding filed scope fixed, a characteristic that raises the acceptance index lowers the complier share, but the grant rate increases only by one to two percentage points across the quality distribution, so quality moves the acceptance index little. At fixed filed scope on the quality-scope surface, the complier share is lower for the highest than for the lowest quality band in all eight scope bands, and the one-dimensional profile

decreases from 7.2% to 6.7% between the 10th and the 90th percentile. The grant rate and the density of applications at the grant margin, the two terms through which part (i) works, account for three fifths of that decline, the grant rate alone for a sixth. The remaining two fifths come from the location of the leniency dose, which reflects the stricter examiners that high-quality applications tend to be assigned to, as Section 5.5 shows. The proposition holds the dose fixed across applications, so this part lies outside of it.

Part (ii) disciplines the two-sided flip rate of Equation (12). Every cell of the two-dimensional surface has a grant probability above one half, in a range of 60% to 78%, so the data lie entirely on one side of the peak at $P = \frac{1}{2}$. The flip rate should therefore increase as the acceptance index falls, but never reach its maximum. I show this in Appendix Figure D.5. The flip rate is lowest in the cell combining the highest quality with the narrowest claims, at 7.8%, and highest in the cell combining the lowest quality with the broadest, at 11.3%. Across the surface, the flip rate is higher where the cell grant rate is lower.[25] Comparing this to the share of compliers among granted patents, which is divided by $\Phi(t)$, shows that a larger part of the gradient in the complier share comes from the denominator $\Phi(t)$. Along filed scope, the grant rate across all applications, decreasing from 77% to 63%, accounts for a little over half of the log change in the share from 5.9% to 8.6%, the density for about a third, and the location of the leniency dose for the remaining eighth. I show this decomposition at every decile of filed scope in Appendix Table L.4. For the applicant, broad filed scope is therefore less likely to be granted, and when it is, the chances are higher that a lenient examiner granted it.

An alternative interpretation attributes the scope gradient in the share of compliers to claims that examiners find hard to assess: broad claims might be harder to evaluate, so that their grant depends more on which examiner reads them. This "hard to judge" reading is a statement about the variance of the examiner's assessment. In Equation (6), I hold σ fixed across applications for simplicity. However, this is not necessary for the design and σ can be separately estimated. A probit of the grant outcome on the leniency deviation, run within cells of quality and filed scope, has slope β/σ, where β is the pass-through of leniency into the acceptance index. The art unit $\times$ year pool effect enters that index additively rather than through β, so its coefficient in the same probit varies with σ alone. The two slopes together separate β from σ. In Appendix B.1, I report the estimates in Table B.1. The results show that assessment noise falls by about a sixth across quality and increases by about a sixth across filed scope. Weighting applications by the flip rate of Equation (12) puts the compliers' assessment noise 0.32% above the population mean, so assessment noise cannot explain the observed scope gradient in the share of compliers. In Appendix B.1, I additionally, show that assessment noise is not only too small to create the observed complier distribution, its effect even points in a different direction.

[25] The flip rate and the grant rate have a between-cell Spearman rank correlation of -0.85.

7 A model of prosecution

The comparison of examiners holds the scope that applicants requested fixed and it cannot show how applicants would respond to a stricter patent office. The model adds the applicant's side. Its purpose is to simulate a stricter patent office, holding fixed the set of filed inventions, and to split each result into the examiners' decisions under a new rule set and the applicants' response.

Examination in the model is a game between an applicant and an examiner in which only the applicant makes choices. The examiner does not optimize. Each examiner applies the office's grant rule, shifted by their own leniency, which is the examiner's type. The rule weighs the examiner's assessment of the invention's quality, which improves with a noisy signal in every round, against the filed scope. Applicants differ in two types, drawn independently of each other: the quality of their invention and their cost of a revision round. The costs to revise may differ between applicants, for example, between an individual inventor and a large firm. Applicants neither directly observe the quality of their invention nor the leniency of their examiner. They hold a prior over their own quality and know the distribution of leniency across examiners, and each rejection updates their joint belief over the two. The model explains the requested scope and the outcomes of applications given their quality and the leniency of the examiner they are assigned to.

The applicant chooses the scope to file at the start and again after every rejection, when they either file a claim no broader than the one rejected or abandon the application. A broader claim is worth more if granted but decreases the chance of a grant in each round, so it takes more rounds of revision in expectation. At the USPTO, even a final rejection can be reopened by a request for continued examination, so a rejection does not end an application in the model. It ends with a grant or when the applicant abandons, so prosecution lasts as long as the applicant is willing to pay for another round. Quality enters this trade-off on the applicant's side. A better invention raises the value of a grant relative to the cost of each round, so applicants with better inventions continue despite more rejections and can afford broader claims. I call these two effects persistence.

Three sets of parameters govern this trade-off. The price of scope is how much a broader claim decreases the chance of a grant in each round, and it is identified from how the grant rate changes across filed scope among applications of the same quality before examiners of the same leniency. The cost of a revision round is identified from how long applicants keep negotiating. How much more a grant is worth to better inventions relative to that cost comes from how filed scope and the number of rounds vary with quality, and the grant rate by quality adds to it. Given the price and the cost, the level of filed scope reveals the value of scope relative to these costs. The other parameters describe how the examiner weighs and learns about quality, how the chance of a grant changes across rounds, how much of the filed scope the examiner grants, and how filed scope departs from the applicant's optimum. Section 7.1 sets out the primitives and the timing,

Sections 7.2–7.4 the two sides and the solution, Section 7.5 takes the model to the data, and Section 7.7 runs the counterfactuals, including the stricter office.

7.1 Environment, states, and timing

Time within an application is discrete, indexed by revision rounds $r = 0, 1, \ldots, R-1$, with $R = 12$, and a common discount factor $\beta = 0.95$. R is a finite-horizon approximation chosen to be well above the observed round distribution, not a statutory limit.

An application starts from three draws made before filing: the quality of the invention q, the applicant's revision-cost type $\zeta \sim \mathcal{N}(0, \sigma_c^2)$ and a filing taste $u_s \sim \mathcal{N}(0, 1)$. Quality enters the model as a rank, $q \in [0.01, 1]$, which is the percentile $Q_p \in [1, 100]$ of Section 3 rescaled to the unit interval. I perform the same transformation on scope $s \in [0.01, 1]$. When the application is assigned, the examiner's leniency l is drawn from a kernel-density estimate $\hat{F}_l$ of the leave-one-out grant rate of Section 5, clipped to the grid support $[0.01, 0.99]$, so l is a probability-scale object. Neither the applicant nor the examiner directly observes q. The examiner learns about it from a noisy signal in each round. The applicant holds a subjective Gaussian density over their own quality, centered at q up to a bias q_{err} and with dispersion q_{unc}.[26] The applicant knows their cost type ζ, their filing taste u_s and the distribution $\hat{F}_l$, but observes neither the examiner's leniency nor the examiner's signals.

Only the applicant makes choices. At filing they choose the scope s_0 to claim, which is their requested scope. In each round r the examiner receives a signal, updates their assessment, and grants if the acceptance index clears zero. A grant ends prosecution with a patent of the offered scope $\hat{s}_r \leq s_r$. After a rejection the applicant updates a joint belief over (q, l) and either abandons, for an outside value normalized to zero, or pays the cost of another round and files a claim $s_{r+1} \leq s_r$, no broader than the claim just rejected. Prosecution ends at a grant, at abandonment or after round $R - 1$, whichever comes first. Claiming scope s in a round costs κs^{χ}, a grant pays νs, and continuing after a rejection costs $C(s, r)$. The examiner makes no choice and has no payoff. The grant rule and the offer, set out next, are taken as given.

7.2 The examiner's grant rule

The examiner has no objective function and does not optimize. Both the grant rule and the scope offer that follow are a grant technology, taken as given by the applicant and disciplined by the moments of Section 7.5.

[26] This applicant-side prior is the counterpart of the prior in Section 5.1, and is distinct from the examiner's prior $\mathcal{N}(\bar{q}, \sigma_q^2)$ of Section 7.2.

Learning. The examiner holds a conjugate Gaussian prior $q \sim \mathcal{N}(\bar{q}, \sigma_q^2)$. Through round r, they have observed $r+1$ signals $\xi_j = q + \eta_j$, $\eta_j \sim \mathcal{N}(0, \sigma_\eta^2)$, i.i.d. across rounds. The examiner's posterior mean is a weighted average of the signal mean $\bar{\xi}_r = (r+1)^{-1} \sum_{j \le r} \xi_j$ and the prior, $\mathbb{E}_r[q] = f_r\, \bar{\xi}_r + (1 - f_r)\, \bar{q}$, with precision weight

$$f_r \;=\; \frac{(r+1)/\sigma_\eta^2}{1/\sigma_q^2 + (r+1)/\sigma_\eta^2}, \qquad \mathbb{E}\Big[\mathbb{E}_r[q] \,\Big|\, q\Big] \;=\; f_r\, q \;+\; (1 - f_r)\, \bar{q}. \tag{15}$$

The second expression is the posterior mean an examiner holds on average given true quality q.

The precision weight f_r is increasing in r, so later rounds include a better-informed examiner.

Grant rule. The examiner grants in round r if the acceptance index clears zero,

$$\begin{aligned} &\underbrace{\alpha_q\, \mathbb{E}_r[q]}_{\text{quality premium}} \;-\; \underbrace{b_q\,(q - \tfrac{1}{2})}_{\text{quality barrier}} \;-\; \underbrace{b_{0q}\,(q - \tfrac{1}{2})\,\mathbf{1}[r=0]}_{\text{first-round quality barrier}} \;-\; \underbrace{\Pi(s_r)}_{\text{scope penalty}} \\ &+\; \underbrace{(l - \mu)}_{\text{leniency}} \;-\; \underbrace{b_r\, r}_{\text{round drift}} \;-\; \underbrace{b_0\, \mathbf{1}[r=0]}_{\text{first-round barrier}} \;+ \varepsilon_r \;\ge\; 0, \qquad \varepsilon_r \sim \mathcal{N}\Big(0,\; \sigma^2_{\text{grant}}\Big), \end{aligned} \tag{16}$$

with $\Pi(s) = \alpha_1\, s + \alpha_2\, s^2 + \alpha_3\, s^3$ the scope penalty of Section 5.1, now given its estimated cubic form. Quality enters in two ways. Through the examiner's realized posterior mean $\mathbb{E}_r[q]$ quality benefits the application as the examiner learns about the true underlying quality of the invention. Through $b_q\,(q - \frac{1}{2})$ true quality enters directly in every round, and b_{0q} lets its weight in the first round differ from later rounds. Positive values of b_q and b_{0q} make a grant harder for higher-quality applications. I allow a separate first-round term because, relative to each round's allowance rate, the allowance rate falls with quality far more in the first round, by 13% against 4% in the second, as Appendix Figure M.4 shows. ε_r is an idiosyncratic case-specific shock, i.i.d. across rounds and independent of the signals. Here, μ is the acceptance threshold, and $l - \mu$ makes explicit that it varies with the specific examiner's leniency. b_r captures a drift against long-running applications, and b_0 is a reluctance term against allowing in the first round. Together they govern how the grant rates change across revision rounds. The applicant never observes the examiner's signals, so the model is solved and simulated with the grant rule averaged over them, as Appendix M sets out.

Granted scope. Conditional on a grant, the granted scope is the filed claim scaled by the examiner's *offer ratio* $\pi(l, s, q)$,

$$\hat{s}_r \;=\; \min\Big\{\, s_r,\; s_r \cdot \pi(l, s_r, q) \cdot \gamma_{\text{E}}^{\,r} \,\Big\}, \qquad \pi(l, s, q) \;=\; \gamma_s + \gamma_l \left(\max\{l, \underline{l}\} - \tfrac{1}{2}\right) + \gamma_{s\text{-sl}}\,(s - \tfrac{1}{2}) + \gamma_q\,(q - \tfrac{1}{2}), \tag{17}$$

$\gamma_E \leq 1$ a per-round decay of the offer. The leniency term is floored at $\underline{l} = \frac{1}{2}$. An examiner stricter than the median offers the median examiner's ratio, and only more lenient examiners offer more. The floor follows the data's offer ratio, which is flat in leniency below the median examiner and rises above it. γ_s is the default offer ratio, evaluated at the grid midpoint, and the three gradients are symmetric deviations from it: γ_l in leniency, $\gamma_{s\text{-sl}}$ in filed scope, which separates the slope of granted on filed scope from the level set by γ_s, and γ_q in quality. The offer is conditional on true quality q rather than on the examiner's posterior, so γ_q is a reduced-form term, discussed in Section 7.5.

7.3 The applicant's problem

The applicant solves a finite-horizon optimal-stopping problem. In each round, the applicant chooses how broad their filed scope should be, or whether to abandon the application. They do not observe q or l, so every term is taken in expectation against the round-r joint belief w_r over (q, l).

$$\bar{P}_r(s) \;=\; \int P_r(q, s, l)\, \mathrm{d}w_r(q, l) \tag{18}$$

is the belief-averaged grant probability implied by Equation (16). The value of entering round r with scope s is

$$V_r(s) \;=\; \omega(q)\Big[\bar{P}_r(s)\,\nu\, s \;-\; \kappa\, s^{\chi}\Big] \;+\; \Big(1 - \bar{P}_r(s)\Big) \cdot \underbrace{\max\Big\{\,0,\;\; \beta \max_{s' \leq s} V_{r+1}(s') \;-\; C(s, r)\,\Big\}}_{\text{continue or abandon}}, \tag{19}$$

where νs is the value of a grant of scope s for an invention of median quality, κs^{χ} the cost of claiming scope with curvature χ, and $C(s, r) = \left(c_0 + c_s\, s + c_r\, r^{0.85}\right) e^{\zeta}$ the revision costs in round r, scaled by the applicant's cost type e^{ζ} with $\zeta \sim \mathcal{N}(0, \sigma_c^2)$. The factor $\omega(q) = e^{\,c_q(q - \frac{1}{2}) + c_{q^2} \max(q - \frac{1}{2}, 0)^2}$, with $\omega(\frac{1}{2}) = 1$, scales the stakes of a claim with the quality of the invention, through c_q and, in the top half of quality, through c_{q^2}. Both the value of a grant and the cost of claiming scope rise by this factor, while the revision costs do not. Because the applicant's choices depend on the stakes only relative to the revision costs, this is the same as revision costs that fall with quality by the factor $\omega(q)$. An applicant with a better invention faces the same grant rule but has more at stake relative to each round that an application takes. An applicant with a higher-quality invention therefore responds to more rejections instead of abandoning. A broader filed scope takes more rounds, so the higher stakes also let higher-quality applicants request more scope. The counterfactuals of Section 7.7 call these two effects persistence. The grant payoff is computed on filed scope, not on granted scope $\hat{s}$ of Equation (17), so the applicant does not anticipate the examiner's offer.

Appendix M.1 discusses this restriction. Unlike the static model of Section 5.1, where ν values granted scope, ν here is a value per unit of filed scope.

The stakes of a claim are multiplicatively separable in quality and scope. A granted unit of filed scope is worth $\nu\,\omega(q)$, which is the value of scope conditional on quality. Because $\omega(q)$ also scales the cost of claiming scope, quality leaves the trade-off between the value of a grant and that cost unchanged. Dividing Equation (19) by e^{ζ} shows that, apart from the chance of a grant, applicants differ only in their stakes relative to the cost of a round, $\omega(q)\,e^{-\zeta}$. This index combines a part that measured quality explains, $\omega(q)$, with a log-normal part that it does not, the cost type. Quality thus moves the applicant's optimal scope only through the chance of a grant and through persistence. Matcham and Schankerman (2026) also value a claim with a Cobb-Douglas form, the applicant's padding times a log-normal claim value, which they motivate by the complementarity of the two. Their attorney cost of padding does not scale with the claim value, so it restrains the padding of a more valuable claim less. Here, the cost of claiming scope scales with quality, and the only cost that does not is the cost of a round. Measured quality therefore takes over part of the role of their unobserved claim value, and the log-normal cost type plays the role of the rest.

An applicant abandons the application if the discounted continuation value does not cover the revision cost. Therefore, the model generates endogenous abandonment and a distribution of stopping rounds. The outer maximization, $\max\{0, \cdot\}$, captures the abandonment decision for an outside option normalized to 0. The inner maximization over $s' \leq s$ is a one-way ratchet with respect to scope. A claim may be narrowed after a rejection but is never broadened again, so the filing decision in each round caps the next round. Broad filing therefore is the option-preserving choice for the applicant.[27] The continuation is evaluated at the next round's belief w_{r+1}. After a rejection, the applicant updates w_r by Bayes' rule. Each rejection is bad news about both leniency and quality, so revision decisions respond to accumulated rejections.

Filed scope. s_r^* is the optimal scope that solves the applicant's problem (19) in round r, ignoring the cap. Filed scope departs from it by a displacement Δ_r and cannot exceed the previous round's claim,

$$(20)\quad s_r \;=\; \min\Big\{\, s_r^* + \Delta_r,\; s_{r-1} \Big\}, \quad s_{-1} \equiv 1, \qquad \Delta_r \;=\; \Big[\, \sigma_s\, u_s \;+\; \beta_s\,(l - \tfrac{1}{2}) \;+\; \beta_q\,(q - \tfrac{1}{2}) \,\Big]\,\gamma_{\mathrm{A}}^{\,r},$$

where $s_0 = s_0^* + \Delta_0$ is the original filing, u_s is a standard normal filing taste shock drawn once per application and scaled by σ_s, β_q and β_s are the quality and leniency terms of the displacement, and $\gamma_{\mathrm{A}} \in [0, 1]$ is the round-by-round concession rate at which the displacement decays back toward the optimum as prosecution proceeds. Filed scope is rounded to the scope grid and kept

[27] The solution updates the applicant's belief along one path, the sequence of optimal claims from the round-0 optimum, rather than along a separate path for every claim cap.

within $[0.01, 1]$. The applicant observes neither quality nor leniency directly before filing, so β_q and β_s are reduced-form terms, discussed in Section 7.5. The cap $s_r \leq s_{r-1}$ is the scope ratchet of Equation (19). The applicant's concession rate γ_{A}, the ratchet and the examiner's offer decay γ_{E} together govern the scope narrowing across rounds.

7.4 Optimal policy and solution

The examiner's posterior evolves through Equation (15) from private signals that depend on the true quality but do not depend on the applicant's scope choice. Therefore, the applicant can forecast the examiner's posterior without reference to their own strategy, and I solve the applicant's finite-horizon decision problem against that technology as a single-agent dynamic program.

Proposition 2 (Existence of an optimal policy)**.** *Fix the grant technology* (16)–(17) *and the belief dynamics above. Then for every round r and belief state w_r the applicant's problem* (19) *admits a maximizer, and the value function V_r is bounded and continuous in w_r. The resulting policy maps the applicant's type (q, ζ) into a filed-scope and stopping rule which, together with the leniency l of the assigned examiner, determines (a) the grant probability, (b) the number of revision rounds, and (c) the granted scope $\hat{s}$ conditional on a grant. Beliefs are updated by Bayes' rule on the path.*

Proof in Appendix C.[28]

7.5 Estimation

Indirect inference. The model has no closed-form likelihood, so I estimate the vector of the model's free parameters, θ, by indirect inference (Gouriéroux, Monfort, and Renault, 1993; Smith, 1993) on auxiliary statistics computed identically in the data and in the simulation. These are the grant rate, average filed scope, granted scope, and revision rounds within bins of quality, filed scope, or leniency, the variances of selected outcomes within the same bins, the allowance rate by quality in the first two rounds, and four profiles across revision rounds: the share of applications granted in each round, the abandonment hazard, filed scope, and the slope of filed scope in leniency. Appendix Table M.7 lists the moment sets. Appendix M.2 details the weighting, identification, and inference of the estimator. On the quality and round axes, these are the descriptive objects of Section 4. The complier shares are deliberately excluded from the estimation and instead used as untargeted moments. Stacking the data statistics in m^{data} and their simulated counterparts,

[28] Existence does not require quasiconcavity. The objective in (19) is continuous on a compact action set, and, in the solved model, on a finite grid. $\partial V/\partial s$ is not monotone, since the value of a grant and the scope conceded both increase in s while the probability of obtaining one falls, so the balance can reverse across the domain. The local optima this creates are resolved by the grid search, and the solved policy is pure.

computed from N_{sim} applications drawn with the empirical quality and leniency distributions, in $m^{\text{sim}}(\theta)$, the estimator is

$$\hat{\theta} \;=\; \arg\min_{\theta\in\Theta} \Big[m^{\text{data}} - m^{\text{sim}}(\theta)\Big]' W \Big[m^{\text{data}} - m^{\text{sim}}(\theta)\Big], \tag{21}$$

over the bounds Θ. Appendix Table M.1 describes all parameters. The estimator requires that the map $\theta \mapsto m^{\text{sim}}(\theta)$ be locally injective at $\hat{\theta}$, which the sensitivity analysis of Appendix M.2 checks. W is diagonal rather than efficient. Each moment set is divided by the standard deviation of its data values across bins, so the criterion targets how an outcome changes across bins rather than its precision in any one bin. Additionally, a small set of weights across moment sets keeps sets with little variation but an informative slope, such as the grant rate by leniency, from being overwhelmed by other moment sets.

Identification. Equation (21) determines θ jointly, but the parameters fall into three blocks. Each of these three blocks is disciplined by a different source of variation in the data.

The first block is the examiner's grant rule. Examiner leniency is assigned as good as randomly within art unit and year, so the grant rate across leniency traces the distribution of the acceptance index directly. Among applications of the same quality and filed scope, the grant probability is therefore a probit in leniency, $\Phi\Big((a(q,s)+l)/\sigma_{\text{grant}}\Big)$, with slope coefficient $1/\sigma_{\text{grant}}$ and constant $a(q,s)/\sigma_{\text{grant}}$. The slope identifies σ_{grant}, and with it the constant gives the acceptance index $a(q,s)$ in its own units. How $a(q,s)$ changes with filed scope at given quality identifies the scope penalty Π. Since $\Pi(0)=0$, what remains of $a(q,s)$ once scope and the quality terms below are accounted for is the threshold μ. The grant hazard across rounds disciplines b_r and b_0. The quality terms of the grant rule – b_{0q}, b_q, and α_q – are identified by the allowance rate across quality in the first two rounds together with the eventual grant rate across quality. At the calibrated signal noise the three enter differently: b_{0q} shifts only the first round, b_q every round, and the weight of α_q grows from round to round with f_r.

The second block is the applicant's payoff. It contains the cost of a revision round, $c_0 + c_s s + c_r r^{0.85}$, its dispersion across applicants, σ_c, the quality factor of the stakes, c_q and c_{q^2}, the value of a unit of scope at the median quality, ν, the cost of claiming scope, κ, and the applicant's uncertainty about the quality of their own invention, q_{unc}. The applicant's problem is homogeneous of degree one in ν, κ, c_0, c_s and c_r, since scaling all five leaves every choice unchanged, so I measure them relative to ν, and only these ratios enter the counterfactuals. Locally, most of the information on this block comes from the requests applicants file. A broader claim takes more rounds, so the cost of a round and its rise with scope, c_s, move how broadly applicants request, and the cost of claiming scope, κ, sets the level of the request directly. How the stakes rise with quality, c_q and c_{q^2}, moves how the request varies with quality, and q_{unc} governs how closely the request tracks the quality

of the invention. Two profiles carry most of the information on c_q: filed scope by quality among granted applications and revision rounds by quality. The quality term of the filing displacement, β_q, also moves how filed scope varies with quality, and that slope informs mostly β_q. The slopes that move c_q most are those of rounds by quality and of the allowance rate by quality in the second round, and they move it in opposite directions. The grant rate by quality adds to c_{q^2}. Table M.5 reports how much the standard error of c_q rises when each set is dropped and how far each slope moves it. Granted scope by quality follows the requested scope and adds little on its own. The cost intercept c_0 draws mainly on how the request varies with leniency across termination rounds and on the grant rate by filed scope.

The third block covers how the requested scope departs from the applicant's optimum via the displacement term Δ_r and how the examiner's offer $\pi(l, s, q)$ turns the filed scope into granted scope. The dispersion of the displacement, the filing taste σ_s, is identified by the variance of the requested scope within a quality bin. The quality term β_q is identified by how the requested scope varies with quality. Applicants do not know the examiner their application will be assigned to when they file, so the requested scope is affected by leniency only through the leniency term β_s, which the request by leniency therefore pins. Both β_q and the cost terms c_q and c_{q^2} capture how the requested scope varies with quality. The cost terms addiitonally moves with the number of rounds and the grant rate by quality, which separately identifies the terms. In the examiner's offer, the base ratio γ_s captures the level of granted scope relative to the filed scope and $\gamma_{s\text{-sl}}$ captures how that ratio changes with filed scope. The decay γ_{E} captures how granted scope falls with leniency, since stricter examiners take more rounds before they grant. The data contain the initially filed scope for every application but not the narrowed scope after each rejection. The narrowing of claims across rounds is never observed directly and the concession rate γ_{A} is mainly based on the requested scope by quality.

The three blocks describe where the identifying variation for each parameter comes from. They are not a proof of identification, which a model of this size does not admit in closed form. Locally, identification requires that combinations of parameters are pinned. No combination can change without moving some moment, so that the Jacobian of the moments with respect to the parameters has full rank. With ν held fixed as the numeraire of the payoff scale, the Jacobian has full rank in the other 26 parameters. However, some combinations are pinned far less than others, as Appendix M.2 shows.

Reduced-form components. Three parameters fit slopes in the data without a choice or a belief behind them: the offer's quality gradient γ_q as well as the quality and leniency terms β_q and β_s of the filing displacement. The offer is conditional on true quality, so γ_q matches the slope of granted scope across quality without explaining how that slope arises during prosecution. An alternative

structure would be an offer condtional on the examiner's posterior, which sharpens with every round. Since the data record only initially filed scope and granted scope, but no narrowing in intermediate rounds, the two forms differ only in a single testable implication: how the quality slope of the ratio of granted to filed scope varies with the round of grant. The observed round profile rejects the posterior-based offer at every level of signal noise and offer decay. I discuss this in more detail in Appendix M.3. Both the reduced-form offer condition and the offer conditioned on the examiner's posterior are pinned down only by the slope pooled over rounds, which γ_q matches.[29] γ_q enters neither the grant rule nor the applicant's problem. Therefore, no counterfactuals of grants, the requested scope or the number of rounds are affected by it. Only counterfactual granted scope can be affected by it.

The two filing terms average over a distinction in the data that the model does not contain. Section 4 groups applications by whether the filed scope of the claims are the applicant's first request or whether they depend on a previous application. Among free-standing applications, filed scope varies little with invention quality, while continuations and national-stage entries file more narrowly the higher their quality, as Appendix Figure D.2 shows. The model has a single filing regime, which features an applicant who chooses scope from a continuous set. Therefore, there is no parent application or earlier search report on the claims made in the filed scope. The moment behind β_q is the profile of filed scope across quality among granted applications, pooled over the three groups and net of differences in level between filing routes. The displacement of Equation (20) therefore stands for the combined filing behavior of the three groups and describes none of them directly.[30]

Inference. Under standard indirect-inference regularity conditions, $\hat{\theta}$ is asymptotically normal around the minimizer of the population criterion in Equation (21). In Appendix Table M.3, I show that the over-identification test rejects. Therefore, I interpret the minimizer as the best-fitting member of the parametric family under the stated W, not a true parameter the data were generated from. The estimand is therefore defined relative to W, so the across-profile weights are part of it.

I estimate the variance of the empirical auxiliary statistics by a block bootstrap that resamples examiners, the level at which leniency varies. I combine the variance with the binding-function Jacobian in a sandwich estimator, which stays valid when the model is an approximation and the

[29] The estimate of γ_q is small. It captures the mild decline of residualized granted scope across quality, which is a targeted moment. That decline is an order of magnitude too small to be visible in the unconditional profile of Section 4.

[30] Solving for an optimum and estimating a parametric departure from it is the structure of Hortaçsu, Luco, Puller, and Zhu (2019), who embed a cognitive-hierarchy model in an auction model to price firms' deviations from Nash bidding. There, the departure is an attribute of the bidder. Here it is an attribute of the application, since absorbing applicant fixed effects strengthens the gradient rather than removing it.

weighting is not efficient The value of a unit of scope ν is held fixed as the numeraire of the payoff scale. Of the other 26 parameters, 25 lie more than two standard errors from zero. The exception is the offer's slope in filed scope $\gamma_{s\text{-sl}}$, which is close to its lower bound of zero, as Appendix M.2 discusses.

Calibrated constants. The horizon, the discount factor and eight further quantities are fixed rather than estimated. The horizon and three belief quantities follow from the model's construction. The horizon $R = 12$ lies well above the observed number of rounds, and Appendix M.1 shows that it does not bind. The examiner's prior moments, $\bar{q} = \frac{1}{2}$ and $\sigma_q = 1/\sqrt{12}$, are the mean and standard deviation of a uniform distribution on the unit interval, so the examiner's prior matches the support of the quality rank by construction. The belief parameter q_{err} is set to zero because it is collinear with q_{unc}.

The two constants of the offer, the leniency gradient γ_l and its floor $\underline{l}$, are pinned by the round-zero granted-to-filed ratio,[31] and Appendix M.4 varies both.

The signal noise σ_η is set to 0.5, the value with the lowest criterion in the profile of Appendix M.1. At this value the precision weight is $f_r = (r+1)/(r+4)$, a quarter at filing, one half in the third round and two thirds by the sixth,[32] so the examiner's posterior moves with true quality within the rounds the data cover. The exponent of the round cost is fixed at 0.85.

The discount factor $\beta = 0.95$ applies per round. Successive office actions in the estimation sample are a median of 5.9 months apart, so β corresponds to an annual discount factor of $0.95^{12/5.9} \approx 0.90$.[33] The scope-cost curvature χ enters only through the cost κs^χ over the scope range prosecution realizes, where a change in curvature is nearly a change in the level κ.

7.6 Estimates and fit

Appendix M.3 reports the fit of all sixteen targeted moment sets, eight of them in Figures M.2, M.3 and M.4. The model reproduces the level moments the offer mechanism targets, as well as the sign of the two filed-scope profiles in the data. Two misfits stand out. The larger by far is the abandonment hazard by round. The model abandons too few applications early in prosecution and too many late: at $r = 4$ it abandons just under half of the applications still pending, against under three in ten in the data. This moment set accounts for three quarters of the over-identification

[31] No auxiliary statistic in Equation (21) is satisfied by calibration alone: the round-zero ratio is not itself a moment, and granted scope by leniency remains a weighted target the estimates must fit.

[32] With $\sigma_q = 1/\sqrt{12}$, $f_r = 4(r+1)/\big(12 + 4(r+1)\big) = (r+1)/(r+4)$.

[33] The gap is stable between 5.3 and 6.4 months across the calendar years and between 5.6 and 6.8 across technology centers, over 5.5 million applications. Measuring instead the span from the first office action to disposal divided by the number of rounds, which adds allowance processing, the gap is 10.3 months and the implied annual factor 0.94. The queue before the first office action is excluded throughout, as it precedes the negotiation the model discounts over.

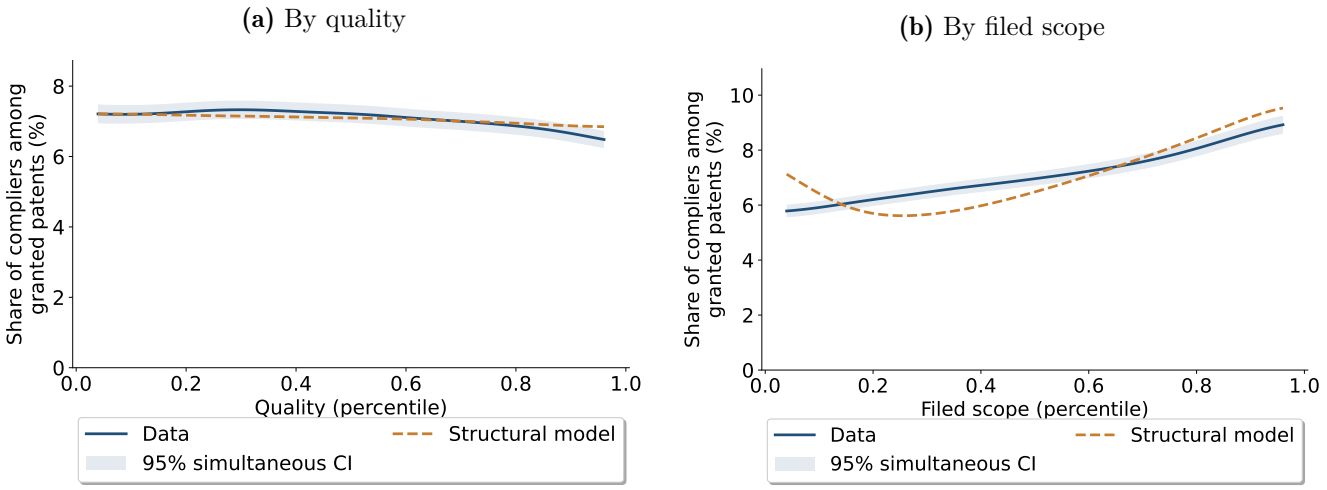
statistic. The misfit reaches the results of Section 7.7 unequally. Under the stricter office, applicants who stop sooner remove 1.8% of baseline grants and narrower claims that are now allowed add 0.9% back, so the applicants' part of the grant loss, 0.9 of 4.0%, rests on the exit decision the model fits worst. The narrowing of granted scope works through what applicants request, and exits account for 3% of it. The second is filed scope by the number of rounds an application takes, which rises in both series but almost twice as steeply in the model, and accounts for about a sixth of the statistic. In quality, the model matches the small rise of the grant rate and the fall of the allowance rate in the first round, but it understates the fall in the second round (1.1 against 1.9 percentage points). Both quality barriers are estimated to be positive and, in the first six rounds, outweigh the weight the posterior mean puts on quality. At a given filed scope, better inventions are therefore not granted faster in those rounds, and the difference fades as the examiner learns. The rise of the eventual grant rate runs through persistence: with $\omega(q)$ replaced by its harmonic mean over applicants, the grant rate would fall with quality instead of rising, and requests would fall with quality instead of staying flat. At the estimates, $\omega(q)$ is about 15% higher at the 90th percentile of quality than at the 10th. Relative to the cost of a round, a grant is therefore worth about 15% more to the better of the two inventions.

The fit above is about the targeted moments. Three moments the objective never saw serve as checks. The conditional complier shares carry the most weight and are taken up below. The secon is the scope concession by quality, which is reported in Table M.9. The third is the joint grant-rate surface in quality and leniency, depicted in Appendix Figure M.5. The surface tests the absence of an interaction between quality and leniency, not the two marginal profiles, which enter the estimation. The grant rule carries no quality-by-leniency term, so if the quality profile tilted across leniency quintiles in the data, the model would have no parameter to follow it. The levels line up in all five quintiles, but the data's quality profile flattens toward the most lenient quintile more than the model's, so the separability that Proposition 1 requires holds in levels, as Appendix M.3 reports.

The share of granted patents whose grant depends on the examiner assigned to them is not among the estimation targets, though not literally an out-of-sample check. The targeted grant-rate profiles constrain it through the threshold structure, so the check is informative about joint shape rather than independent of the fit. Figure 5 plots the model's share of compliers among granted patents against the estimates of Section 6, along quality in Panel a as well as along filed scope in Panel b. In the model the share of compliers among grated patents is exact rather than estimated: every simulated application is paired with the reference examiner, and the compliers are the granted applications that examiner would refuse.

In quality, both the estimated and the simulated shares fall. The estimated share falls from 7.2% among the lowest-quality applications to 6.7% among the highest, which is a 7.6% decline.

Figure 5: *Complier share: model versus data*



Notes: Solid line, the share of compliers among granted patents estimated in Section 5.4, with its 95% simultaneous band. Dashed line, the same share in the model at the estimates. In the model the share is computed exactly: every simulated application is paired with the same application under the reference examiner, the examiner whose grant rate equals the average, and the share is that of the model's granted applications that the reference examiner would refuse. Both curves are smoothed with the same Gaussian kernel ($h = 0.07$) over the same percentile grid. The model curve is expressed in the data's units by one factor for both panels. Panel (a) conditions on quality, Panel (b) on the within-sample percentile of filed scope. The complier share is not a targeted moment.

The decline in the simulated share is 4.6%. The simulated curve stays inside the estimate's simultaneous band except above the 92nd percentile of quality, where the estimate falls more steeply. In filed scope, broader filings are more likely to owe their grant to the examiner they were assigned to in both the model and the data. Between the 10th and the 90th percentile of filed scope the share rises by 43% in the model and by 46% in the data. However, the shapes differ: the estimated share rises throughout. The simulated share starts higher among the narrowest filings, falls to a minimum near the 26th percentile, and from the median on rises more steeply than the estimate. The simulated curve lies outside the estimate's simultaneous band below the 56th percentile, except where the two cross near the 14th, and again between the 75th and the 84th percentiles.

7.7 Counterfactuals

The main counterfactual I report is a stricter patent office. Reform reports call for stricter examination so that fewer patents are granted on obvious inventions or with overly broad claims (U.S. Federal Trade Commission, 2003; Executive Office of the President, 2013; Productivity Commission, 2016). Every examiner's leniency falls by the gap between today's median and 45th-percentile examiner. The median examiner becomes as strict as today's 45th percentile and the dispersion of

leniency is unchanged. In the grant rule of Equation (16), this is the same as raising the acceptance threshold μ by that gap. I also run a larger shift, which uses the gap to the 25th percentile. Applicants know about this change, so their beliefs adjust to the stricter grant rule. I hold the set of filed inventions fixed, because the data contain only filed applications and nothing in them disciplines how the decision to file a patent application would respond to a stricter office. I present every result by decomposing it into two parts. First, the rule columns apply only the stricter examiners, keeping applicants' requests and their decisions to continue or abandon as they are today. Second, the response columns are the rest of the total change, which comes from applicants choosing their requests and their persistence anew. Changes by quality and by filed scope are measured within fixed groups of today's applications, so they are not confounded by the change in filed scope.

The counterfactuals report behaviour only: what applicants request, how long they persist, how many applications are granted, how broad the granted patents are, and which inventions receive them. I attach no dollar value or welfare measure to these changes. Scope is a within-year percentile rank, so changes in mean filed and granted scope are changes in a mean rank, while changes in grants and rounds do not depend on this scale.

I present the counterfactual results in Table 1. Under the stricter office, granted patents fall by 4.0%, and by 25.2% under the larger shift. The rule alone accounts for 3.2 of the 4.0% and the applicants' response for the rest. The response consists of applicants who stop sooner, which removes 1.8% of today's grants, slightly counteracted by narrower requests that examiners now grant.

Requests, the mean filed scope s_0, narrow by 3.7% and by 19.5% under the larger shift. Almost all of this is due to the applicants' response. The rule change alone moves filed scope by at most 0.6%, because filed scope depends on the assigned examiner only through the leniency term β_s of the filing displacement. Granted patents become 4.3% narrower. The rule change alone accounts for 13% of this narrowing and the response for the other 87%. The applicants' responses can be decomposed further: 77 percentage points are due to changes in applicants request, 7 from changed narrowing behavior after a rejection, and 3 from applicants who abandon sooner. Across the offer calibrations of Appendix M.4, the narrowing ranges from 4.2% to 4.4% and the rule's share from 10.3% to 15.4%. Patents therefore become narrower mostly because applicants ask for less, not because examiners refuse more claims of the same filed scope. An evaluation that holds today's applications fixed sees only the rule part, which would account for most of the decrease in granted patents but little of the scope narrowing.

Over the distribution of filed scope, the stricter office takes most from both ends. Table M.13 reports the losses, split into groups at the data's quintiles of filed scope. The narrowest group loses 5.0% of its patents, the second and third 3.4%, the fourth 4.1% and the broadest 5.2%. The two ends lose for different reasons: under the rule change alone, the broadest group loses the most,

Table 1: *The office's levers: requests, rounds, grants and granted scope*

Status quo	Requested scope 0.556		Rounds 1.336		Grants 0.710		Granted scope 0.426	
Change against the status quo (%)	Total	Rule	Total	Rule	Total	Rule	Total	Rule
Stricter examiners								
Median at today's 45th percentile	−3.7	+0.1	−0.4	+3.1	−4.0	−3.2	−4.3	−0.6
Median at today's 25th percentile	−19.5	+0.6	−6.2	+17.3	−25.2	−19.5	−22.1	−3.2
Consistent examiners								
Every examiner at the median	+6.5	−0.1	+11.9	−3.4	+15.8	+7.2	+8.0	−0.3
Every examiner at the mean	+1.7	0.0	+12.4	+0.5	+10.8	+3.3	+2.3	−1.2
Judge-level compression	+1.6	0.0	+11.4	+0.5	+9.8	+3.0	+2.0	−1.1
Limit on rounds								
At most 3 rounds	−0.5	0.0	−10.8	−10.7	−6.3	−6.4	−1.0	−0.5
At most 4 rounds	−0.2	0.0	−2.9	−2.8	−1.5	−1.5	−0.2	0.0
Fee per round ($\times \bar{c}_1$)								
0.25	−0.8	0.0	−5.5	0.0	−4.3	0.0	+1.2	0.0
0.5	−1.4	0.0	−9.9	0.0	−7.4	0.0	+1.3	0.0
1	−2.6	0.0	−19.4	0.0	−14.5	0.0	+1.9	0.0

Notes: Percentage change against the status quo, for each lever. "Total" re-solves the applicants' policy under the lever, so their beliefs are consistent with it. "Rule" keeps the status-quo policy and applies only the office's new rule, and the difference between the two is the applicants' response. Requested scope is the mean filed scope, rounds the mean number of revision rounds, grants the number of grants, granted scope the mean granted scope among grants. The status-quo row gives the levels, with the grant rate for grants. Scope is a within-year percentile rank, so the changes in requested and granted scope are percentage changes of a mean rank. A fee leaves the office's rule unchanged, so its rule column is 0 by construction and its total is entirely the response. The fee is a multiple of the mean first-round revision cost $\bar{c}_1$. Filed inventions are held fixed, so the change in grants is also the change in the grant rate, and every difference is paired on one simulated population. Levers are reported at their own sizes and are not ranked.

while the response falls most on the narrowest group. Under the larger acceptance rate shift the two ends lose the same share, about 31%, against 22% in the middle groups.

I present additional counterfactual scenarios with the same split between the rule and the response. Collapsing the examiner leniency distribution, such that every examiner applied the median examiner's standard leads to grants increasing by 15.8%. They would increase by 10.8% if every examiner applied the mean standard, and by 9.8% if every examiner's distance from the mean leniency shrank to about a quarter, which corresponds to the spread found among bail judges and prosecutors (Dobbie et al., 2018; Agan et al., 2023). The response accounts for about 55% of the increase in the first case and about 70% in the other two. Requests and granted scope increase only through the response. Fees and caps act on the persistence of applicants, because they change the value of another round. A fee per round of a quarter, a half or all of the mean first-round revision cost decreases grants by 4.3%, 7.4% and 14.5%. The number of mean rounds decreases by 5.5%, 9.9% and 19.4%, while requests move by at most 2.6%. The 2013 increase in the

fee for a request for continued examination and the 2025 surcharge on continuations are adopted counterparts of the fee arms. A cap of three rounds, the highest welfare-improving counterfactual in Matcham and Schankerman (2026), lowers grants by 6.3%, almost all through the rule. Table 1 reports every lever at its own size, and the levers are not ranked.

An office that tests quality. Over the distribution of quality, the stricter office removes about 4% of the grants for every quintile of invention quality. I present the results in Table 2. To test how a different rule set would impact allocation of patents across invention quality, I build an office that tests the quality of the invention instead of the scope of the request. In this counterfactual, filed scope does not enter the acceptance index of Equation (16). Instead, quality enters with a weight chosen so that it varies the index as much as the scope penalty does at today's filings, averaged over the rounds in which today's patents are granted. The examiners of this alternative office see quality with the same noise of 0.5 as the existing office, but it applies no quality barrier. The acceptance threshold keeps the average first-round acceptance index at today's filings unchanged, and applicants respond to the new rule.

Before any change in examiner strictness, the quality-testing office grants 73.4% of applications against 71.0% in the status-quo. The alternative office grants the highest quintile of quality far more often than the lowest, 88% against 57%. Both offices then become stricter by the same shift in leniency, with the median examiner moving to today's 45th percentile, each based on its own specific grant rate. In the office that tests quality, the shift takes away 6.9% of the granted patents of the lowest quality quintile and 2.4% of those of the highest, which corresponds to 1.60 and 0.55 times its overall fall of 4.3%. Under the larger shift, to today's 25th percentile, today's office still takes away about a quarter of the grants in every quintile, while the office that tests quality takes away 40% in the lowest quintile and 18% in the highest. A stricter office therefore changes which inventions receive patents far less than an office that tests quality does when it becomes equally strict. Without the gradient of the stakes in quality, the stricter office would take slightly more of its grants from higher-quality inventions. With $\omega(q)$ replaced by its harmonic mean over applicants, as in Section 7.6, the highest quintile would lose 4.15% of its grants and the lowest 4.04%, against 4.01% and 4.07% today.

These results carry three caveats from the estimated model. First, net of quality and leniency, the model's grant rate rises with filed scope up to about the 36th percentile, while the data's rises only up to the 17th, and among the narrowest filings the model grants less often, 0.68 against 0.73 at the 5th percentile. The larger loss of the narrowest filings under a stricter office sits on this rising part, which the data show over a much narrower range, so it is the less firmly grounded end of the profile. The larger loss of the broadest filings sits on the falling part, which the data share. Second, among the broadest fifth of filings, granted scope rises less steeply with filed scope

Table 2: *Grant losses by invention quality under a stricter office and under an office that tests quality*

	Today's office			Office that tests quality		
	Grant rate	Change in grants (%)		Grant rate	Change in grants (%)	
	before	45th	25th	before	45th	25th
Lowest fifth of quality	0.704	−4.1	−25.3	0.574	−6.9	−39.6
Second fifth	0.707	−4.1	−25.3	0.663	−5.7	−34.6
Third fifth	0.709	−4.0	−25.2	0.745	−4.4	−29.1
Fourth fifth	0.713	−4.1	−25.1	0.819	−3.3	−23.4
Highest fifth	0.720	−4.0	−24.8	0.879	−2.4	−17.7
All applications	0.710	−4.0	−25.2	0.734	−4.3	−27.8

Notes: Change in the number of grants against each office's grants before the shift, in percent, by fifth of invention quality. Both offices become stricter by the same downward shift of every examiner's leniency, which moves the median examiner to today's 45th percentile (the main shift) or 25th percentile, each from its own grant rate before the shift. Applicants re-solve their requests and their persistence under the stricter rule. In the office that tests quality, the scope penalty is switched off, and quality enters the acceptance index with a weight chosen so that, averaged over the rounds in which today's patents are granted, it varies the index as much as the scope penalty does at today's filings. Its examiners see quality with today's signal noise, $\sigma_\eta = 0.5$, and it applies no quality barriers. Its threshold keeps the mean first-round acceptance index at today's filings unchanged. Filed inventions are held fixed, and every difference is paired on one simulated population.

in the model than in the data, 0.63 against 0.89. A given fall in requests therefore narrows granted patents less in the model than the data's slope implies, so the narrowing among broad filings is, if anything, understated. Third, the applicants' part of the fall in grants works through applicants who stop sooner, and the abandonment hazard is the profile the model fits worst, as Section 7.6 shows.

Individual-level monotonicity, the compositional interpretation of the filed-scope gradient, and the fixed set of filed inventions are assumptions the data cannot test, and Appendix N discusses each.

8 Conclusion

This paper asks whether the outcome of patent examination depends on the quality of the invention or on the scope of the request. I measure the quality of an invention from the title and abstract, which do not overlap with the claims, and the scope of the request from the claims, for refused applications as well as granted ones. I show that the grant rate varies far more with scope than with quality. Two sides produce this outcome: the office decides on requests that applicants choose. The office'S decisions show the same pattern. Among granted patents, the share that depends on the assigned examiner increases from about 5.9% for the narrowest scope applications to about 8.6% for the broadest and falls slightly with the quality of the invention. The rules

of examination explain why broad requests are likelier to depend on the assigned examiner. The reasons for rejection test whether the claims are novel, are not obvious, and cover no more than the application discloses. A broader claim has more ways to fail these tests, so broader requested scope is closer to the examiner's acceptance threshold. This holds at every level of quality, although an invention further from earlier work leaves more room for broad claims. So the pattern is a finding about how examination works, not a consequence of patent law.

Because applicants choose the scope, the examiner assignment shows how examiners decide but not what a stricter patent office would change. In my model, a stricter office narrows patents mostly because applicants ask for less scope. Only 10.3% to 15.4% of the narrowing comes from stricter decisions on today's requests. The stricter office also changes which inventions receive patents far less than an equally strict test of invention quality would. Two implications follow. First, evaluations of a stricter patent standard that analyze applications filed before the change capture most of the fall in granted patents but little of the scope narrowing. Second, studies that use the examiner assignment estimate what a patent does based on the grants that depend on the assigned examiner. These grants differ from other grants far more in scope than in the quality of the invention.

The analysis holds fixed which inventions are filed. A stricter office may also change which inventions are developed and filed to the office, which is a margin that the model does not capture. Because patents block follow-on innovation in some technologies (Galasso and Schankerman, 2015), a natural next step is to study what narrower patents do for follow-on innovation.

Like a patent office, many gatekeepers judge the quality of an application and the size of the request. My results are specific to patents, but similar patterns appear in other settings. On a hiring platform where candidates must post the salary they ask for, the ask matters much more for firms' salary offers for a given job than the résumé does, and when the platform changed how it asked candidates for their salary, women asked for more and the gender gap in offers closed (Roussille, 2024). In blinded review of research proposals, reviewers give higher scores to proposals described in broad words, although such wording does not predict better outcomes later (Kolev, Fuentes-Medel, and Murray, 2019). Drug regulators decide on the indications a sponsor proposes, and their guidance suggests that a sponsor can make a favorable assessment more likely by limiting the population it asks for (U.S. Food and Drug Administration, 2023). Where the gatekeeper instead sets its terms by who the applicant is, as a lender sets the minimum down payment by the borrower's credit category, a higher requirement changes who borrows as well as how much, since it screens out riskier borrowers and makes loans smaller (Einav, Jenkins, and Levin, 2012). Wherever applicants choose what to ask for, decisions on today's requests show how strictly a gatekeeper decides, but what a stricter gatekeeper would achieve also depends on how applicants change their requests. The patent office already lets applicants choose how much

protection to ask for, and a broader request lowers the chance of a grant, but the request barely tracks the quality of the invention, so this choice does not make protection follow the invention.

References

Agan, A., J. L. Doleac, and A. Harvey (2023): "Misdemeanor Prosecution," *Quarterly Journal of Economics*, 138, 1453–1505. [20, 26, 45, A-63]

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, 132, 1553–1592. [A-84, A-88, A-91]

Aneja, A., O. Reshef, and G. Subramani (2026): "Attrition and the Gender Patenting Gap," *Review of Economics and Statistics*, 108, 1017–1026. [4]

Angelova, V., W. Dobbie, and C. S. Yang (2026): "Algorithmic Recommendations and Human Discretion," *Review of Economic Studies*, 93, 2250–2283. [22]

Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67. [A-23, A-24]

Arnold, D., W. Dobbie, and P. Hull (2022): "Measuring Racial Discrimination in Bail Decisions," *American Economic Review*, 112, 2992–3038. [22]

Autor, D., A. Kostøl, M. Mogstad, and B. Setzler (2019): "Disability Benefits, Consumption Insurance, and Household Labor Supply," *American Economic Review*, 109, 2613–2654. [5]

Avivi, H. (2025): "Are Patent Examiners Gender Neutral?" Working Paper, University College London. [4]

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450. [A-23]

Bryan, K. A. and H. L. Williams (2021): "Innovation: Market Failures and Public Policies," in *Handbook of Industrial Organization*, ed. by K. Ho, A. Hortaçsu, and A. Lizzeri, Elsevier, vol. 5, 281–388. [16]

Caillaud, B. and A. Duchêne (2011): "Patent Office in Innovation Policy: Nobody's Perfect," *International Journal of Industrial Organization*, 29, 242–252. [4]

Carley, M., D. Hegde, and A. Marco (2015): "What Is the Probability of Receiving a U.S. Patent?" *Yale Journal of Law and Technology*, 17, 203–223. [6]

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2025): "Binscatter Regressions," *Stata Journal*, 25, 3–50. [13, 14]

Dahl, G. B., A. R. Kostøl, and M. Mogstad (2014): "Family Welfare Cultures," *Quarterly Journal of Economics*, 129, 1711–1752. [20]

de Jong, P., M. Lindeboom, and B. van der Klaauw (2011): "Screening Disability Insurance Applications," *Journal of the European Economic Association*, 9, 106–129. [5]

de Rassenfosse, G., W. E. Griffiths, A. B. Jaffe, and E. Webster (2020): "Low-Quality Patents in the Eye of the Beholder: Evidence from Multiple Examiners," *Journal of Law, Economics, and Organization*, 37, 607–636. [10, 26, A-24]

Dobbie, W., J. Goldin, and C. S. Yang (2018): "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108, 201–240. [20, 26, 45]

Einav, L., M. Jenkins, and J. Levin (2012): "Contract Pricing in Consumer Credit Markets," *Econometrica*, 80, 1387–1432. [48]

Executive Office of the President (2013): "Patent Assertion and U.S. Innovation," Report,

The White House, Washington, DC, prepared by the President's Council of Economic Advisers, the National Economic Council, and the Office of Science & Technology Policy. [1, 43]

Farre-Mensa, J., D. Hegde, and A. Ljungqvist (2020): "What Is a Patent Worth? Evidence from the U.S. Patent 'Lottery'," *Journal of Finance*, 75, 639–682. [1, 3, 4, 6, 8, 24, 27, A-26, A-29, A-32, A-59, A-60, A-62, A-63, A-64, A-65, A-66, A-67, A-72, A-73, A-74, A-76, A-78, A-79]

Feng, J. and X. Jaravel (2020): "Crafting Intellectual Property Rights: Implications for Patent Assertion Entities, Litigation, and Innovation," *American Economic Journal: Applied Economics*, 12, 140–181. [2, 4, 9, 28]

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," *Review of Economics and Statistics*, 99, 550–563. [4, 6, 22]

——— (2020): "Procrastination at the Patent Office?" *Journal of Public Economics*, 183, 104140. [6]

——— (2023): "Investing in Ex Ante Regulation: Evidence from Pharmaceutical Patent Examination," *American Economic Journal: Economic Policy*, 15, 151–183. [6]

——— (2024): "Deadlines Versus Continuous Incentives: Evidence from the Patent Office," Working Paper 32066, National Bureau of Economic Research. [6]

Frandsen, B., L. Lefgren, and E. Leslie (2023): "Judging Judge Fixed Effects," *American Economic Review*, 113, 253–277. [25, A-26, A-64]

Freilich, J. and S. Kim (2026): "Is the Patent System Sensitive to Incorrect Information?" *The Review of Economics and Statistics*, 108, 542–552. [17]

Frumkin, J., N. A. Pairolero, A. Tesfayesus, and A. A. Toole (2024): "Patent Eligibility after Alice: Evidence from USPTO Patent Examination," *Journal of Economics & Management Strategy*, 33, 748–769. [5]

Galasso, A. and M. Schankerman (2015): "Patents and Cumulative Innovation: Causal Evidence from the Courts," *Quarterly Journal of Economics*, 130, 317–369. [1, 10, 48]

Gao, T., X. Yao, and D. Chen (2021): "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 6894–6910. [10]

Gaulé, P. (2018): "Patents and the Success of Venture-Capital Backed Startups: Using Examiner Assignment to Estimate Causal Effects," *Journal of Industrial Economics*, 66, 350–376. [4, 6]

Gilbert, R. and C. Shapiro (1990): "Optimal Patent Length and Breadth," *RAND Journal of Economics*, 21, 106–112. [4]

Godard, M., P. Koning, and M. Lindeboom (2024): "Application and Award Responses to Stricter Screening in Disability Insurance," *Journal of Human Resources*, 59, 1353–1386. [5]

Goldsmith-Pinkham, P., P. Hull, and M. Kolesár (2026): "Leniency Designs: An Operator's Manual," *Journal of Economic Perspectives*, 40, 213–240. [3, 4, 21, 22, 24, 25, A-26, A-28, A-29, A-30, A-32]

Gouriéroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, S85–S118. [37, A-80, A-86]

Hansen, N. (2006): "The CMA Evolution Strategy: A Comparing Review," in *Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms*, Springer, vol. 192 of *Studies in Fuzziness and Soft Computing*, 75–102. [A-80]

Harhoff, D. and S. Wagner (2009): "The Duration of Patent Examination at the European

Patent Office," *Management Science*, 55, 1969–1984. [4, 15]

Heckman, J. J. and E. J. Vytlacil (2007): "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6B, 4875–5143. [30, A-58, A-60]

Hegde, D., A. Ljungqvist, and M. Raj (2022): "Quick or Broad Patents? Evidence from U.S. Startups," *Review of Financial Studies*, 35, 2705–2742. [4]

Higham, K., G. de Rassenfosse, and A. B. Jaffe (2021): "Patent Quality: Towards a Systematic Framework for Analysis and Measurement," *Research Policy*, 50, 104215. [10]

Hopenhayn, H., G. Llobet, and M. Mitchell (2006): "Rewarding Sequential Innovators: Prizes, Patents, and Buyouts," *Journal of Political Economy*, 114, 1041–1068. [1]

Hortaçsu, A., F. Luco, S. L. Puller, and D. Zhu (2019): "Does Strategic Ability Affect Efficiency? Evidence from Electricity Markets," *American Economic Review*, 109, 4302–4342. [40]

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475. [21]

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021): "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 3, 303–320. [9, 10, 11, 12, 13, 15, 28, A-13, A-53, A-68]

Klemperer, P. (1990): "How Broad Should the Scope of Patent Protection Be?" *RAND Journal of Economics*, 21, 113–130. [4]

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez (2011): "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark," *Econometrica*, 79, 651–692. [5]

Kline, P., R. Saggio, and M. Sølvsten (2020): "Leave-Out Estimation of Variance Components," *Econometrica*, 88, 1859–1898. [A-19, A-23, A-24]

Kline, P. and C. R. Walters (2019): "On Heckits, LATE, and Numerical Equivalence," *Econometrica*, 87, 677–696. [30]

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017): "Technological Innovation, Resource Allocation, and Growth," *Quarterly Journal of Economics*, 132, 665–712. [9, 11, 12, 13, 15, A-13, A-48, A-68]

Kolesár, M. (2013): "Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity," Working paper, Cowles Foundation, Yale University, Version 1.3.1, November 5, 2013. [22, A-23, A-24]

Kolev, J., Y. Fuentes-Medel, and F. Murray (2019): "Is Blinded Review Enough? How Gendered Outcomes Arise Even Under Anonymous Evaluation," Working Paper 25759, National Bureau of Economic Research, Cambridge, MA. [48]

Kuhn, J. M. and N. C. Thompson (2019): "How to Measure and Draw Causal Inferences with Patent Scope," *International Journal of the Economics of Business*, 26, 5–38. [2, 9, A-55]

Lemley, M. A. and C. Shapiro (2005): "Probabilistic Patents," *Journal of Economic Perspectives*, 19, 75–98. [10]

Llobet, G., Á. Parra, and J. Suarez (2026): "Optimal Patent Screening with Imperfect Enforcement," *International Journal of Industrial Organization*, 105, 103248. [4]

Lu, Q., A. Myers, and S. Beliveau (2017): "USPTO Patent Prosecution Research Data: Un-

locking Office Action Traits," Economic Working Paper 10, United States Patent and Trademark Office. [9]

Marco, A. C., J. D. Sarnoff, and C. A. deGrazia (2019): "Patent Claims and Patent Scope," *Research Policy*, 48, 103790. [7, 15]

Matcham, W. and M. Schankerman (2026): "Screening Property Rights for Innovation," *Econometrica*, 94, 1849–1886. [4, 5, 36, 46]

Melero, E., N. Palomeras, and D. Wehrheim (2020): "The Effect of Patent Protection on Inventor Mobility," *Management Science*, 66, 5485–5504. [4]

Morris, C. N. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–55. [22, A-19]

Munnell, A. H., G. M. B. Tootell, L. E. Browne, and J. McEneaney (1996): "Mortgage Lending in Boston: Interpreting HMDA Data," *American Economic Review*, 86, 25–53. [1]

Ni, J., G. Hernández Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang (2022): "Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models," in *Findings of the Association for Computational Linguistics: ACL 2022*, Association for Computational Linguistics, 1864–1874. [10, A-46]

Okada, Y., Y. Naito, and S. Nagaoka (2018): "Making the Patent Scope Consistent with the Invention: Evidence from Japan," *Journal of Economics & Management Strategy*, 27, 607–625. [15]

Pairolero, N. A. and C. A. W. de Grazia (2026): "Automated Routing, Patent Examiner Specialization, and Quasi-Random Assignment at the USPTO," *Research Policy*, 55, 105460. [2]

Pennington, J., R. Socher, and C. Manning (2014): "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ed. by A. Moschitti, B. Pang, and W. Daelemans, Doha, Qatar: Association for Computational Linguistics, 1532–1543. [10]

Productivity Commission (2016): "Intellectual Property Arrangements," Inquiry Report 78, Productivity Commission, Canberra. [1, 43]

Righi, C. and T. Simcoe (2019): "Patent Examiner Specialization," *Research Policy*, 48, 137–148. [6, 24, 25]

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954. [A-22]

Romano, J. P. and M. Wolf (2005): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73, 1237–1282. [A-60, A-73]

——— (2016): "Efficient Computation of Adjusted p-Values for Resampling-Based Stepdown Multiple Testing," *Statistics & Probability Letters*, 113, 38–40. [A-60, A-73]

Roussille, N. (2024): "The Role of the Ask Gap in Gender Pay Inequality," *Quarterly Journal of Economics*, 139, 1557–1610. [1, 48]

Sampat, B. and H. L. Williams (2019): "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome," *American Economic Review*, 109, 203–236. [2, 4, 6, 27, A-26, A-41, A-59, A-60, A-61, A-62, A-63, A-64, A-65, A-66, A-67, A-72, A-73, A-75, A-77]

Schankerman, M. and F. Schuett (2022): "Patent Screening, Innovation, and Welfare," *Review of Economic Studies*, 89, 2101–2148. [1]

Schuett, F. (2013): "Inventors and Impostors: An Analysis of Patent Examination with Self-Selection of Firms into R&D," *Journal of Industrial Economics*, 61, 660–699. [4]

Scotchmer, S. (2004): *Innovation and Incentives*, Cambridge, MA: MIT Press. [18]

Shi, M. (2024): "Monitoring for Waste: Evidence from Medicare Audits," *Quarterly Journal of Economics*, 139, 993–1049. [5]

Smith, A. A. (1993): "Estimating Nonlinear Time-Series Models Using Simulated Vector Autoregressions," *Journal of Applied Econometrics*, 8, S63–S84. [37]

Toole, A. A. and N. A. Pairolero (2020): "Adjusting to Alice: USPTO Patent Examination Outcomes after Alice Corp. v. CLS Bank International," IP Data Highlights 3, United States Patent and Trademark Office, Office of the Chief Economist, Alexandria, VA. [5]

U.S. Federal Trade Commission (2003): "To Promote Innovation: The Proper Balance of Competition and Patent Law and Policy," Report, Federal Trade Commission, Washington, DC. [1, 43]

U.S. Food and Drug Administration (2023): "Benefit-Risk Assessment for New Drug and Biological Products," Guidance for industry, U.S. Department of Health and Human Services, Food and Drug Administration, Silver Spring, MD. [48]

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017): "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc., vol. 30, 5998–6008. [10]

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. [30]

Zhou, X. and Y. Xie (2019): "Marginal Treatment Effects from a Propensity Score Perspective," *Journal of Political Economy*, 127, 3070–3084. [A-60]

Appendix

Table of Contents

A Summary statistics A-3

B Proof of Proposition 1 A-4

B.1 Assessment noise and the "hard to judge" reading A-5

C Proof of Proposition 2 A-7

D Additional output A-11

E Robustness: many-examiner instruments and the UJIVE estimator A-17

F Instrument validity: monotonicity and complier balance A-26

G Robustness: sample restrictions A-40

H Robustness: repeat filers A-42

I Quality measure: construction, validation, and decomposition A-45

J Scope measure: construction robustness and normalization dependence A-54

K What the examiner assignment identifies: marginal treatment effect curves A-58

L Regression tables underlying the figures A-68

L.1 Conditional correlations A-68

L.2 Complier shares A-69

L.3 Marginal treatment effects A-73

M The model: specification, identification, and fit A-80

M.1 Parameters A-80
M.2 Identification and inference A-84
M.3 Fit A-96
M.4 Sensitivity to the offer's leniency gradient and floor A-105
M.5 Counterfactual losses by filed scope A-105

N Limitations of the design A-106

A Summary statistics

Table A.1 reports application-level summary statistics for the estimation window of Section 3.1. The leniency rows show the raw material of the design: the examiner's running grant rate disperses widely (SD 0.185) around the same mean as its art-unit expectation, and their difference, the recentered deviation Δl, is centered on zero with a standard deviation of 0.129. The scope rows are reported in words per claim, the raw inverse of scope, before the within-year percentile transformation of Section 3.2. Granted claims carry more words than filed claims on average. The granted mean runs over the granted subsample while the filed mean runs over all applications, so the contrast mixes the narrowing prosecution produces with selection into grant, and the paired within-application evidence is in Section 3.2. Forward citations to the publication are defined for granted and rejected applications alike, which is the coverage property the validation exercises of Section 3.4 use.

Table A.1: *Summary statistics, estimation sample*

	N	Mean	SD	P10	Median	P90
Granted	5,367,642	0.711	0.453	0.000	1.000	1.000
Examiner leniency (leave-one-out)	5,367,642	0.661	0.185	0.400	0.692	0.878
Expected leniency (art-unit pool)	5,367,642	0.661	0.133	0.485	0.679	0.813
Leniency deviation Δl	5,367,642	-0.000	0.129	-0.170	0.014	0.149
Invention quality (within-cell percentile)	5,364,971	0.506	0.289	0.110	0.510	0.910
Filed scope (words per claim)	5,367,642	55.7	40.8	27.8	47.3	89.3
Granted scope (words per claim)	3,772,516	69.8	49.8	33.5	58.5	113.8
Revision rounds (non-final rejections)	5,363,782	1.27	0.89	0.00	1.00	2.00
Forward citations (to the publication)	5,367,642	10.28	35.62	0.00	3.00	22.00
Realized renewals	3,807,772	-0.35	0.71	-1.00	0.00	0.00

Notes: Application-level summary statistics for the estimation window, publication years 2005–2020 (5,367,642 applications; the examiner-dummy estimators drop a further three observations, see Section 3.1). *Granted* is the grant indicator. *Examiner leniency* is the grant rate over the assigned examiner's earlier-filed applications, each counted with its final outcome; *expected leniency* averages the leniency of examiners deciding in the same art unit within an eighteen-month window on either side of the decision date; the *deviation* is their difference, the recentered variation the design uses. *Invention quality* is the text-based measure of Section 3.3, expressed as a within-art-unit $\times$ publication-year percentile rank on the unit interval. *Filed* and *granted scope* are the average number of words per claim in the application and the issued patent respectively; more words delineate a narrower claim, so word count is the inverse of scope (Section 3.2), and the granted row conditions on grant. *Revision rounds* counts non-final rejections. *Forward citations* count citations to the application publication. *Realized renewals* are observed for granted patents only, computed as the maintenance fees paid minus the fee decisions observable by June 2023, so zero means every observable fee was paid.

B Proof of Proposition 1

Write the standardized acceptance index at the pool mean leniency as $t(q,s) = \big(a(q,s) + \bar{l}\big)/\sigma$, so that the grant probability is $P = \Phi(t)$ and the first-stage sensitivity to leniency is $\psi = \partial P/\partial l\,|_{l=\bar{l}} = \sigma^{-1}\varphi(t)$. The model's grant probability is not linear in leniency, so assumption (iii) of Section 5.3 holds in the model to first order in the deviation of leniency from the pool mean, and the proposition concerns this first-order share. Equation (11) writes the share of compliers among granted patents as the sensitivity times the dose over the grant rate. In the units of the model's leniency l, the share is $\phi_c^{D=1} = \psi\,\bar{d}\,/\,P$, with $\bar{d} = \mathbb{E}[(\Delta l)_+]$ the favorable deviation from the pool mean, a constant that does not depend on (q,s) under quasi-random assignment. Substituting,

$$\phi_c^{D=1}(q,s) \;=\; \frac{\bar{d}}{\sigma}\,\frac{\varphi\big(t(q,s)\big)}{\Phi\big(t(q,s)\big)} \;=\; \frac{\bar{d}}{\sigma}\,\lambda\big(t(q,s)\big),$$

where $\lambda(t) = \varphi(t)/\Phi(t)$ is the inverse Mills ratio.

Part (i). Since $\bar{l}$, σ and $\bar{d}$ are common to all applications, the share depends on (q,s) only through $t(q,s)$, and hence only through $a(q,s)$. The inverse Mills ratio is strictly decreasing on $\mathbb{R}$: $\lambda'(t) = -\lambda(t)\big(t + \lambda(t)\big) < 0$, since $\lambda(t) > 0$ and $t + \lambda(t) > 0$ for all t. The second inequality holds because $-\lambda(t) = \mathbb{E}[Z \mid Z < t]$ for a standard normal Z, and this conditional mean lies below t. Hence $\phi_c^{D=1}$ is a strictly decreasing function of t, and therefore of the acceptance index: for any two applications, $\phi_c^{D=1}(q,s) > \phi_c^{D=1}(q',s')$ if and only if $a(q,s) < a(q',s')$. Since $P = \Phi(t)$ is strictly increasing in t, the same ordering holds with the grant probability in place of the acceptance index, and $\phi_c^{D=1} = (\bar{d}/\sigma)\,\lambda\big(\Phi^{-1}(P)\big)$ is a strictly decreasing function of P. □

Part (ii). The treated-complier mass, the probability that an application is granted *because* the assigned examiner's leniency exceeded the expected leniency, is $\psi\,\bar{d} = (\bar{d}/\sigma)\,\varphi\big(t(q,s)\big)$, with $\bar{d} = \mathbb{E}[(\Delta l)_+]$ as above. This is the one-sided object. The two-sided flip rate ϕ_c of (12) replaces $\bar{d}$ with $\mathbb{E}[|\Delta l|]$ and is twice as large under mean-zero recentering. Both are proportional to $\varphi\big(t(q,s)\big)$, so the shape argument is common to them and only the level differs. The standard normal density φ is strictly single-peaked at 0, so each is single-peaked in the acceptance index and attains its maximum where $t(q,s) = 0$, i.e. where $P(q,s,\bar{l}) = \Phi(0) = \frac{1}{2}$. □

Remark. The proposition holds the assigned examiner's leniency fixed at the pool mean leniency $\bar{l}$, at which the profiles are evaluated. Additive separability of l, together with a common noise scale σ, is what makes ψ depend on (q,s) only through a. A leniency-characteristic interaction would multiply ψ by the pass-through of leniency into the index, and a noise scale that varies with (q,s)

would change σ across applications. Either would make the composition depend on (q, s) beyond the index, and Appendix B.1 estimates both. Neither part is new as mathematics. Part (i) is the monotonicity of the inverse Mills ratio and part (ii) the unimodality of the normal density. What is specific to this setting is that the argument of both is an acceptance index in which the applicant's own filed scope enters, so the composition is partly a statement about applicant behavior and not only about the screening technology.

B.1 Assessment noise and the "hard to judge" reading

The Remark above holds the assessment scale σ fixed and names a leniency-characteristic interaction as the testable deviation. Table B.1 measures both. A probit of the grant outcome on the recentered leniency deviation, run within a bin of quality or of filed scope, has slope β/σ and cannot on its own say which of the two moves. The art unit $\times$ year pool effect enters the acceptance index additively rather than through β, so its slope in the same probit is $\bar{\sigma}/\sigma$, which varies with σ alone. The ratio of the two slopes is $\beta/\bar{\sigma}$, so both are recovered up to one common constant and their variation across bins can be read separately. The probit identifies the total scale σ of Equation (6) and does not separate the signal component $\alpha_q f \sigma_\eta$ from σ_{grant}. The argument below applies to either. The separation of β and σ is not a refinement. Imposing $\beta \equiv 1$ returns a σ that rises by seven percent across quality, and freeing β reverses that to a fall of fifteen percent, with the pass-through falling by twenty.

Neither dimension supports the "hard to judge" reading. Noise falls across quality rather than rising, so the weakest applications carry the most assessment noise. Weighting applications by the flip rate puts the compliers' noise 0.32% above the population mean, against a spread of half across cells. The reading fails on the mechanism before it fails on the magnitude. Since $\psi = \varphi(t)/\sigma$, the derivative $\partial\psi/\partial\sigma$ carries the sign of $t^2 - 1$. A noisier assessment therefore adds to the margin only more than one standard deviation from the threshold, which means a grant probability outside $[0.16, 0.84]$, and no cell of the surface is there. Over the range the data cover, noise thins the margin rather than supplying it.

Two further senses of "hard to judge" concern the quality measure rather than the examiner, and Table I.5 in Appendix I reports both. The first is imprecision, the disagreement between scores built on two independent halves of the text corpus at fixed construction, which loads on compliers negatively or not at all. The second is reference dependence, the disagreement between the art-unit and the coarser technology-center comparison pool, on which compliers rank 0.55 percentage points higher once ranked within art unit $\times$ year. Both are properties of the measure rather than of the examiner's assessment, and both are an order of magnitude below the filed-scope gradient that carries the composition.

Table B.1: *Assessment noise, leniency pass-through, and compliers against the population*

	β/σ	$\bar{\sigma}/\sigma$	σ	β
Assessment noise across quality, filed scope pooled				
lowest quality bin	2.484	0.523	1.000	1.000
2nd quality bin	2.401	0.555	0.942	0.910
3rd quality bin	2.365	0.582	0.898	0.854
4th quality bin	2.309	0.605	0.863	0.802
5th quality bin	2.323	0.602	0.869	0.812
6th quality bin	2.310	0.610	0.856	0.796
7th quality bin	2.327	0.611	0.855	0.801
highest quality bin	2.331	0.616	0.848	0.796
Assessment noise across filed scope, quality pooled				
narrowest scope bin	2.326	0.646	1.000	1.000
2nd scope bin	2.396	0.592	1.093	1.125
3rd scope bin	2.422	0.591	1.094	1.139
4th scope bin	2.406	0.573	1.129	1.168
5th scope bin	2.389	0.564	1.147	1.177
6th scope bin	2.366	0.563	1.147	1.167
7th scope bin	2.372	0.557	1.161	1.184
broadest scope bin	2.262	0.552	1.171	1.139
Compliers less population, weights from Figure D.5				
quality (percentile points)	−0.68			
filed scope (percentile points)	+0.81			
assessment noise σ (percent)	+0.317			
examiner bootstrap s.e.	(0.036)			
95% interval	[+0.25, +0.39]			
Dispersion of σ, and the sign of $\partial\psi/\partial\sigma$				
across cells, max/min	1.533			
across cells, s.d. (percent of mean)	8.6			
cells where noise raises the margin	0 of 64			

Notes: The Table reports the assessment-shock scale σ of Equation (5) and the leniency pass-through β, estimated on the estimation sample ($n = 5,353,106$ applications, 12,698 examiners, art unit × publication year pools of at least 30). Within each bin, a probit of the grant outcome on the recentered leniency deviation has slope β/σ, which on its own cannot say which of the two varies. The art unit × year pool effect enters the acceptance index additively rather than through β, so its slope in the same probit is $\bar{\sigma}/\sigma$, which varies with σ alone, and the ratio of the two slopes is $\beta/\bar{\sigma}$. The pool effect is measured as the leave-one-out pool grant rate on the probit scale, so an application never enters its own regressor. Both implied columns are relative to the first row of their block and are identified up to one common constant, so levels carry no units and only ratios across bins are interpretable. Complier means are the modified-outcome form $\mathbb{E}[X \mid \text{complier}] = \sum_i \psi_i X_i / \sum_i \psi_i$ with ψ the surface of Figure D.5, matched to applications on quality and filed scope. The same weights return the complier quality gap of Section 6.3 to within a tenth of a percentile point, which is the check that they are being applied correctly. Standard errors resample examiners with replacement, $B = 300$. They hold ψ and the pool effect at their point values, so the sampling error of the complier surface itself is not propagated. The last block reports the dispersion against which the complier displacement should be read, and the count of cells in which a noisier assessment would raise rather than thin the complier margin. That count follows from $\partial\psi/\partial\sigma$ carrying the sign of $z^2 - 1$ at $z = \Phi^{-1}(P)$, so noise adds to the margin only where the cell grant rate falls outside $[0.16, 0.84]$. Quality is a percentile within art unit × publication year, so every pool enters every quality bin equally and the quality block is balanced on pool composition by construction. Filed scope is ranked within publication year, so pools enter its bins unequally and the scope block should be read as directional.

C Proof of Proposition 2

Setup and solution concept. There are two parties. An applicant A holds a joint prior $w^0 \in \Delta(Q \times L)$ over own quality $q \in Q = [0,1]$ and examiner leniency $l \in L = [0,1]$. An examiner E holds a prior $\phi^0 \in \Delta(Q)$ over application quality and applies a threshold grant rule g_E together with the offer (17), both taken as given: the examiner has no objective function and does not optimize. The solution concept is correspondingly one-sided, and deliberately so. A solution consists of three objects: (i) an applicant policy σ_A^* mapping each history h_r to an action in $S_A(h_r)$, (ii) a belief system μ_A^* assigning posterior $w_r^A \in \Delta(Q \times L)$ at each h_r via Bayes' rule, and (iii) an examiner belief system μ_E^* assigning $\phi_r^E \in \Delta(Q)$ at each h_r via Bayes' rule from signals $\xi_r = q + \eta_r$. Optimality requires σ_A^* to maximize the applicant's expected payoff at every information set given μ_A^* and g_E. This is a single-agent stochastic control problem with an exogenous, empirically disciplined environment, not an equilibrium of a game. No mutual best-response condition is imposed or claimed.

Decoupling. Because the examiner forms quality beliefs exclusively from the private signal $\xi_r = q + \eta_r$–and not from the applicant's scope choice s_r–the examiner's posterior ϕ_r^E evolves independently of σ_A. From the applicant's perspective, ϕ_r^E is therefore an exogenous stochastic process that can be forecast without reference to own strategy. This decouples the two belief systems: the applicant can treat the examiner's dynamic threshold as a feature of the environment rather than a strategic response. The applicant's problem reduces to a finite-horizon single-agent stochastic control problem, which is solved by backward induction over rounds $r \in \{0, \ldots, R-1\}$, the text's convention, with $R-1$ the terminal round.

State and action spaces. The applicant's state at round r is the pair $(w_r^A, \bar{s}_r)$: the posterior $w_r^A \in \mathcal{P} \equiv \Delta(Q \times L)$, endowed with the weak topology and compact and convex, and the *scope ceiling* $\bar{s}_r \in [0, \bar{s}_A]$. The action space at round r is

$$S_A(\bar{s}_r) \;=\; [0,\, \bar{s}_r], \qquad \bar{s}_0 = \bar{s}_A, \qquad \bar{s}_r = s_{r-1} \text{ for all } r \geq 1,$$

so prosecution may narrow claims after a rejection but never re-broaden them. Each claim caps the next, which is exactly the constraint the solved model imposes. The new-matter bar caps every amendment at the initial disclosure, and the ratchet tightens that cap to the claim last rejected. The permanence of granted-scope concessions (Sections 2 and 5.1) operates through the offer (17), not through the action set. $\bar{s}_A$ is the upper bound on filed scope imposed by existing prior art, and 0 is the outside option of abandoning the application. Each $S_A(\bar{s}_r)$ is compact and convex,

and the correspondence $\bar{s} \mapsto S_A(\bar{s})$ is continuous (indeed both upper and lower hemicontinuous, being an interval with continuous endpoints), which is what Berge's theorem requires below. The grant probability $d_r(s, w_r^A)$, the granted scope $\hat{s}(s)$, and the revision cost $C_A(s)$ are continuous in s by the smoothness of the normal CDF and the continuity of the cost structure, and payoffs are bounded by construction.

Backward induction – base case ($r = R - 1$). At the terminal round the game ends whatever happens, so both the continuation branch and the abandonment branch of (19) are worth zero and the applicant maximizes the terminal instance of that objective,

$$U_A^{R-1}(s \mid w_{R-1}^A) \;=\; \omega(q)\Big[d_{R-1}(s, w_{R-1}^A) \cdot \nu s \;-\; \kappa s^{\chi}\Big],$$

with κs^{χ} the cost of claiming scope. d_{R-1} is jointly continuous in (s, w_{R-1}^A): the grant probability integrates a bounded function of (q, l, s), continuous in s by the smoothness of the normal CDF, against w_{R-1}^A, and integration of a bounded continuous function is continuous in the measure under the weak topology. Hence U_A^{R-1} is jointly continuous in state and action, and continuous in s on the compact set $S_A(\bar{s}_{R-1})$. By the Extreme Value Theorem the maximum is attained, so the argmax correspondence $\sigma_{R-1}^*(w_{R-1}^A, \bar{s}_{R-1}) = \operatorname{argmax}_{s \in S_A(\bar{s}_{R-1})} U_A^{R-1}(s \mid w_{R-1}^A)$ is non-empty at every state. By Berge's Maximum Theorem, applicable because $S_A(\cdot)$ is a continuous, compact-valued correspondence and the objective is jointly continuous, the value function $V_{R-1}(w_{R-1}^A, \bar{s}_{R-1}) = \max_{s \in S_A(\bar{s}_{R-1})} U_A^{R-1}(s \mid w_{R-1}^A)$ is bounded and continuous in the state.

Backward induction – inductive step. Suppose at round $r+1$ there exists a non-empty argmax correspondence σ_{r+1}^* and a bounded, continuous value function $V_{r+1}(w_{r+1}^A, \bar{s}_{r+1})$. Upon a rejection at round r the applicant updates beliefs via Bayes' rule:

$$w_{r+1}^A \;=\; B^A(w_r^A,\, s,\, \text{reject}) \;\propto\; w_r^A(q, l) \cdot \Pr(\text{reject} \mid q, l, s),$$

where the rejection likelihood is the marginal over the examiner's signal history ϕ_r^E, which the applicant never observes: from the grant rule (16),

$$\begin{aligned}\Pr(\text{reject} \mid q, l, s) &= \mathbb{E}\Big[\Pr(\text{reject} \mid q, l, s, \phi_r^E) \,\Big|\, q\Big] \\ &= \Phi\left(\frac{\mu - \alpha_q\, \mathbb{E}[\mathbb{E}_r[q] \mid q] + b_q(q - \tfrac{1}{2}) + b_{0q}(q - \tfrac{1}{2})\mathbf{1}[r = 0] + \Pi(s) + b_r\, r + b_0\, \mathbf{1}[r = 0] - l}{\sigma_{\text{tot}}(r)}\right),\end{aligned}$$

the second equality exact under the Gaussian learning structure because the dispersion of the realized posterior mean around its average given true quality is precisely the $\sigma_e(r)$ component

of $\sigma^2_{\text{tot}}(r)$ in (34), with Π the scope penalty. Rejection is the event the applicant conditions on. Whether they then revise or abandon is their own decision and carries no further information about (q, l). The likelihood is bounded and continuous in (q, l, s) by the smoothness of Φ and of Π, so the grant probability $d_r(s, w_r^A) = 1 - \int \Pr(\text{reject} \mid q, l, s)\, \mathrm{d}w_r^A(q, l)$, now defined solely from the stated applicant state, is jointly continuous in (s, w_r^A) under the weak topology, and the Bayes map $(w_r^A, s) \mapsto B^A(w_r^A, s)$ is weakly continuous, its normalizer bounded away from zero because Φ is interior on the compact support. The round-r objective is

$$U_A^r(s \mid w_r^A, \bar{s}_r) \;=\; \omega(q)\Big[d_r(s, w_r^A)\cdot \nu s \,-\, \kappa s^\chi\Big] \,+\, \Big(1 - d_r(s, w_r^A)\Big)\max\Big\{0,\; -C_A(s, r) + \beta\, V_{r+1}\Big(B^A(w_r^A, s),\; s\Big)\Big\},$$

the objective (19), with κs^χ the cost of claiming scope, $C_A(s, r)$ the round-r revision cost, and the outer maximum the abandonment option. Since d_r, C_A, V_{r+1}, and B^A are continuous, so is their composition, and the pointwise maximum of two continuous functions is continuous. The ceiling is the previous claim, part of the state, and the constraint correspondence $\bar{s} \mapsto [0, \bar{s}]$ is continuous. Hence U_A^r is jointly continuous in state and action, and continuous on the compact set $S_A(\bar{s}_r)$. By the Extreme Value Theorem the argmax $\sigma_r^*(w_r^A, \bar{s}_r)$ is non-empty at every state, and by Berge's Maximum Theorem, whose hypotheses of a continuous compact-valued constraint correspondence and a jointly continuous objective are now verified, the resulting value function V_r is bounded and continuous, closing the induction.

Remark on set-valued maximizers. Continuity and compactness deliver existence outright, with no appeal to randomization: quasiconcavity is not required by the Extreme Value Theorem and is not assumed here. Where U_A^r fails to be strictly quasiconcave in s the argmax may be set-valued, and any selection from it, or any distribution over it, is optimal. In the solved model the action set is a finite grid, so a measurable pure selection always exists and is what the numerical solution returns. Non-monotonicity of $\partial U_A^r / \partial s$ therefore at most generates multiple local optima, never a requirement to mix, and the existence and pure-selection conclusions stand regardless.

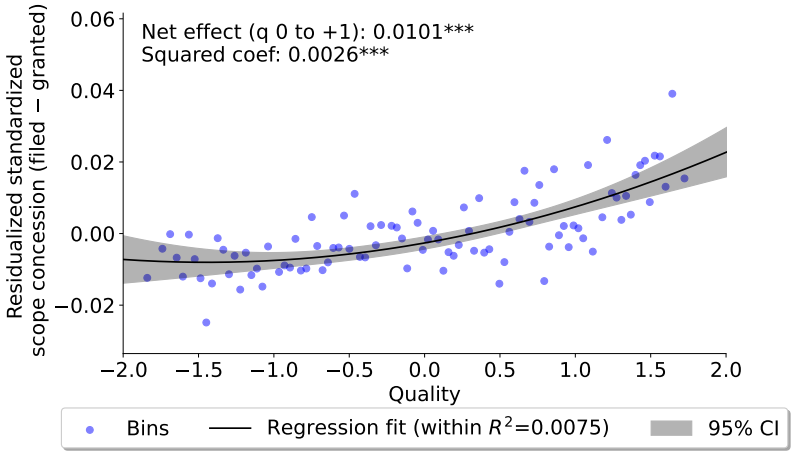
Belief consistency. The applicant's belief system μ_A^* is constructed by applying B^A at every on-path history under σ_A^*, so it is consistent with σ_A^* and g_E by construction. The examiner's belief system μ_E^* follows from the signal process independently of σ_A^*, and is similarly consistent by construction. Beliefs at histories reached with probability zero are unrestricted and play no role, since the examiner's rule g_E is exogenous and does not respond to them.

Conclusion. Backward induction delivers a non-empty optimal policy σ_r^* at every round r and state $(w_r^A, \bar{s}_r)$, together with a bounded, continuous value function. The triple $(\sigma_A^*, \mu_A^*, \mu_E^*)$ satisfies

sequential optimality for the applicant and on-path Bayes consistency for both parties, which is Proposition 2. □

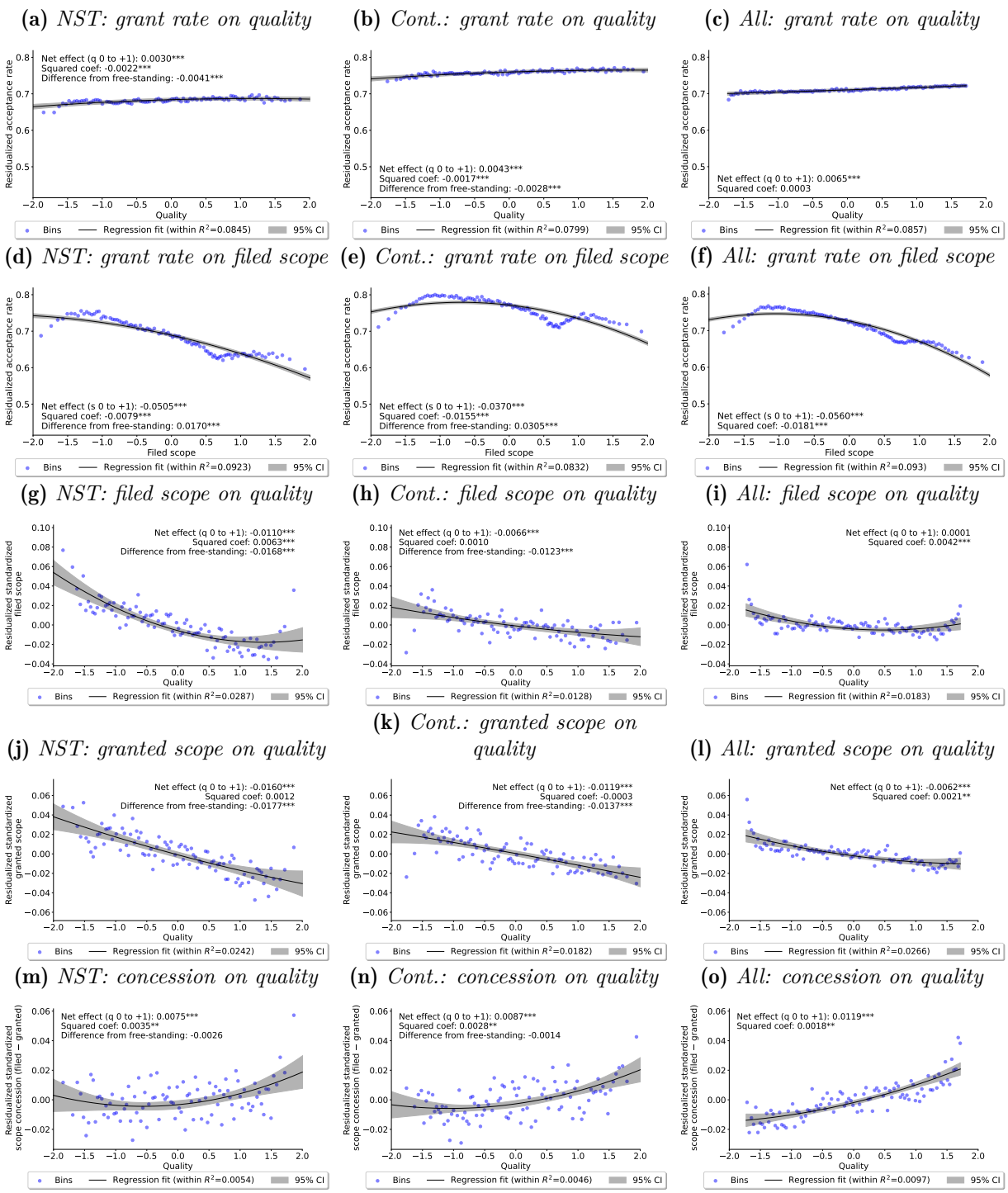
D Additional output

Figure D.1: *Conditional correlation: difference between filed and granted scope by quality*



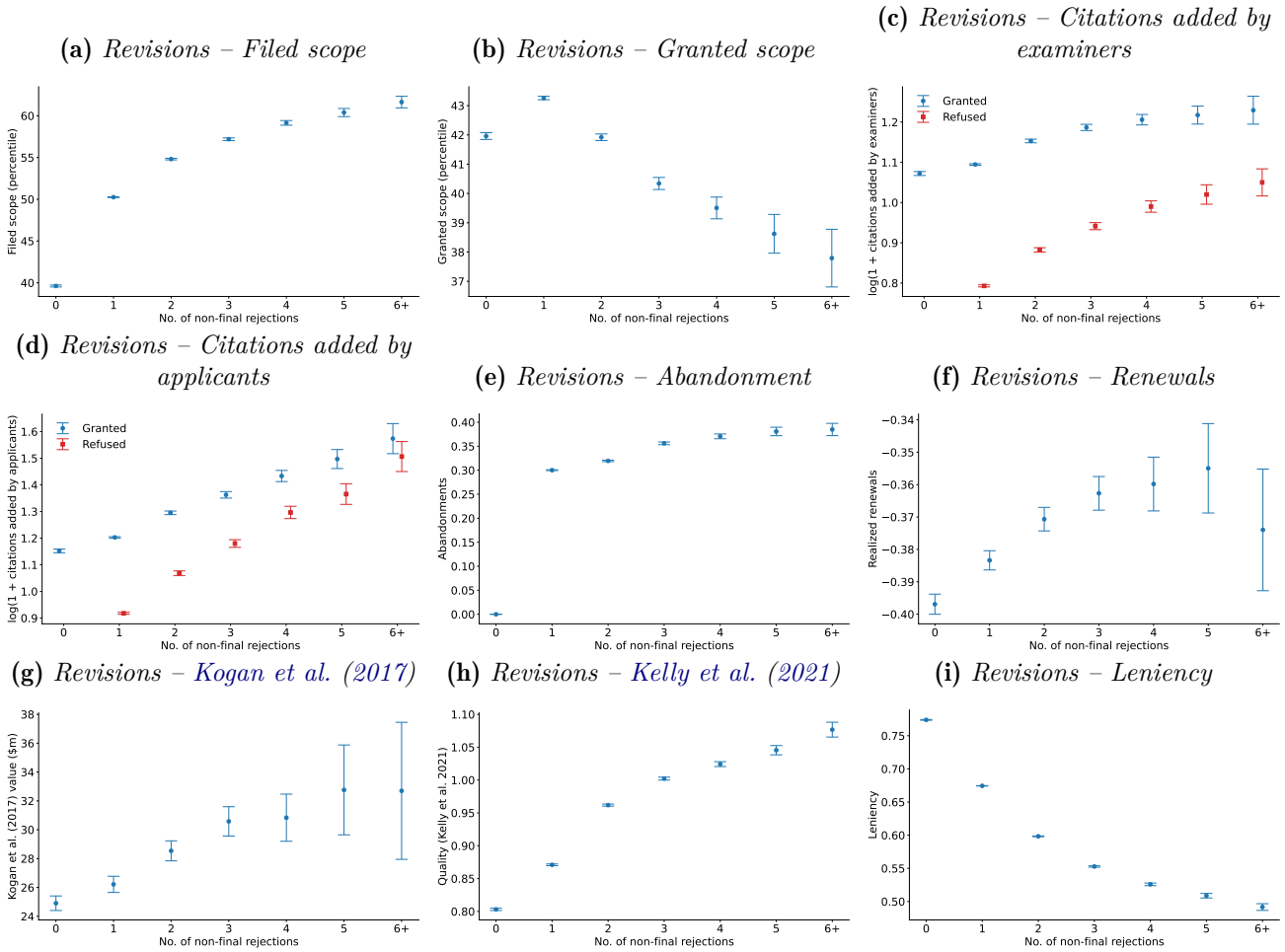
Notes: Binned scatter of the residualized difference between filed and granted scope against standardized invention quality for free-standing granted applications (n = 2,024,630), on the specification of Figure 2. Both scopes are percentiles among the filed and granted claims of the same year, and their difference is standardized to unit standard deviation, so a positive value is a narrowing of the claims between filing and issue. The linear and squared terms and the net effect are reported in column (5) of Table L.2, and Figure D.2 shows the other filing groups.

Figure D.2: *Conditional correlations by filing group*



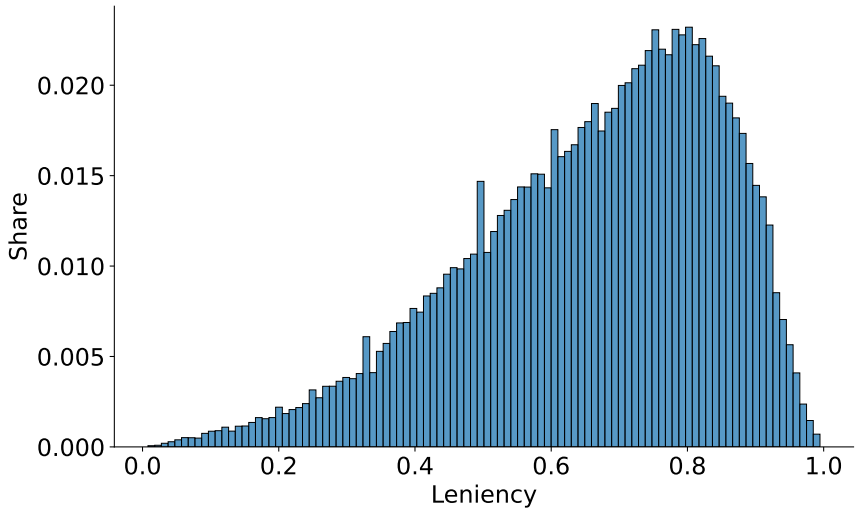
Notes: The figure repeats Figure 2 and Figure D.1 for the other filing groups. The left column shows national-stage entries without a US parent, the middle column continuations, divisionals and continuations-in-part, and the right column all applications. The regressions are the interacted model of Figure 2, estimated on all applications, and each annotation in the left and middle columns adds the difference of the group's net effect from the free-standing one, with standard errors from the joint covariance of the stacked model. Each row shares one vertical scale with the corresponding free-standing panel, and the two grant-rate rows share one scale.

Figure D.3: *Additional patent and patent application attributes by the number of revisions*



Notes: The Figure shows the mean values of patent and patent application attributes by the number of review rounds the patent application had to go through before being granted, abandoned, or receiving a final rejection. Because the number of patent applications going through multiple revision rounds decreases rapidly, I top-code the number of revisions at 6. The error bars span 3.2 standard errors of the mean on either side, a Bonferroni-adjusted 95% level for 35 simultaneous intervals. Panels D.3c and D.3d show citations to the invention added by examiners and by applicants, constructed as Panel 3c of Figure 3. Panels D.3f, D.3g and D.3h show the round-0 mean plus each round's difference from round 0 within issue year, from a regression of the outcome on round dummies with issue-year fixed effects among granted applications, because these outcomes depend on the time since issue and applications with more rounds issue later. Their error bars use the heteroskedasticity-robust standard errors of these differences. The panels complete the main-text Figure 3.

Figure D.4: *Distribution of examiner leniency*



Notes: The Figure plots the distribution of examiner leniency, measured as the grant rate over the same examiner's earlier-filed applications, each counted with its final outcome. The sample is the 2005–2020 estimation sample described in Section 3.1. The measure is undefined for an examiner's first-filed application, and leniency values of exactly zero or one are excluded. The dashed vertical line marks the sample median. The interquartile range spans approximately 26.5 percentage points, against a mean grant rate of approximately 66%. This measure sets the level of the complier shares of Section 6 and supplies the leniency distribution of the model in Section 7.

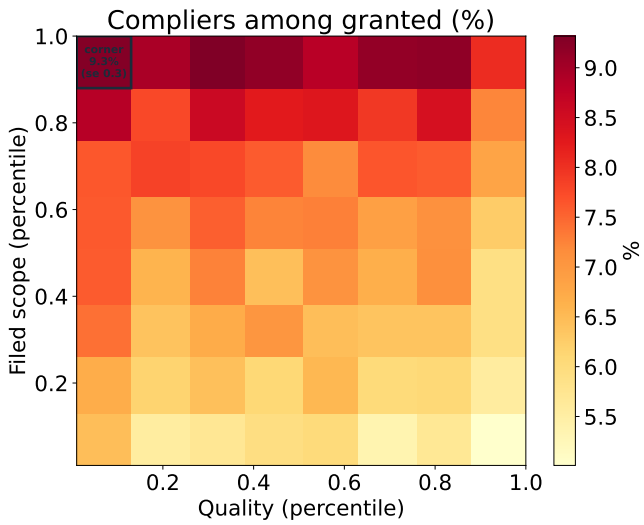
Table D.1: *Unconditional first stage*

	FE-OLS	UJIVE
Leniency pass-through $\bar{\psi}$	0.667	1.062
	(0.001)	(0.004)
Art unit × year FE	✓	✓
Controls	✓	✓
Examiner leave-out		✓
Observations	5,364,969	5,364,969

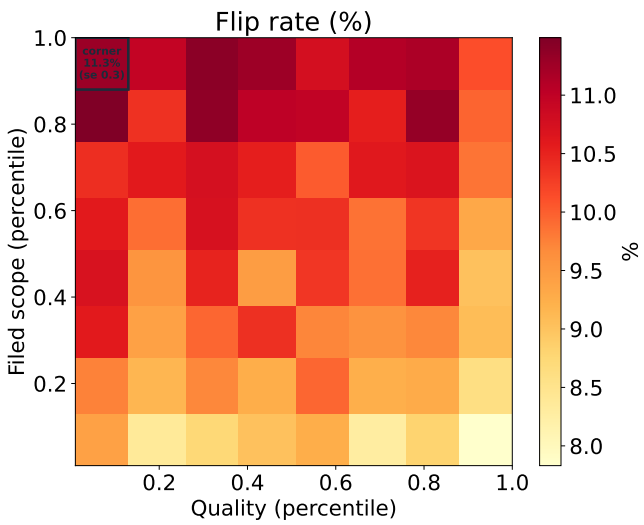
Notes: Unconditional first stage of the grant indicator on examiner leniency, on the 2005–2020 core sample (publication years; largest connected set of 12,705 examiners over 9,701 art unit × year cells). Column 1 is the fixed-effects OLS slope of grant on the recentered leniency deviation Δl_p, controls absorbed by Frisch–Waugh. Column 2 instruments that deviation with the examiner leave-out fit (UJIVE). The two constructions do different work: the leave-out fit, shared by both columns, removes the own-observation bias that inflates a raw leniency regression, while instrumenting the noisy leave-one-out leniency measure corrects its measurement-error attenuation, which is why Column 2 is larger and Column 1 is the conservative floor. Controls are small- and micro-entity indicators, continuation-type dummies, examiner experience, and a first-year-active indicator. Standard errors clustered on art unit × publication year (9,701 clusters). The two columns are the two pass-through conventions discussed in Appendix D; the main text uses the FE-OLS figure, the conservative of the two. Mean favourable deviation $\bar{d} = 0.0503$ and grant rate 0.7106.

Figure D.5: *Complier composition over the quality-scope plane*

(a) *Compliers among granted patents (%)*



(b) *Flip rate: share of all applications (%)*



Notes: The Figure plots the complier distribution over the joint quality × filed-scope plane, from the estimator of Section 5.3 on the 8 × 8 grid of equal-mass marginal quantile bins of Section 5.4, with the instrument demeaned within cell × four-by-four block of the plane. Both panels are multiplied by one factor so that the grant-weighted average of Panel (a) equals the aggregate share of 7.1%. Panel (a) reports the share of compliers among granted patents, Equation (11), and Panel (b) the flip rate, Equation (12), the share of all applications in the cell whose grant status depends on the examiner assignment in either direction. Both axes are in-sample percentile ranks. The low-quality × broad-scope cell is outlined and annotated with its standard error, clustered on the art unit × year cell. Cells with fewer than 200 applications are masked. None are, since the smallest cell holds 74,726 applications. Across the sixty-four cells, the flip rate is higher where the cell grant rate is lower (rank correlation −0.85), as part (ii) of Proposition 1 predicts for cells whose grant probability lies above one half.

The complier surface. Panel D.5a plots the two dimensional complier surface. Holding quality fixed, the share of compliers among granted patents is higher for the broadest than for the narrowest filings in every quality band, by 2.8 to 3.8 percentage points. In the lowest quality band it rises from 6.5% to 9.3%. Holding filed scope fixed, the share declines with quality: the highest quality band lies below the lowest in all eight scope bands.

Panel D.5b plots the flip rate on the two dimensional space. The figure highlights that concentration is sharper among granted patents than among applications. The flip rate of Equation (12), the share of all applications whose grant status depends on the examiner assignment, is lowest in the cell combining the highest quality with the narrowest claims, at 7.8%, highest in the cell combining the lowest quality with the broadest, at 11.3%, and ranges from 7.8% to 11.5% across the sixty-four cells, while the share among granted patents ranges from 5.0% to 9.3%. The mass of applications at the grant margin therefore varies less across the plane than the share it represents among granted patents, because the cell grant rate falls from about 77% in the narrowest filings to about 60% in the broadest. Section 6.3 shows that this is the signature of the threshold model. The numerator and the denominator of the inverse Mills ratio move differently, and over the range

of acceptance indices the data cover, it is the denominator that moves more: along filed scope, a little over half of the log change in the share among granted patents is the grant rate, about a third the density at the margin, and the rest the location of the leniency dose.

E Robustness: many-examiner instruments and the UJIVE estimator

The complier shares of Section 6 rest on the estimator of Sections 5.3 and 5.4. This appendix gives its sample form, its implementation and the derivation of its objects, and validates it on simulated data with a known target. It then describes the pass-through estimator, which conditions on examiner experience and which the robustness checks of the later appendices hold fixed, and reports two robustness exercises: a semiparametric partially linear estimator and a many-instrument comparison that measures the overfit of the examiner-dummy first stage.

Sample form of the estimator. Let $g(p)$ be the conditioning group of application p: its cell $\times$ ventile of the characteristic for the one-dimensional profiles, and its cell $\times$ block of the four-by-four partition of the quality $\times$ filed-scope plane for the joint profile. U_p^c is the instrument demeaned within the cell and $\tilde{U}_p$ the instrument demeaned within $g(p)$. The sample objects are

$$\hat{\psi}(B) = \frac{\sum_{p\in B} \tilde{U}_p D_p}{n_B\, \hat{V}(B)}, \qquad \hat{V}(B) = \sum_a w_{aB}\, \hat{V}_a, \qquad \hat{V}_a = \frac{1}{n_a} \sum_{p\in a} U_p^c D_p, \tag{22}$$

where n_B is the number of applications in B, a indexes art units, w_{aB} is the share of B's applications in art unit a, and $\hat{m}_B$ is the mean of U_p^c in B. Every within-group covariance carries the factor $n_g/(n_g-1)$ for a group of n_g applications, which makes it unbiased in groups of a few dozen applications, the typical size of a cell $\times$ ventile. A group with one application or one examiner-year carries no comparison and drops out of the numerators, while its applications still count in the grant rate. $\hat{V}_a$ is the variance of leniency in art unit a. Art units with fewer than 2,000 applications are pooled within their technology center, and centers still below that size form one unit. Under normality of leniency within an art unit, the dose of a bin is

$$\hat{d}^0(B) = \sum_a w_{aB} \left[\hat{V}_a^{1/2}\, \varphi\!\left(\hat{m}_B / \hat{V}_a^{1/2}\right) + \hat{m}_B\, \Phi\!\left(\hat{m}_B / \hat{V}_a^{1/2}\right) \right],$$

and over all applications it is $\hat{T}^0 = (2\pi)^{-1/2} \sum_a (n_a/n)\, \hat{V}_a^{1/2}$. The lower bound of Section 5.3, $\underline{\hat{T}}$, is the within-cell covariance of $\mathbf{1}\{U_p^c > 0\}$ with D_p, and the upper bound is $\overline{\hat{T}} = n^{-1} \sum_p (U_p^c)_+$. The point estimate is $\hat{T} = (\underline{\hat{T}}\,\overline{\hat{T}})^{1/2}$. A profile multiplies $\hat{\psi}(B)\,\hat{d}^0(B)$ by the factor $\hat{c} = \hat{T} \big/ \sum_B (n_B/n)\, \hat{\psi}(B)\, \hat{d}^0(B)$, which sets its grant-weighted average to the aggregate share. The same factor scales the dose of refusals, $\hat{d}^0(B) - \hat{m}_B$, and the two-sided dose of the flip rate, $2\hat{d}^0(B) - \hat{m}_B$. The lower end of the bracket, $\underline{\hat{T}}_1$, replaces the sign of U_p^c by the sign of the career instrument, on the applications where both instruments are defined. If $\hat{T}$ falls below $\underline{\hat{T}}_1$, I report

the bracket $[\hat{\underline{T}}_1, \hat{\overline{T}}]$ and no point. The shares replace the objects of Equations (11) and (12) by these estimates and by the grant rate of the applications in B.

Implementation: smoothing, resolution, and inference. The complier profiles of Section 5.4 are reported in two resolutions. The binned version uses twenty equal-mass bins of the percentile axis, the ventiles that also define the conditioning groups. The smoothed version replaces the bin indicator in Equation (22) by a Gaussian kernel weight $K_h(X_p - x)$ of bandwidth $h = 0.07$ over the unit percentile interval, comparable to a natural cubic spline with seven degrees of freedom. The kernel weight enters the numerator, the mass n_B, the art-unit shares w_{aB}, the location $\hat{m}_B$ and the grant rate, and the instrument stays demeaned within cell $\times$ ventile. The curve is evaluated at one hundred points from the first to the 99th percentile. Numerator and mass share the kernel, so the leading smoothing bias common to both cancels. The ratio keeps an $O(h^2)$ term in the curvature of the two, which survives where they curve differently, so the reported bands cover the smoothed profile rather than the underlying one and are correctly centered only to that order. Results are reported for $h \in \{0.05, 0.07, 0.10\}$, which bounds the practical importance of that term without eliminating it, and the outermost percentiles are kernel-edge regions and are trimmed. The joint profile masks cells of the plane with fewer than 200 applications. Standard errors come from the joint influence function of every estimated component: the numerators, the kernel masses and grant rates, the art-unit variances $\hat{V}_a$, the locations $\hat{m}_B$ and the two bounds. They are clustered on the art unit $\times$ year cell. The bands hold the scaling factor of each profile fixed, which leaves every studentized statistic and every test of the shape unchanged. The instrument is treated as given. Applications that share an examiner's leave-out estimate across cells are dependent, and clustering on the cell does not capture that dependence. Because the profiles are read as curves, I also report simultaneous bands: a Gaussian multiplier bootstrap with 2,000 draws gives a studentized sup-t critical value and, from the same draws, a joint test of the null that the curve is flat.

Derivation of the complier objects. Fix a cell c and write $\bar{D}_c$ for the pool's average grant rate on the cell's applications, so that an examiner with leniency ℓ_e grants a share $\bar{D}_c + \ell_e$ of them and the reference examiner a share $\bar{D}_c$. The reference examiner need not be one of the cell's examiners, and monotonicity is assumed for the leniency order extended to it. For $\ell_e > 0$, monotonicity gives $D_p(\ell_e) \geq D_p(0)$ for every application, so the difference of the two grant rates, ℓ_e, is the probability of the event $D_p(\ell_e) = 1$, $D_p(0) = 0$. For $\ell_e \leq 0$ the event cannot occur. Random assignment makes the assigned examiner independent of the potential outcomes within the cell, and averaging over assignments and cells gives Equation (8). The reverse event has mass $\mathbb{E}[(\ell_{e(p)})_-]$, which equals $\mathbb{E}[(\ell_{e(p)})_+]$ because ℓ has mean zero within every cell, and the two masses together are $\mathbb{E}\,|\ell_{e(p)}|$.

Within a set B, assumption (iii) writes the grant probability of an application in B before

an examiner of leniency ℓ as $\alpha_{cB} + \psi(B)\,\ell$. The reference examiner's grant rate in B is then α_{cB}, and the same argument, applied to the applications in B, gives Equation (9), and its two-sided version Equation (12). The instrument recovers both components. The leave-out noise of U_p is independent of D_p given the assigned examiner, and each examiner's grant rate is $\bar{D}_c + \ell_e$ by the definition of ℓ, so $\mathbb{E}[U_p^c D_p \mid c] = \mathrm{Var}(\ell_{e(p)} \mid c)$ without assumption (iii). Within a conditioning group, $\mathbb{E}[\tilde{U}_p D_p \mid g] = \psi(B)\,\mathrm{Var}(\ell_{e(p)} \mid g)$ under (iii). Assumption (iv) makes the group's variance the same fraction of the cell's in every bin, so the ratio with the cell's variance is $\psi(B)$ up to a factor common to all bins, which the scaling to the aggregate removes.

Without assumption (iii), the ratio estimates a weighted average, over the critical leniencies v in the range of the pool, of the density of ℓ_p^* at v among the applications in B. The weight of v is proportional to $\mathbb{E}[\ell_{e(p)}\,\mathbf{1}\{\ell_{e(p)} \geq v\}]$, the leniency deviation of the examiners beyond v, which is positive on both sides of the pool mean. $T(B)$ integrates the same density with the weight $\Pr(\ell_{e(p)} \geq v)$ over $v > 0$, the lenient side alone. The two agree when the density is constant over the range of the pool, which is assumption (iii) and which the arm check of Section 5.4 tests.

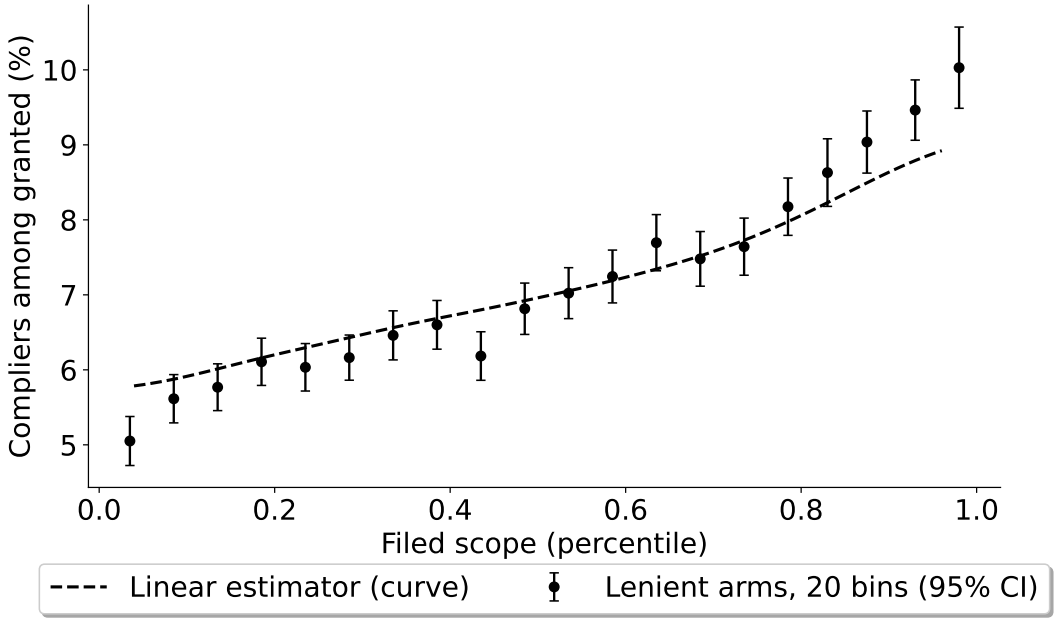
For the bounds, write $U_p = \ell_{e(p)} + \varepsilon_p$ with noise of mean zero given the assigned examiner. The within-cell covariance of $\mathbf{1}\{U_p > 0\}$ with D_p equals $\mathbb{E}[\ell\,\mathbf{1}\{U > 0\}]$, which is at most $\mathbb{E}[\ell\,\mathbf{1}\{\ell > 0\}] = \mathbb{E}[(\ell)_+]$, since the indicator of $\ell > 0$ maximizes $\mathbb{E}[\ell\, a]$ over all $a \in [0, 1]$. Jensen's inequality gives $\mathbb{E}[(U)_+] = \mathbb{E}\left[\mathbb{E}[(\ell + \varepsilon)_+ \mid \ell]\right] \geq \mathbb{E}[(\ell)_+]$, for the same reason that estimation noise raises the variance of estimated firm effects (Kline, Saggio, and Sølvsten, 2020). If ℓ and ε are jointly normal, as in parametric empirical Bayes (Morris, 1983), with reliability $\lambda = \mathrm{Var}(\ell)/\mathrm{Var}(U)$, which is also the factor by which empirical Bayes would shrink U, the lower bound is $\lambda\,\mathbb{E}[(U)_+] = \sqrt{\lambda}\,\mathbb{E}[(\ell)_+]$ and the upper bound is $\mathbb{E}[(\ell)_+]/\sqrt{\lambda}$, so their geometric mean is $\mathbb{E}[(\ell)_+]$.

Validation on simulated data. I validate the estimator on simulated applications with a known target. Each simulated art unit has cells of publication years and a pool of examiners whose leniency drifts across years, and a small share of examiners decides in two art units, which connects the design as in the data. Both instruments are built with the code that builds them in the data, and the target objects are computed from the potential outcomes. On a design close to linearity, the estimator recovers the twenty-bin shares within simulation error. With curvature planted in the grant probability, the arm check rejects and the lenient-arm estimator moves toward the target. Under that curvature, the aggregate still equals the target, as Equation (8) implies, and the bracket contains the target in every replication up to sampling error. An end-to-end run on fifteen art units of the data reproduces the instrument job's own sums exactly. The simulation also showed that the normal dose alone understates the aggregate in every replication, by 2% to 5% on average depending on the size of the pools. A cell's leniencies are a small discrete set, and the grant-probability scale compresses their tails. The geometric mean of the bounds errs in both

directions, by about one percent or less on average in each simulated design, and the career-sorted lower bound is valid in every replication and tighter than the year-sorted one. On this evidence, the geometric mean is the point estimate of Section 5.3. The choice was made on the simulation, before the estimator was run on the data.

The lenient-arm profile along filed scope. Along filed scope, the arm check of Section 5.4 rejects linearity. The sensitivities estimated from the three most lenient and the three strictest arms differ jointly (Wald statistic 22.2, $p = 0.0005$), and the lenient profile differs from the linear one by up to 5.8% in a quintile. Figure E.6 reports the profile estimated from the lenient arms alone, which counts the treated compliers the way $T(B)$ does. In the twenty bins it rises from 5.6% in the four narrowest bins to 9.3% in the four broadest, by about 65%, against about 47% for the same bins of the linear estimator. The two estimators agree on the direction of the profile and differ in how steeply it rises. The main text reports the linear curve throughout.

Figure E.6: *Compliers among granted patents by filed scope, lenient arms*



Notes: Dots show the share of compliers among granted patents in twenty equal-mass bins of filed scope, with the sensitivity estimated from the three most lenient of five within-cell leniency arms. Bars are 95% confidence intervals from a bootstrap over cells. The dose and the grant rate are those of the linear estimator. The dashed line is the linear curve of Panel 4b. Both are multiplied by the factor that scales the linear profile to the aggregate share.

Dispersion of leniency within bins. Equation (22) divides the covariance of the instrument with the grant within cell × bin by the variance of leniency in the art unit. The shape of a profile therefore requires assumption (iv): within a cell, the variance of leniency among the applications of a bin is the same fraction of its variance in the cell for every bin. A fraction common to all bins

drops out when each profile is scaled to the aggregate share, so only differences across bins matter. To test for them, I split each examiner × year at random into three folds, rebuild the instrument on each fold, and predict it for the applications of the other two folds. Every application thus receives two measures of its examiner's leniency from fits that share no rows. Because their estimation noise is independent, the mean product of the two measures within a bin estimates the variance of leniency in that bin. At the level of the art unit, the same product reproduces the variance that the estimator uses to within 0.3%. Dividing the product in each bin by its value in the art unit gives a ratio for each of the twenty bins.

Along filed scope, the ratio is the same in every bin (Wald statistic 15.0 with 19 degrees of freedom, $p = 0.72$). Correcting each bin by its own ratio moves no twenty-bin share by more than 1.1%, and the rise between the four narrowest and the four broadest bins is 47.1% instead of 47.4%. Along quality, the test rejects equality (Wald statistic 82.9, $p < 0.001$). The differences are small but precisely estimated. The ratio averages 0.94. Relative to that mean, it is 0.98 in the lowest-quality bin, about 1.02 in the middle bins and 0.96 in the highest, with standard errors of about 0.01. Correcting each bin by its own ratio raises the share by 0.4% in the four lowest-quality bins and by 2.8% in the four highest. Between them, the share then falls from 7.25% to 6.77%, by 6.6%, against a fall from 7.23% to 6.59%, by 8.8%, without the correction. The correction thus flattens the quality profile and leaves the filed-scope profile unchanged. The main text reports the uncorrected profiles. Along quality, the uncorrected profile is the steeper of the two. With or without the correction, the share rises steeply in filed scope and falls slightly in quality.

The pass-through estimator. The robustness checks of the later appendices hold fixed an estimator that differs from that of Section 5.3 in three ways. First, its controls W also contain a linear term in examiner experience, the years since the examiner's first observed decision top-coded at five, and an indicator for examiners already active in the first year of the data. Its reference examiner is therefore the average examiner of the same experience. Second, its instrument U is built from examiner indicators that pool each examiner's years. Third, its level comes from a pooled regression on the deviation $\Delta l_p = l_p - \tilde{l}_p$ of realized from expected leniency of Section 3,

$$D_{p_{kt}} \;=\; \alpha_{kt} \;+\; \bar{\psi}\,\Delta l_{p_{kt}} \;+\; X_{p_{kt}}\Theta \;+\; \upsilon_{p_{kt}}, \tag{23}$$

and its shape from the ratio

$$\hat{\theta}(B) \;=\; \frac{\sum_p U_p\,\mathbf{1}[X_p \in B]\,D_p}{\sum_p U_p D_p}, \tag{24}$$

the share of the instrument's first-stage covariance that falls in B. Dividing $\hat{\theta}(B)$ by the population share of B gives the representation ratio $\rho(B)$, and the share among granted patents is $\bar{\psi}\,\rho(B)\,\bar{d}/P[D=1 \mid X \in B]$ with the common dose $\bar{d} = \mathbb{E}[(\Delta l_p)_+]$. Averaged over the population it is

$$\text{complier share} \;=\; \frac{\bar{\psi}\cdot\bar{d}}{\Pr(D_p=1)}. \tag{25}$$

The ratio $\hat{\theta}(B)$ divides by the covariance over all applications rather than taking the covariance within the bin, so a bin whose applications are assigned more lenient examiners on average is credited with compliers. The factorization takes the dose to be the same in every bin, which Appendix F tests. The deviation Δl_p averages a finite number of decisions, so the least-squares $\hat{\bar{\psi}} = 0.69$ is attenuated, and instrumenting Δl_p with the examiner indicators raises it to 1.11. With the dose in measured units, the two give aggregate shares of 4.9% and 7.8%.

Semiparametric partially linear estimator. Following the residualization approach of Robinson (1988), I also estimate the conditional first-stage sensitivity $\psi(q)$ from the partially linear system

$$D_{p_{kt}} = \alpha_{kt} + \psi(q)\, l_{p_{kt}} + g(q_{p_{kt}}, s_{p_{kt}}, X_{p_{kt}}) + \upsilon_{p_{kt}} \tag{26}$$

$$l_{p_{kt}} = \gamma_{kt} + m(q_{p_{kt}}, s_{p_{kt}}, X_{p_{kt}}) + \eta_{p_{kt}}, \tag{27}$$

where α and γ are art unit $\times$ year fixed effects and the nuisance functions $g(\cdot)$ and $m(\cdot)$ absorb the nonlinear effects of quality, filed scope, and covariates on both grant and leniency. I estimate the nuisances by random forests with cross-fitting (double machine learning), and identify the varying coefficient $\psi(q)$ as the best linear projection of the cross-fitted signal on a tensor of natural cubic splines in the quality and scope percentiles. This estimator collapses the examiner dummies into the leniency deviation $\Delta l_p = l_p - \tilde{l}_p$ and relaxes the linear-control structure of the pass-through estimator.

Along **filed scope** the two estimators agree: the complier share rises steeply across the filed-scope distribution under both. Along **quality** they do not. The modified-outcome estimator stays close to its mean, within half a percentage point. The partially linear estimator declines. Because this disagreement bears directly on the quality profile, I traced it. It is not binning (the kernel-smoothed modified outcome, matched to the same effective degrees of freedom, reproduces the modified-outcome shape), not the instrument set (the scalar leniency deviation run through the modified outcome agrees with the examiner-dummy version), not the control set (re-residualizing the same instrument on the double machine learning conditioning set leaves the divergence intact), and not the flexibility of the learner (replacing the random forest with linear nuisances leaves

the partially linear profile essentially unchanged). What remains is the estimand itself. The partially linear conditional effect is a variance-weighted projection, proportional to $\mathrm{Cov}(\Delta l, D \mid q)/\mathrm{Var}(\Delta l \mid q)$, whereas the complier mass is the covariance alone. The complier mass is the object of Section 5.3. The estimator of the main text takes the same covariance within cell $\times$ bin and divides it by the variance of leniency in the art units of the bin, which, unlike $\mathrm{Var}(\Delta l \mid q)$, contains no measurement noise of the instrument. Once the instrument imbalance is removed, the modified-outcome profile also declines mildly, in the direction of the partially linear estimator. Nor does the choice between a common and a conditional dose reopen the question: reweighting the profile point by point by the conditional favorable dose of Appendix F moves it by at most four percent anywhere, so the dose-weighted profile tracks the pass-through profile.

Many-instrument robustness. Both instruments are leave-out fits on thousands of examiner indicators. With instruments this numerous, in-sample fitted values are overfit, and the first stage is biased toward the ordinary least-squares association (Bound, Jaeger, and Baker, 1995; Angrist, Imbens, and Krueger, 1999). I measure the overfit for the examiner indicators of the pass-through estimator by comparing estimators that differ only in how they handle many-instrument and own-observation bias.

Estimators. Let D_p be the grant indicator, Z the matrix of examiner dummies, and W the controls (art unit $\times$ year fixed effects and the covariates of the pass-through estimator). Two-stage least squares (2SLS) projects D on $[Z, W]$ in sample. With many instruments its fitted first stage is overfit and biased toward the OLS association (Bound et al., 1995). The jackknife instrumental variables estimator (JIVE; Angrist et al., 1999) replaces each fitted value with its leave-one-out analogue, removing the own-observation component. The unbiased jackknife estimator (UJIVE; Kolesár, 2013) additionally leaves the application out of the *control* projection, making it consistent under many-instrument and many-control asymptotics, the relevant regime here, where the art-unit $\times$ year fixed effects are themselves numerous. The leave-out fitted values require the leverages of the joint examiner-by-cell design, which is a bipartite two-way fixed-effects problem with no closed form. I compute them by the Johnson-Lindenstrauss random-projection estimator used by Kline et al. (2020), restricting the sample to the largest leave-out- connected component of the examiner-cell graph (which retains all but two observations).

Result. On the 2005–2020 core sample (n = 5,364,969, 9,701 art-unit $\times$ year cells, 12,705 examiners and 22,416 design parameters), the examiner dummies are jointly strong (joint first-stage $F \approx 30.3$). The mean leverage is only $K/n \approx 0.004$, so the own-observation and many-instrument distortions are small. Table E.2 reports the result: the honest leave-out (UJIVE) first stage re-

Table E.2: *Many-instrument robustness of the first stage*

	2SLS	JIVE	UJIVE
First-stage signal (relative to 2SLS)	1.000	0.940	0.966

Notes: The table reports the first-stage signal $\sum_i \hat{D}_{-i} D_p$, the (leave-out) covariance of the fitted treatment with the realized grant, for the examiner-dummy instrument set, normalized to the in-sample 2SLS value. 2SLS uses in-sample fitted values. JIVE (Angrist et al., 1999) and UJIVE (Kolesár, 2013) use leave-one-out fitted values, with UJIVE additionally jackknifing the control projection. Leverages are computed by Johnson-Lindenstrauss random projection (Kline et al., 2020). Sample: publication years 2005–2020; $n = 5,364,969$; 12,705 examiner instruments; 9,701 art-unit × year cells; mean leverage $K/n = 0.004$; joint first-stage $F = 30.3$. Absolute first-stage signal: 2SLS 63,788, JIVE 59,941, UJIVE 61,647. Entity status as recorded in the PatEx snapshot for every application. Controls W as in the pass-through estimator.

covers 96.6% of the in-sample 2SLS explained variation in grant, and JIVE 94.0%.[34] The implied overfit of the examiner-dummy first-stage projection is thus at most three to four percent, so the design is not materially contaminated by the many-instrument problem at this sample size. The ratio holds the instrument set fixed and compares in-sample with leave-out fits, so it diagnoses that projection and not the leniency deviation of Equation (23).

The level of the complier share. The two levels of the pass-through estimator bracket the share in true units in the same way as the bounds of Section 5.3. The noise in Δl_p scales the least-squares pass-through by the reliability λ of the deviation and, under normality, the measured dose by $1/\sqrt{\lambda}$, so the lower level equals $\sqrt{\lambda}$ times the share in true units and the upper level $1/\sqrt{\lambda}$ times it. Their geometric mean, 6.2%, is the counterpart of the point estimate of Section 5.3. That estimate measures the dose in grant-probability units from the instrument itself, so it needs neither the pass-through regression nor an estimate of λ from the ratio of two coefficients. The point estimate of 7.1% is 5.9 times the 1.2% benchmark of de Rassenfosse et al. (2020). Both numbers measure inconsistency in the application of one office's own standard, but they are not the same estimand, and the gap has a construction-based reading rather than a substantive one. Theirs is a rate of *detected* disagreement. A granted US patent enters their numerator only if a twin application was refused at another office, so the rate is bounded above by the share of families with mixed grant–refusal outcomes, 25.2% of US grants in their sample, and a marginal US grant whose twins were all granted contributes nothing to it by construction. My estimate counts grants that flip with the identity of the examiner assigned inside an art unit, which does not require the outcome to be observed a second time. I therefore read their figure as a floor on the same phenomenon rather than as a competing estimate of it. I do not rest this on sample selection: their

[34] JIVE lies below UJIVE because, unlike UJIVE, it does not leave the application out of the high-dimensional control projection and therefore over-corrects. UJIVE is the appropriate benchmark with many controls (Kolesár, 2013).

twin applications are of higher value than the average filing, but the authors' own comparison of the estimated inventive-merit distributions across offices finds selection to be immaterial, and they argue that population rates are unlikely to be much higher than their sample estimates.

The mirror image: compliers among rejected applications. Equation (8) counts the applications granted only because they were assigned an examiner more lenient than the reference examiner. Leniency has mean zero within every cell, so the mass of applications refused only because they were assigned a stricter examiner is the same, $\mathbb{E}[(\ell_{e(p)})_-] = \mathbb{E}[(\ell_{e(p)})_+]$. Divided by the granted pool, the mass gives the share of Section 6.1, and divided by the refused pool, $1 - \Pr(D_p = 1) = 0.289$, it gives 17.4%. Roughly one rejection in six is owed to the examiner assignment, against one grant in fourteen. The two figures are one fact reported over two denominators, not two effects of different size.

For an individual application, neither share is the relevant object. The relevant measure is the flip rate of Equation (12), the probability that the examiner assignment decides the outcome in either direction. It counts both the grant a lenient examiner creates and the rejection a strict one forces, which makes it roughly twice the treated-complier mass. Panel D.5b plots it over the quality–scope plane: it is 7.8% in the cell combining the highest quality with the narrowest claims and 11.3% in the cell combining the lowest quality with the broadest, against the 4.3 points separating the same cells in the share among granted patents. In the terms of Proposition 1, the larger part of the gradient in the composition is therefore the denominator $\Phi(t)$, and the smaller part is the density of applications at the margin. Table L.4 reports the split along filed scope: between the 10th and the 90th percentile of filed scope, the grant rate carries 54% of the log gradient, the density 34%, and the location of the leniency dose 12%.

F Instrument validity: monotonicity and complier balance

The examiner-leniency design requires that assignment be as-good-as-random within art unit × year (addressed in Section 5), that leniency affect outcomes only through the grant decision (exclusion), and that no application is granted by a stricter examiner but rejected by a more lenient one (monotonicity, or "no defiers"). The joint test of exclusion and monotonicity of Frandsen et al. (2023) requires a second-stage outcome and is therefore reported for the Sampat and Williams (2019) and Farre-Mensa et al. (2020) subsamples in Appendix K (p-values 1.000 and 0.570). This appendix reports the tests of Section 5.5 for the instrument of Section 5.3, and then the diagnostics of the pass-through estimator of Appendix E, which the robustness appendices hold fixed.

Monotonicity. Under monotonicity, the first-stage effect of leniency on grant must be non-negative in every subpopulation. A negative first stage in a well-populated subgroup would signal defiers. I provide three pieces of evidence. First, re-estimating the first stage of the pass-through estimator within subgroups, the sensitivity of grant to the recentered leniency deviation is positive across all sixteen publication years in the sample (range 0.66–0.89) and across all one hundred quality percentiles (range 0.60–0.70). It is non-negative in 94.2% of the 657 art units, with 38 negative point estimates, and the 5th percentile across art units is −0.03. Table F.1 summarizes. Each subgroup enters with the median of its estimate across five runs with different seeds, since the random-forest nuisances make a single run depend on its seed. The two extremes of the art-unit column, −2.95 and 45.8, are estimated on eight and five applications and do not differ significantly from zero. Of the 38 negative estimates, 20 are individually significant at the five percent level. Art units are disjoint subsamples, so the 632 studentized slopes with non-degenerate standard errors are independent. The count of significant negatives is then close to the 15.8 that chance would produce were every true slope exactly zero, neither significantly fewer ($p = 0.88$) nor significantly more ($p = 0.17$). The minimum studentized slope is −7.0, against an expected minimum of −3.1 across 632 independent tests of an exact zero, so non-negativity in every unit is rejected ($p < 10^{-4}$). Four art units with between 1,465 and 12,429 applications have studentized slopes below −5, and the twenty units with significant negative estimates hold 1.8% of the applications, so in a small number of art units the first stage of the pass-through instrument runs the wrong way. What the design requires is the average-monotonicity condition, whose test follows, not unit-level exactness. The scalar-type diagnostic below carries the same reading: monotonicity is an approximation holding almost everywhere, not an exact restriction. Art units enter this table only if the subgroup estimator returns a non-degenerate fit, which is why 657 of the 755 units in the sample appear. Second, I test the implication that the complier mass is non-negative in each of ten equal-mass bins (Goldsmith-Pinkham et al., 2026). With the instrument of Section 5.3, the

smallest studentized bin mass is 38.7 on quality and 66.0 on filed scope, against critical values of 2.74 and 2.85, and the joint test gives $p = 1.00$ on both axes, as Table F.3 shows. The career instrument and the instrument that controls for examiner experience give the same result. Third, I test the scalar-type assumption directly. A two-dimensional examiner, lenient on quality but strict on scope, would generate defiers along filed scope even where the average condition holds within coarse subgroups. Within each art unit $\times$ period cell, I split every examiner's caseload at the cell median of filed scope (average words per claim, where broad means fewer words), compute the examiner's grant rate separately on the two halves, and compare the within-cell ranking of examiners implied by the broad half against that implied by the narrow half. The benchmark is a placebo that randomly re-halves each examiner's caseload at exactly the scope split's half sizes, so the sampling-noise structure of the two grant-rate estimates is identical by construction and the placebo correlation is the agreement that noise alone permits. Table F.2 reports both correlations under four-year and two-year period definitions. The rankings agree at 96% of the noise ceiling, and the gap of 0.03 does not move when the period is halved, which points away from within-cell examiner drift as its dominant source without proving the residual is zero. The difference is precisely estimated ($z = -12.8$, clustered on art unit), so an exactly scalar type is rejected at this sample size. Reading the magnitude as an upper bound on any scope-specific dimension rests on a sign assumption: residual within-cell drift, and any other component shared across the split, loads on the scope split and not on the placebo, inflating the gap rather than masking it. Under that assumption the 0.03 is a ceiling. Without the assumption, it is an estimate. What the design requires is that defiers along filed scope be rare, and a ranking displacement of three to four percent of the noise ceiling is the scalar approximation holding to within a few percent.

Table F.1: *Monotonicity: pass-through first stage across subgroups*

Subgroup partition	# subgroups	Min	Median	Max	Share ≥ 0
Publication year	16	0.66	0.71	0.89	100.0%
Quality percentile	100	0.60	0.64	0.70	100.0%
Art unit	657	−2.95	0.58	45.79	94.2%

Notes: Sensitivity of the grant to the recentered leniency deviation Δl_p, re-estimated within each subgroup by the partially linear model of Appendix E with random-forest nuisances. Each subgroup enters with the median of its coefficient across five runs with different seeds. Under monotonicity the coefficient is non-negative in every subgroup. Across the five single runs, no publication-year or quality-percentile entry moves by more than 0.01. The art-unit minimum ranges from −3.27 to −1.84, the maximum from 42.69 to 48.29, the share of non-negative estimates from 93.8% to 94.2%, the number of negative estimates from 38 to 41 and the number of significant negatives from 16 to 20. The art-unit minimum and maximum are estimated on eight and five applications. Partitions that condition on the instrument itself, such as the leniency percentile, leave no residual variation and are uninformative for this check.

Table F.2: *Scalar leniency: examiner rankings across filed scope*

	4-year cells	2-year cells
Rank correlation, scope split	0.756	0.680
Rank correlation, placebo split (noise ceiling)	0.782	0.710
Agreement ratio	0.966	0.959
Difference (s.e.)	−0.027 (0.002)	−0.029 (0.002)
z	−12.8	−12.8
Cells	1,861	2,864
Art units	577	557
Examiner-cells	30,893	41,398

Notes: Within each art unit × period cell, every examiner's decided applications are split at the cell median of filed scope (average words per claim, where broad means fewer words), and the examiner's grant rate is computed separately on each half. Reported are the caseload-weighted means of the within-cell Spearman correlations between the examiner rankings implied by the two halves. The placebo split randomly re-halves each examiner's caseload at the same half sizes (twenty draws), so its correlation is the ranking agreement that sampling noise alone permits. Examiners qualify with at least twenty decisions per half and a non-degenerate overall grant rate. Cells qualify with at least eight qualifying examiners. Weights are qualifying examiners per cell. The standard error of the difference is clustered on the art unit. Sample: publication years 2005–2020, 5,367,642 applications.

Table F.3: *Monotonicity: non-negative complier mass by bin*

	Quality			Filed scope		
Instrument	Min. t	Crit.	p	Min. t	Crit.	p
Examiner × publication year, headline	38.7	2.74	1.00	66.0	2.85	1.00
Examiner, whole career	36.9	2.70	1.00	62.4	2.76	1.00
Examiner, experience in the controls	37.3	2.67	1.00	58.0	2.81	1.00

Notes: Test of Goldsmith-Pinkham et al. (2026) that the complier mass is non-negative in each of ten equal-mass bins of the characteristic. *Min. t* is the smallest studentized bin mass, *Crit.* the one-sided simultaneous critical value from 2,000 multiplier draws, and p the p-value of the joint test. Instruments as in Table F.5.

Complier characteristics. Under the estimator of Section 5.3, the treated compliers have a mean quality percentile of 0.50 against a population mean of 0.51, and a mean filed-scope percentile of 0.53 against 0.51. Table F.4 reports the complier means of the pass-through estimator from the modified-outcome estimator of Goldsmith-Pinkham et al. (2026), which pools the means computed from the granted and from the refused applications. It gives the same two means and adds a predetermined firm characteristic, whether the applicant is publicly listed: 34.5% of compliers are listed firms against 28.7% of all granted applications ($z = 17.7$).

Table F.4: *Complier characteristics of the pass-through estimator: means versus the reference population*

Characteristic	Reference	Compliers	Gap (z)
Invention quality (mean)	0.51 (pop.)	0.50	−4.5
Filed scope (mean)	0.51 (pop.)	0.53	19.1
Publicly listed (share)	0.287 (granted)	0.345	17.7

Notes: Complier means via the modified-outcome estimator of Goldsmith-Pinkham et al. (2026), with the examiner-dummy UJIVE instrument of the pass-through estimator. Each mean is the efficient pooling of the treated- and untreated-arm estimates of the same complier mean. The statistic z is the studentized pooled-mean-minus-reference gap, with standard errors clustered on art unit × publication year. The reference column is the population mean for quality and scope and the granted-sample mean for public listing. Quality and scope are the paper's central dimensions, the subject of Section 6. Public listing is a predetermined firm characteristic, defined only among granted applications, so its row is the treated-only coverage estimand on the same UJIVE variation. No untreated arm exists to pool for it. The quality gap is precisely estimated and small, 0.6 percentile points.

Balance. Table F.5 reports the balance test of Section 5.5 for three instruments. Under random assignment, the mean characteristic of the compliers is the same whether it is computed from the granted or from the refused applications. With the instrument of Section 5.3, the two means differ by 1.9 percentiles of quality and by 1.5 percentiles of filed scope, with z-statistics of −17.9 and −14.3. Pooling each examiner's years reduces the gaps to 0.5 and 1.1 percentiles. With examiner experience in the controls, the quality gap is +1.4 percentiles and the scope gap −2.0. Table F.6 tests assumption (i) within cell × bin on three characteristics fixed at filing. The correlations with the instrument are at most 0.012 in absolute value, and residualizing the instrument on all three moves no twenty-bin share by more than 0.22% and the aggregate share by 0.03%.

Size of the imbalance. The z-statistics of Table F.5 are large because the sample is large. Table F.7 measures the gaps in standard deviations of the characteristic. With the instrument of Section 5.3, the quality gap is 0.066 standard deviations and the scope gap 0.051. In their reanalysis of Farre-Mensa et al. (2020), Goldsmith-Pinkham et al. (2026) find a gap of 0.109 standard deviations in one of three patent-class indicators and read their results as support for quasi-random assignment. Per application, the two tests are about equally precise. The standard error times the square root of the number of applications is 8.5 and 8.4 standard deviations here and between 7.6 and 8.3 in their reanalysis. The tests differ in the number of applications. With the 32,514 applications of the reanalysis, the test here would detect the quality gap with probability 0.29 and the scope gap with probability 0.20. Detecting them with probability 0.8 requires about 130,000 and 211,000 applications.

The gaps also differ across cells. Clustering the standard errors on the art unit × year cell multiplies them by 5.2 for quality and by 1.7 for scope. The z-statistics fall to −3.5 and −8.1, and detecting the gaps with probability 0.8 requires about 3.5 million and 646,000 applications. The quality gap is thus an average over cells whose gaps differ in size and sign, not a small gap

common to every cell.

Measured against the complier means of the same fit, the quality gap is 2.1 times the difference between the mean quality of the compliers and that of the population. The scope gap is 0.58 times the corresponding difference in filed scope. For the three class indicators of the reanalysis, the same ratio is 1.1, 2.2 and 0.64. The balance test therefore cannot separate the mean quality of the compliers from the population's, whereas the mean filed scope of the compliers exceeds the population's by more than the gap.

The imbalance in a characteristic enters its own profile in one way only. The sensitivity is estimated within cell × bin, which removes the imbalance in the profile's own characteristic. The imbalance therefore enters through the dose, as the mean leniency of the examiners assigned to the applications at a given value of the characteristic. Sorting on the other characteristic moves no profile by more than 0.05%. Panel C of Table F.7 sets that mean to zero and scales each profile to the same aggregate share. The share of compliers among granted patents then rises from 5.85% to 8.76% between the 10th and the 90th percentile of filed scope, by 50% instead of 46%. It falls from 7.09% to 6.76% across quality, by 4.6% instead of 7.6%. Under the definition of compliers in Section 5.3, the term belongs to the estimand, since an application of a kind that tends to be assigned stricter examiners is less often a complier. With or without it, the share rises steeply in filed scope and falls slightly in quality.

Table F.5: *Balance of the instrument against quality and filed scope*

	Quality			Filed scope		
Instrument	Gap	z	Bins	Gap	z	Bins
Examiner × publication year, headline	−1.88	−17.9	9	−1.50	−14.3	10
Examiner, whole career	−0.51	−4.7	6	−1.07	−9.8	7
Examiner, experience in the controls	+1.44	+12.1	13	−1.98	−16.6	12

Notes: Balance test of Goldsmith-Pinkham et al. (2026). *Gap* is the mean percentile of the compliers computed from the granted applications minus the mean computed from the refused applications, in percentile points, which is zero under random assignment. z uses heteroskedasticity-robust standard errors. *Bins* counts the twenty equal-mass bins of the characteristic in which the same difference is significant at the five percent level. Each instrument is a leave-out UJIVE fit with art unit × year fixed effects, small and micro entity status and the type of continuation as controls, the last row adding examiner experience. All rows use the same 5,357,836 applications.

Table F.6: *Balance within cell × bin on characteristics fixed at filing*

	Within cell × quality bin			Within cell × scope bin		
Characteristic fixed at filing	Corr.	z	Bins	Corr.	z	Bins
Number of claims, log	+0.0072	+10.5	19	+0.0072	+10.3	18
Length of the first claim, log	−0.0105	−16.2	20	−0.0123	−20.0	19
Month of publication	−0.0074	−9.2	19	−0.0100	−11.7	19
Largest change of a 20-bin share (%)		0.05			0.22	
Change of the aggregate share (%)			−0.03			

Notes: *Corr.* is the correlation of the instrument of Section 5.3, demeaned within art unit × year cell and twenty equal-mass bins of the characteristic on the axis, with the characteristic in the first column, also demeaned within cell × bin. z clusters on the cell. *Bins* counts the twenty bins in which the correlation is significant at the five percent level. The last two rows residualise the instrument on all three characteristics within cell × bin and re-estimate: the largest change of a twenty-bin share among granted patents along each axis, and the change of the aggregate share, both in percent of the value.

Table F.7: *Size of the imbalance*

Panel A: size and precision of the balance test

	Applications	Gap (s.d.)	z	Detectable gap (s.d.)	s.e. $\times\sqrt{N}$ (s.d.)	Gap / complier difference
This paper						
Quality	5,360,465	−0.066	−17.98	0.010	8.47	2.06
Filed scope	5,360,465	−0.051	−14.12	0.010	8.42	0.58
Farre-Mensa et al., reanalysed by Goldsmith-Pinkham et al.						
Patent class group I	32,514	−0.109	−2.35	0.129	8.33	1.08
Patent class group II	32,514	+0.022	+0.52	0.120	7.70	2.20
Patent class group III	32,514	+0.064	+1.52	0.118	7.62	0.64

Panel B: power of the test in this paper

	Quality	Filed scope
z, robust standard errors	−17.98	−14.12
z, standard errors clustered on art unit × year	−3.49	−8.07
Applications for 80% power, robust	130,000	211,000
Applications for 80% power, clustered	3,455,000	646,000
Power with 32,514 applications, robust	0.29	0.20

Panel C: compliers among granted patents (%)

	10th percentile	90th percentile	Change
Filed scope, as reported	5.91	8.64	+46.2%
Filed scope, location term set to zero	5.85	8.76	+49.7%
Quality, as reported	7.20	6.65	−7.6%
Quality, location term set to zero	7.09	6.76	−4.6%

Notes: Panel A expresses the balance gap of Table F.5 in standard deviations of the characteristic, for the instrument of Section 5.3, next to the same statistics for the reanalysis of Farre-Mensa et al. (2020) by Goldsmith-Pinkham et al. (2026), computed from their Tables 1, 2 and 4. All rows of Panel A use heteroskedasticity-robust standard errors. The smallest detectable gap is the gap that a two-sided test at the five percent level detects with probability 0.8, which is 2.80 standard errors. The standard error times the square root of the number of applications measures the precision of the test per application. The last column divides the gap by the difference between the pooled complier mean and the population mean of the same fit. Of the eight rows of their Table 2, the three class-group indicators are shown. The three rows in logs are left out because their Table 1 reports the standard deviation of the level only, and the two rows on European and Japanese approval cover only the 3,287 and 1,206 applications also filed abroad. Panel B assumes that the standard error falls with the square root of the number of applications. Clustered standard errors group the applications of each art unit × publication year cell. Panel C sets to zero the mean leniency of the examiners assigned to the applications at each point of the characteristic, which is the only route by which the imbalance in a characteristic enters its profile in Figure 4, and scales each profile so that its grant-weighted average equals the aggregate share of 7.06%. All panels use the 5,360,465 applications of the instrument. Table F.5 uses the 5,357,836 applications common to its three instruments, on which the gaps are −1.88 and −1.50 percentiles.

The pass-through estimator. The remainder of this appendix concerns the pass-through estimator of Appendix E, whose instrument is the recentered leniency deviation Δl and its examiner-dummy UJIVE counterpart. It shows why that instrument is imbalanced in filed scope, in which direction the imbalance moves its profiles, and how far its common dose departs from the conditional one.

Distributional balance and caseload composition in measured leniency. The pass-through estimator is identified from the *bin-conditional* mean of its instrument, so its balance is tested bin by bin. Let B index a bin of the filed-scope percentile and let $g_B = \mathbf{1}[s \in B]$ be the corresponding modified outcome. The treated and untreated arms of the estimator are

$$\hat{\theta}_T(B) = \frac{\overline{\Delta l \, g_B D}}{\overline{\Delta l \, D}}, \qquad \hat{\theta}_U(B) = \frac{\overline{\Delta l \, g_B (1-D)}}{\overline{\Delta l \, (1-D)}}, \tag{28}$$

both of which identify the complier density $f_c(B)$ when assignment is as-good-as-random, so their difference is a per-bin balance test. Because Δl is mean zero by construction, that difference has the closed form

$$\hat{\theta}_T(B) - \hat{\theta}_U(B) \;=\; \frac{\Pr(s \in B)\; \mathbb{E}[\Delta l \mid s \in B]}{\overline{\Delta l \, D}}, \tag{29}$$

so the statistic is the bin-conditional instrument mean, rescaled by the first stage. Two features of Equation 29 matter for reading it. The rescaling is by $\overline{\Delta l \, D} \approx 0.0095$, which magnifies the underlying imbalance roughly a hundredfold. Since $\sum_B \Pr(B)\; \mathbb{E}[\Delta l \mid B] = \mathbb{E}[\Delta l] = 0$, the per-bin deviations sum to zero and are therefore not independent tests.

This diagnostic rejects on the scope axis and not on the quality axis, and the contrast survives two-way clustering on customer number $\times$ art unit, eleven of twenty scope bins reject (maximum $|z| = 3.9$) against one of twenty quality bins (maximum $|z| = 2.2$). Clustering instead on art unit $\times$ publication year gives twelve and five. Panel A of Table F.9 reports both. The asymmetry has a simple explanation. Measured leniency is examiner j's realized grant rate, $l_j = n_j^{-1} \sum_{i \in j} D_i$, and the grant rate falls steeply in filed scope (0.769 to 0.578 from the narrowest to the broadest bin, against 0.701 to 0.714 across quality). An examiner whose caseload happens to be scope-heavy therefore records

$$\mathbb{E}[l_j \mid \bar{s}_j] \;=\; \lambda_j \;-\; b\,(\bar{s}_j - \mathbb{E}[\bar{s}]), \qquad b > 0, \tag{30}$$

where λ_j is the examiner's true strictness and $\bar{s}_j$ the mean filed scope of the caseload: they are *measured* as less lenient without being stricter. Caseloads are persistent in scope (cross-half correlation of the examiner's mean filed scope within a cell, 0.35), so an application's own scope is

positively correlated with its examiner's caseload mean, and $\mathbb{E}[\Delta l \mid s]$ inherits a negative slope in s. Empirically the bin-level imbalance correlates 0.89 with the bin grant rate. A characteristic thus breaks distributional balance only if it is *both* caseload-persistent and grant-shifting. Filed scope is both. Quality is if anything more caseload-persistent (0.84) but nearly flat in grants, rising by a point where scope falls by nineteen.

Sign of the bias. The consequence for the paper's central object follows from decomposing the representation ratio. With a mean-zero instrument,

$$\rho(B) \;=\; \underbrace{\frac{\mathrm{Cov}(\Delta l,\, D \mid B)}{\overline{\Delta l\, D}}}_{\text{complier mass}} \;+\; \underbrace{\frac{\mathbb{E}[\Delta l \mid B]\;\Pr(D=1 \mid B)}{\overline{\Delta l\, D}}}_{\text{contamination}}. \tag{31}$$

The contamination term carries the sign of $\mathbb{E}[\Delta l \mid B]$, which is negative in broad-scope bins and positive in narrow ones. Estimated complier mass is therefore displaced down exactly where the complier share is highest, and

$$\operatorname{plim} \partial_s\, \hat{\rho}(s) \;<\; \partial_s\, \rho(s)\,: \tag{32}$$

the estimated scope gradient is attenuated. Contamination of this form cannot manufacture the rising profile documented in Section 6.2. It can only flatten it.

Verification. I confirm the sign two independent ways. Netting the contamination term of Equation 31 out algebraically turns the profile measured with the pass-through instrument, which runs from 1.048 to 1.034 over the $p2.5$–$p97.5$ range of filed scope (nearly flat, 0.99×), into one running from 0.834 to 1.156 (a rising 1.39×). Alternatively, rebuilding leniency from scratch as a leave-one-out examiner mean of $D_i - \hat{g}(s_i)$, grants purged of the caseload's scope composition, collapses the bin-level imbalance correlation from 0.93 to 0.49 while leaving quality unchanged (0.72 to 0.72), and steepens the profile from 1.30× to 1.51×. Every figure in this comparison is measured against the same unpurged leave-one-out benchmark, so the starting correlation is its 0.93 rather than the 0.89 of the pass-through instrument reported above. Compliers among granted widen accordingly, from 4.40–7.59% to 4.02–8.08% across the scope range. Panel B reports three levels, and only the first is the pass-through estimator's. Evaluated at Equation 25, the pass-through instrument gives $\bar{\psi} = 0.691$ and an aggregate complier share of 4.9%, recovered here as a check. The two leave-one-out rows sit at $\bar{\psi} \approx 0.86$ not because the attenuation discussed in Appendix E has been undone, but because they are a different instrument on a different scale. Their levels are therefore not comparable to the pass-through estimator's, and what the pair identifies is the difference the purge makes. That difference is negligible: $\bar{\psi}$ moves from 0.8560 to 0.8569 and the implied share stays at 5.49%. The quality profile is untouched throughout, declining by a factor of 0.93 before and 0.94 after the purge. Contamination of measured leniency is thus a statement

about complier *composition*, not about the level, and it runs against the paper's finding rather than toward it. Panel B of Table F.9 collects these comparisons.

One limit confines the purged numbers to evidence on the *direction* of a bias. The exercise is run on the scalar leniency deviation Δl, and that is not the examiner-dummy UJIVE form of the pass-through estimator, which uses the examiner variation jointly rather than collapsing it into one scalar. The two therefore weight examiners differently, and no equivalence result links the weights the purged leniency deviation implies to those UJIVE delivers. That gap closes by computation rather than by a theorem: rebuilding the UJIVE instrument itself from the purged grants, on the same design matrices, the same leave-out construction, and the same Johnson–Lindenstrauss draws, so that the two instruments differ by the purge and by nothing else, steepens the examiner-dummy scope profile from 1.29× to 1.46× over the same p2.5–p97.5 range and leaves the quality profile at 0.99×. The conservative direction of the bias is therefore a property of the examiner-dummy design, not of the scalar convention. The purge still does not produce a replacement value for the gradient, and the levels reported in either column are the purged instrument's own rather than the paper's.

Finally, the magnitude should be read in instrument units rather than in the units of Equation 29. The bin-conditional imbalance $\mathbb{E}[\Delta l \mid B]$ reaches at most 0.0028 against a standard deviation of Δl of 0.119, or 2.3% of a standard deviation. At $n = 5.4$ million applications it is precisely estimated rather than large.

The conditional leniency dose. Appendix E factors the conditional complier share into a scalar dose and a shape, $\psi(x)\,\bar{d} = \bar{\psi}\,\rho(x)\,\bar{d}$, and this is legitimate only if the dose does not itself vary with x. What licenses the factorization is the scalar-index structure of Section 5.2: with leniency entering the grant rule additively and separably, ψ depends on (q, s) only through the acceptance index, and the dose is a property of the assignment process rather than of the application. That is not an innocuous assumption. Were the dose to vary with x, $\rho(x)$ would not be carrying all the conditional variation and the product would misstate the profile's *shape*, not merely its level. The condition is testable and I test it here. I sort applications into twenty equal-mass bins on each conditioning axis, the construction the complier profiles use, and compute the bin-conditional dose, clustering standard errors and the joint test on art unit × year. Two dose statistics are reported because two are in circulation: the object actually pulled out of Equation 11 is the *one-sided* favorable deviation $\mathbb{E}[(\Delta l)_+ \mid X = x]$, while the two-sided $\mathbb{E}[|\Delta l| \mid X = x]$ is the natural flip-rate analogue and is twice as large under mean-zero recentering. The error the factorization commits at x is exactly $\bar{d}(x)/\bar{d} - 1$, so the table reports that ratio in percent.

The construction reproduces the pipeline it is meant to audit: the unconditional $\mathbb{E}[(\Delta l)_+] = 0.05028$ against the 0.0503 used throughout, and $\mathbb{E}[\Delta l] = -1.3 \times 10^{-5}$ confirms the recentering.

Table F.8: *Conditional leniency dose: deviation from the pooled dose*

		Deviation from pooled dose (%)			
Axis	Dose statistic	Max abs.	Max − min	SD across bins	Wald $\chi^2(19)$
Quality	$(\Delta l)_+$	4.04	6.18	1.29	73.6
Quality	$\lvert\Delta l\rvert$	3.46	4.36	1.04	179.0
Filed scope	$(\Delta l)_+$	2.02	3.51	1.08	114.9
Filed scope	$\lvert\Delta l\rvert$	3.18	5.96	1.73	202.5

Notes: Twenty equal-mass bins per axis on the estimation sample (n = 5,364,969 applications in 9,701 art unit × year clusters). Entries are the bin-conditional dose expressed as a percentage deviation from its pooled value, which is exactly the error the factorization $\psi(x)\,\bar{d}$ commits at that bin. The Wald statistic tests equality of the dose across all twenty bins, clustered on art unit × year. The five percent critical value is 30.1.

Two things follow. First, *the joint test rejects on every row and that fact is not informative.* At five and a third million applications a Wald statistic is a statement about precision, not about magnitude, and the same trap recurs wherever this sample is tested bin by bin. What the table establishes is the size of the approximation: the conditional dose is flat to within four percent of itself on quality and three percent on filed scope, and the spread across bins is around one percent. The scalar pull-out is an approximation whose error has now been measured rather than assumed away.

Second, the small residual variation runs in the direction of the caseload-composition bias discussed above. The dose *falls* in quality, from about +3.2% of the pooled value in the bottom decile to −1.6% in the top. Across filed scope it *rises* through the eighteen interior bins, from about −1.5% to +2.0%, with both endpoint bins reverting, so it should not be described as monotone. Because the dose also rises in filed scope, holding it fixed at its pooled value makes the complier profile of the pass-through estimator rise slightly less than it would with a bin-conditional dose. Between the bottom and the top decile of filed scope, the profile would rise by 73% instead of 71%. Between the third and the nineteenth bin, where the dose moves most, it would rise by 68% instead of 62%. As with the caseload composition analyzed above, the approximation attenuates the scope gradient rather than manufacturing it.

Table F.9: *Distributional balance and caseload composition in measured leniency*

	Filed scope	Invention quality
Panel A: preconditions and the balance test		
Caseload persistence (split-half r)	0.35	0.84
Grant rate, lowest → highest bin	0.769 → 0.578	0.701 → 0.714
Bins rejecting, cluster art unit × year	12/20	5/20
max $\lvert z\rvert$	9.3	4.5
Bins rejecting, two-way cluster	11/20	1/20
max $\lvert z\rvert$	3.9	2.2
Max $\lvert\mathbb{E}[\Delta l \mid B]\rvert$, in sd of Δl	0.024	0.023
corr(imbalance, bin grant rate)	0.89	0.13
Panel B: consequence, by leniency measure		
	$\bar{\psi}$ / share (%)	scope ρ, p2.5 → p97.5
Paper instrument Δl	0.6906 / 4.89	1.048 → 1.034 (0.99×)
Leave-one-out grant rate	0.8560 / 5.49	0.868 → 1.125 (1.30×)
Leave-one-out, scope-purged	0.8569 / 5.49	0.794 → 1.196 (1.51×)
Algebraic netting-out	—	0.834 → 1.156 (1.39×)
Quality profile ρ, same three measures		
Paper instrument Δl	—	1.111 → 0.838 (0.75×)
Leave-one-out grant rate	—	0.998 → 0.932 (0.93×)
Leave-one-out, scope-purged	—	0.991 → 0.932 (0.94×)

Notes: Distributional balance of the examiner instrument on predetermined characteristics, and the caseload-composition mechanism behind the filed-scope rejection. Sample: 5,364,971 applications, publication years 2005–2020. 20 equal-mass bins per axis. *Panel A.* Caseload persistence is the cross-half correlation of the examiner mean characteristic within an art unit × year cell (66,568 examiner × cell units with at least 15 applications per half); it measures whether examiner identity predicts the caseload. The balance statistic is the treated-minus-untreated contrast of Equation 28, which equals the bin-conditional instrument mean rescaled by the first stage (Equation 29); rejection is at the five percent level. Two-way clustering is on `customer_number` × `examiner_art_unit`. The imbalance is also reported in standard deviations of Δl (sd = 0.120), since the rescaling in Equation 29 magnifies it by roughly a hundredfold. A characteristic breaks balance only if it is both caseload-persistent and grant-shifting: filed scope is both, quality is more persistent but grant-flat. *Panel B.* $\bar{\psi}$ is the pooled first stage, estimated after partialling out the art unit × year fixed effects and covariates, and the share is $\bar{\psi}\ \mathbb{E}[(\Delta l)_+]/P(D=1)$ (Equation 25). Following that convention the dose $\mathbb{E}[(\Delta l)_+]$ is taken on the recentered but unresidualized instrument; the share is invariant to rescaling the instrument, but not to residualizing it, so the convention has to be stated. On this convention the first row reproduces the paperś headline floor. ρ is the representation ratio $f_{\text{complier}}/f_{\text{pop}}$ over the p2.5–p97.5 range. The scope-purged instrument is a leave-one-out examiner mean of $D_i - \hat{g}(s_i)$; the algebraic netting-out instead removes the contamination term of Equation 31 from the paper instrument. Both routes steepen the scope profile and leave the level and the quality profile unchanged.

A pre-treatment anchor for the expected-leniency benchmark. The recentred instrument Δl_p subtracts from the assigned examiner's leniency the mean leniency of the examiners the application could have been assigned. In the shipped construction that pool is assembled around the application's *decision* date, within a symmetric ±18-month window. Two features of this construction matter. For a granted application the decision date is the issue date, which leniency itself moves, since a more permissive examiner grants sooner, so the benchmark is not strictly predetermined.

Because the window is symmetric, part of the pool consists of examiners deciding up to eighteen months *after* the anchor, which no applicant could have observed ex ante.

I therefore rebuild the benchmark on an anchor that the examiner assignment cannot move. The application's publication date is scheduled by statute at roughly eighteen months from the priority date, so it is fixed independently of who examines the case. Anchoring on it, restricting the window to the eighteen months *preceding* publication, pooling within the art unit that the design randomizes inside of, and leaving out the applicant's own examiner yields a pre-treatment instrument defined for 5,333,672 applications, 99.4% of the core sample, with a median pool of 909 decisions. The two instruments correlate at 0.93. Table F.10 reports both.

The first stage strengthens under the re-anchoring: the grant response per standard deviation rises from +0.0996 to +0.1148, a factor of 1.15. The pre-treatment instrument is therefore usable, and the objection is a robustness question rather than a threat to identification. It does not, however, explain the residual association between measured leniency and the text components documented above. That association is slightly larger on the pre-treatment instrument, not smaller, so re-anchoring can be ruled out as its source. The caseload composition analyzed earlier in this appendix remains the operative explanation.

The composite-quality row concerns the scalar leniency deviation, and its reading is a decomposition. The −0.0380 is estimated on an object and a conditioning set the design never uses: the raw level of the composite measure, with the fixed effects alone. Re-expressing the outcome as the within-cell percentile rank, the object every design exhibit conditions on, takes the association to −0.0106 (0.0026), so nearly three quarters of it is variation in the measure's level that the rank never transmits. Adding the control set of the first stage takes the rank to +0.0002 (0.0027). On the pre-treatment instrument the same two steps run −0.0522 to −0.0139 (0.0028) to −0.0031 (0.0029). With the rank as outcome and the controls of the first stage, the association of the leniency deviation is indistinguishable from zero on either anchor (Table F.10, bottom panel). What the conditioning does not remove is the component loadings: impact and novelty stay individually associated with the instrument under the full control set (+0.0191 and −0.0245 per standard deviation on the shipped instrument, with the same signs and slightly larger magnitudes on the pre-treatment one), which is the caseload-composition channel above.

Three properties of the rebuilt instrument should be stated by anyone using it. First, a backward-looking window under-estimates contemporaneous leniency because grant rates drift upward over the sample, so Δl_p built this way has mean +0.012 rather than zero. It is absorbed by the art unit × year effects but it is not recentred in the sense the main text uses. Second, measured leniency is not an examiner-level constant. Only 1.6% of the 12,709 examiners appearing in the leniency data carry a single value (12,706 of them enter the connected estimation core of Section 5), so any pool average is a weighted average over examiner-periods rather than over

examiners. Third, the publication anchor is not universally pre-decision: for 3.6% of applications the decision precedes publication, and for those the anchor is predetermined by statute but not by timing.

Table F.10: *Expected-leniency benchmark: decision-date versus pre-treatment anchor*

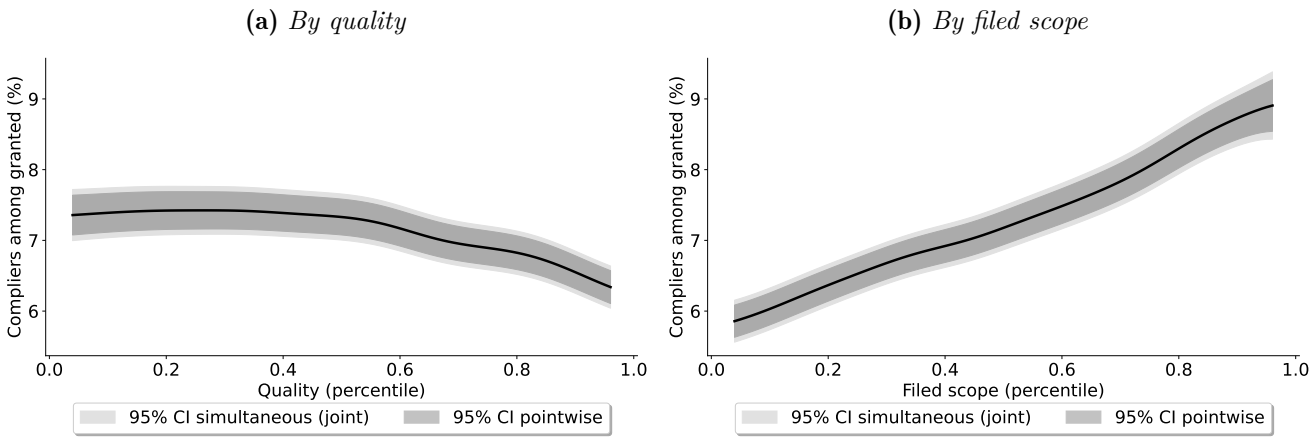
	Shipped Δl_p (decision anchor)	Pre-treatment Δl_p (publication anchor)	Ratio
Standard deviation	0.129	0.144	
Grant *(first stage)*	+0.0996	+0.1148	1.15
	(0.0006)	(0.0004)	
Impact component	+0.0145	+0.0149	1.03
	(0.0017)	(0.0018)	
Novelty component	−0.0231	−0.0269	1.16
	(0.0017)	(0.0018)	
Composite quality	−0.0380	−0.0522	1.37
	(0.0044)	(0.0055)	
Composite quality, reconciled to the design conditioning			
as within-cell rank (design object)	−0.0106	−0.0139	1.31
	(0.0026)	(0.0028)	
with the first-stage control set	−0.0248	−0.0401	1.62
	(0.0046)	(0.0058)	
rank and control set (design conditioning)	+0.0002	−0.0031	—
	(0.0027)	(0.0029)	

Notes: Each cell is the coefficient from a separate regression of the outcome on the recentered instrument, with art unit × publication year fixed effects absorbed and standard errors clustered on the same cell, expressed per standard deviation of the instrument in use. All outcomes other than the grant indicator are standardized. The shipped instrument centres the expected-leniency pool on the applicationś decision date within a symmetric ±18-month window. The pre-treatment instrument centres it on the statutory publication date, uses only the preceding 18 months, pools within art unit, and leaves out the applicantś own examiner; it is defined for 5,333,672 applications, 99.4% of the core sample, with a median pool of 909 decisions. Estimation sample 5,333,618 applications after singleton fixed effects are dropped. The two instruments correlate at 0.93. Ratio is the pre-treatment coefficient over the shipped one. The backward-looking window under-estimates contemporaneous leniency because grant rates drift upward over the sample, so the pre-treatment instrument has mean +0.012 rather than zero; it is absorbed by the fixed effects but is not recentered in the sense used in the main text. The final three rows re-express the composite-quality association on the designś own terms: as the within-cell percentile rank that every design exhibit conditions on, with the control set of the first stage (entity status, continuation type, examiner experience, first-year indicator) added, and with both. They are estimated on the same sample with the same fixed effects and clustering (singleton cells retained, 5,333,672 applications); a ratio of two near-zero coefficients is noise and is not reported.

G Robustness: sample restrictions

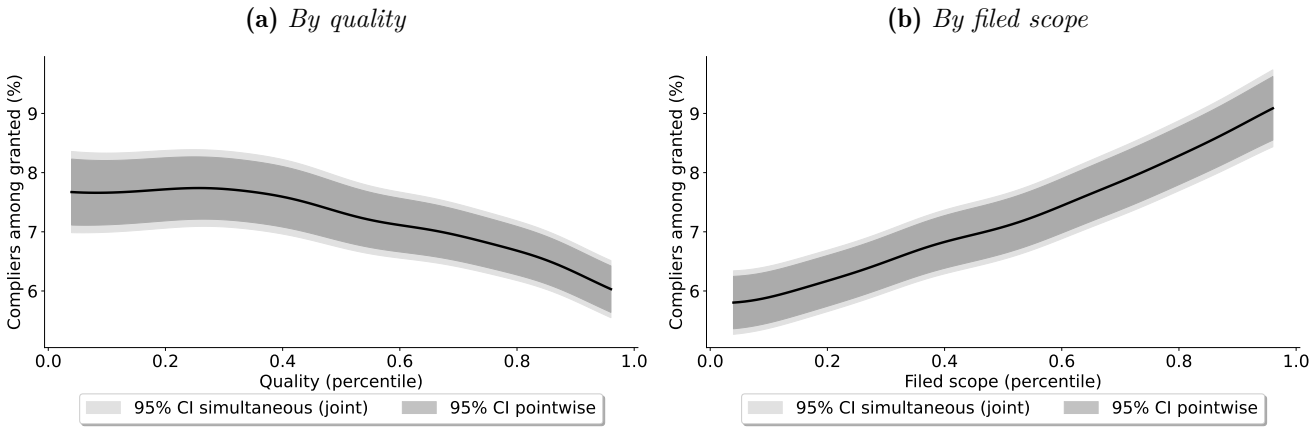
This appendix re-estimates the complier profiles of Figure 4 on two subsamples with the estimator of Section 5.3. The instrument is the headline's, built on the full sample. Only the rows the estimator uses change: the cell demeaning, the art-unit variances of leniency, the aggregate share and the profiles are computed on the subsample, and each profile is scaled to the subsample's own aggregate share. Table L.5 reports both checks at the deciles. Without continuations, 2,144,479 applications, the aggregate share is 7.1%. The share rises from 6.0% to 8.7% between the 10th and the 90th percentile of filed scope, by 45%, and falls from 7.4% to 6.5% across quality, by 11%. In the random-digit cells, 1,208,862 applications, the aggregate share is 7.2%. The share rises from 5.9% to 8.8% across filed scope, by 49%, and falls from 7.7% to 6.3% across quality, by 18%. In the full sample the two changes are 46% and 7.6%. The simultaneous test rejects flatness along both characteristics in both subsamples, with $p < 0.001$.

Figure G.1: *Robustness: complier shares excluding continuations*



Notes: The estimator of Figure 4 on the 2,144,479 applications that are not continuations, with the instrument built on the full sample. Continuation applications, which inherit prosecution history from a parent filing and are frequently assigned to the same examiner, are excluded. Each profile is scaled to the subsample's aggregate share. The shaded areas show 95% pointwise and 95% simultaneous confidence bands, and the two panels share one vertical axis.

Figure G.2: *Robustness: complier shares in random-digit art units*



Notes: The estimator of Figure 4 on the 1,208,862 applications in random-digit art unit × year cells, with the instrument built on the full sample. Random-digit cells are the art unit and docket-year pairs in which supervisors allocated applications to examiners by the terminal digit of the application serial number, which is assigned sequentially and is therefore unrelated to the content of the application (Sampat and Williams, 2019). The cells are identified from the list used in that literature, and an application is assigned to a cell by its publication year, offset by the statutory eighteen-month publication lag (publication year equal to the docket year plus one or two). Within these cells examiner assignment is random by construction, which is why they are commonly used as a robustness benchmark in the examiner leniency literature. Each profile is scaled to the subsample's aggregate share. The shaded areas show 95% pointwise and 95% simultaneous confidence bands, and the two panels share one vertical axis.

The random-digit cells differ from the estimation window in composition. Art units in the 2800, 3600, and 3700 series make up 85% of the subsample against 51% of the window, and the median publication year is 2010 against 2013. The leniency of the assigned examiners is as dispersed inside the cells as outside them, with a mean absolute deviation of 0.103 against 0.100.

H Robustness: repeat filers

This appendix asks whether repeat filers drive the complier composition of Section 6. The customer number identifies the law firm or in-house patent department that prosecutes an application (Section 5.5). The estimation sample contains 33,476 customer numbers, and prosecution is concentrated: the 7,838 customer numbers with at least fifty applications hold 92.1% of the sample, and 4.1% of applications carry no customer number. A repeat filer can enter the design in two ways. The first is through the instrument. The fits behind U_p leave out application p but keep the same firm's other applications, so if a firm's applications share a grant propensity and concentrate on few examiners, part of that propensity enters the instrument of each of them. The second is through the composition of dockets. Firms differ in filing practice, and if they are not spread evenly across the examiners of a cell, an examiner's grant rate partly reflects the firms on the docket. Leaving the whole firm out of the instrument removes the first channel, and conditioning on the firm removes the second. I report both, each against the baseline estimated on the same applications. Both checks hold fixed the pass-through estimator of Appendix E, whose instrument conditions on examiner experience and whose profiles take the dose to be the same in every bin. The levels and shapes below are that estimator's, and what each check measures is how far it moves them.

Leave-customer-out instrument. I rebuild the UJIVE instrument so that the fits for application p leave out every application that shares its customer number, in the examiner projection and in the control projection, $U_p = \hat{D}^{[Z,W]}_{-\mathcal{C}(p)} - \hat{D}^{W}_{-\mathcal{C}(p)}$, where $\mathcal{C}(p)$ is the set of applications with p's customer number. Applications without a customer number are left out one at a time. The sample, the controls, and the pass-through of Equation (23) are unchanged. A leave-out set can now hold up to 100,504 applications, which the random-projection leverages of Appendix E cannot accommodate, so I compute every fit exactly, from one inverse of each design's cross-product matrix and a low-rank update for each customer number. The comparison uses the same exact computation with one application per leave-out set, so the two differ only in the leave-out unit. That comparison reproduces the random-projection baseline to within 0.05% of its explained variation.

Leaving the firm out lowers the explained variation of the examiner dummies by 1.9%, so the leave-out fit recovers 94.9% of the in-sample one, against 96.7%. The deattenuated pass-through of Appendix E falls from 1.10 to 1.08, which lowers the upper bracket of the share of compliers among granted patents from 7.8% to 7.7%. The profiles do not move. At every point of either axis the curve lies within 0.6% of the baseline (Figure H.1). The share rises from 4.0% to 6.7% across filed scope (flatness rejected, $p < 0.001$) and stays between 4.8% and 5.1% across quality (flatness not rejected, $p = 0.11$). The level of both profiles, 4.9%, comes from the leniency deviation of

Equation (23), which does not use the examiner dummies and is the same in both arms.

Customer-number fixed effects. The direct way to absorb the composition of dockets is to add customer-number fixed effects to the controls W, in the instrument and in the pass-through regression alike. I hold the sample fixed. Applications without a customer number form one category, and the 25,638 customer numbers with fewer than fifty applications, which hold 3.9% of the sample, are pooled into another.

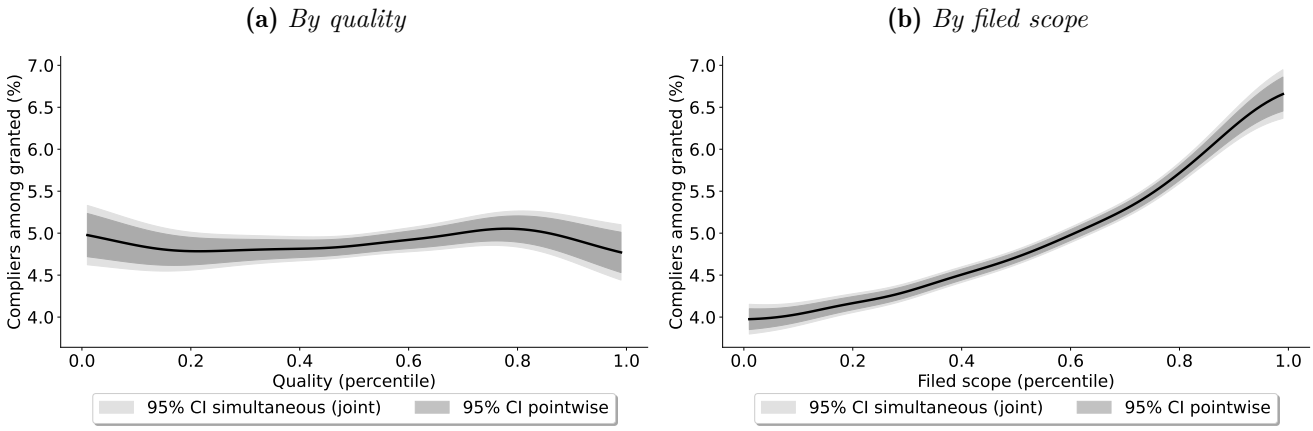
The customer number in the USPTO data is, however, recorded when the data are extracted, not at filing. It is the correspondence address on file at extraction, and the transaction history records changes of correspondence address and of power of attorney during and after prosecution. If the examiner assignment moves the probability of such a change, the fixed effects condition on an outcome of the examiner assignment. I test this by regressing an indicator for a change of correspondence address or power of attorney on the leniency deviation, within art unit $\times$ year cells and with standard errors clustered by examiner, separately for changes before the application is docketed to its examiner, between docketing and the decision, and after the decision. I fixed a tolerance in advance: one within-cell standard deviation of leniency should move the probability of a change by less than 0.1 percentage points in each window. Changes before docketing do not respond to the examiner assignment (-0.04 points per standard deviation, s.e. 0.02), and neither do changes after the decision (-0.04, s.e. 0.03). Changes during prosecution do. A one-standard-deviation more lenient examiner lowers their probability by 0.14 points (s.e. 0.03), against a mean of 21.2%. The sign matches the shorter prosecution a lenient examiner produces, which leaves less time for a change, and whatever the channel, the recorded customer number differs from the one at assignment more often under strict examiners. The effect is small against its mean but exceeds the tolerance, so I read the fixed-effects specification only next to the leave-customer-out instrument, which uses the customer number to define a leave-out set and never as a control.

With the fixed effects, the explained variation of the examiner dummies falls by 6.4% under UJIVE and by 6.5% in sample, so the leave-out fit still recovers 96.8% of the in-sample one. The pass-through of the leniency deviation falls from 0.69 to 0.67, which lowers the level of every profile by 2.8% and the share of compliers among granted patents from 4.9% to 4.8%. The deattenuated pass-through falls to 1.08, and the upper bracket to 7.6%, about as far as when the firm is left out of the instrument (1.08 and 7.7%), so most of the fall comes from the firm's own applications entering the instrument. The design does not separate how much of the rest reflects the post-assignment conditioning documented above and how much reflects firms that are spread unevenly across the examiners of a cell. The profiles keep their shape (Figure H.2). The share rises from 3.9% to 6.6% across filed scope (flatness rejected, $p < 0.001$) and moves from 4.9% to 4.6% across quality (flatness not rejected, $p = 0.24$). Net of the level, the curves lie within 2.6% of the baseline

at every point, with the largest departures at the ends of each axis.

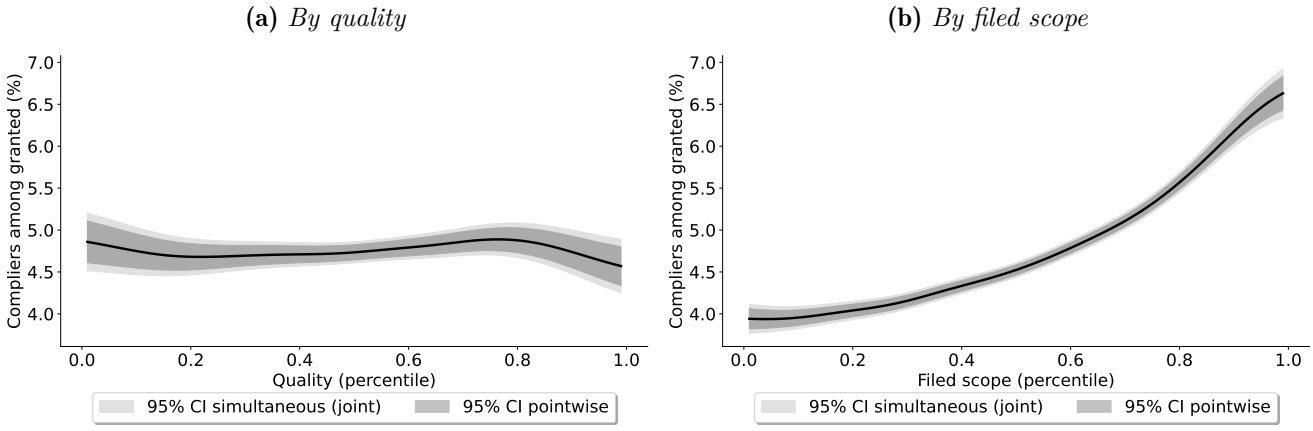
Leaving the firm out of the instrument moves no point of either pass-through curve by more than 0.6%, and the fixed effects lower the level by 2.8% and move the shape by at most 2.6%.

Figure H.1: *Robustness: complier shares with a leave-customer-out instrument*



Notes: The pass-through profiles of Appendix E, with the examiner-dummy UJIVE instrument built by leaving out every application that shares the customer number of the focal application, in both the examiner and the control projection. Applications without a customer number are left out one at a time. Fits are computed exactly. The 318 applications whose customer number accounts for every application of their examiner or of their art unit × year cell have no leave-out fit and are excluded. Sample and controls as in the pass-through estimator. The two panels share one vertical scale. Shaded areas show 95% pointwise and 95% simultaneous confidence bands.

Figure H.2: *Robustness: complier shares with customer-number fixed effects*



Notes: The pass-through profiles of Appendix E, adding customer-number fixed effects to the controls W of the instrument and of the pass-through regression, on the same applications. Applications without a customer number form one category, and customer numbers with fewer than fifty applications are pooled into another. The 284 applications alone in their examiner or their art unit × year cell have no leave-out fit and are excluded here and in the baseline the text compares against, which moves the baseline's explained variation by 0.04%. The two panels share one vertical scale. Shaded areas show 95% pointwise and 95% simultaneous confidence bands.

I Quality measure: construction, validation, and decomposition

This appendix supplements Section 3 with the full discussion of the quality measure's three design choices, summarized there in a single paragraph, and with three robustness exercises for the pre-grant quality measure: sensitivity to construction choices, convergent validity against external measures, and a decomposition into its novelty and impact components.

Design choice 1: titles and abstracts, not claims. Since claims are the legally binding part of a patent and are the main focus of negotiation during the prosecution process, applicants tend to narrow their claims in response to examiner objections. Therefore, the text of the claims at any stage in the prosecution process shows the examiner's decisions as much as the underlying invention. Titles and abstracts, however, are fixed in the filing text, reproduced in the first ("A1") publication at the statutory eighteen months, and remain mostly unchanged during the prosecution process. Using abstracts and titles therefore avoids one specific channel of contamination, as the text Q_p is built from cannot have been rewritten in response to the examiner's objections.

Design choice 2: within-cell percentiles. The percentile transformation $Q_p = P(FS_p/BS_p)$ is calculated separately for each art unit $\times$ publication year cell, employing the same partition that is used as the fixed effects in the first stage (Section 5). The resulting measure ranks each application relative to the cohort which it was evaluated alongside, taking into account both field-level differences in the average rate of development and the density of prior art, as well as year-to-year changes in either of these. An application with $Q_p = 90$ is in the top 10% of its art unit $\times$ year cell as the combination of novelty and impact, not in the top 10% of all USPTO applications. The cost of this within-field construction is in interpretation. Q_p measures quality relative to a technology field rather than on an absolute scale.

Because the measure enters every specification through this rank rather than through the level of FS_p/BS_p, any distortion that applies a common strictly monotone transformation to the *ratio* within an art unit $\times$ year cell leaves Q_p *exactly* unchanged. Distortions outside this class are the subject of the construction-robustness table below and of the balance tests (Section 5.5 and Appendix F).[35]

Design choice 3: orthogonality to the examiner assignment. Decisive for the design in Section 5, Q_p must be orthogonal to the examiner assignment. This requirement is distinct from the timing

[35] Cell-common rescalings of the similarity scale are within the class. Distortions that hit FS_p and BS_p separately need not be, since even a common additive shift to both can reorder a ratio, and truncation of the forward window for the most recent cohorts, bursts in filing volume, and drift in the density of the embedding space are cell-common only approximately. Appendix N draws the same boundary for the rank-defined scope results.

of the text. The two are routinely run together, and only the first is load-bearing. Validity as a control does not require that Q_p be knowable at the time of the decision. It requires that the leniency of the examiner assigned not move it. A measure built entirely from information realized after the decision is a valid control if the examiner assignment does not shift it, and a measure built entirely from information available before the decision is not, if the examiner assignment does. The forward-similarity component FS_p is the part of the construction that could fail this: because it is built from patents filed in the five years after the focal application, the focal grant might itself shape subsequent filings in the art unit, making $\bar{v}_p^+$ endogenous to the examiner decision.

Two tests bear on it, at different levels of aggregation. At the level of the art unit, the exclusion-relevant test is whether leniency moves a unit's forward technological trajectory: regressing the art-unit × year mean forward similarity on the art-unit × year mean leniency with art-unit and year fixed effects, the coefficient is −0.05 standard deviations (s.e. 0.04, 9,896 unit-years), as Table I.7 reports. The raw cross-section of art-unit levels is positive, but that is technology composition, fast-moving fields also differing in grant rates, not a trajectory response. And because the leniency measure predicts the art-unit grant rate almost one-for-one, a poorly measured regressor cannot explain the within-unit null. Examiner leniency does not move the forward technological trajectory of an art unit.

At the level of the individual application, the question is whether the within-cell rank Q_p, the object every estimate conditions on, is balanced against the variation the design uses. It is not exactly balanced. With the instrument of Section 5.3, the mean quality of the compliers computed from the granted applications differs from the mean computed from the refused ones by 1.9 percentiles ($z = -17.9$), and examiner experience accounts for most of the difference. The estimator does not require this balance. Its profiles compare examiners within cell × bin, so sorting on quality enters only the dose. What the design requires of Q_p is that the examiner assignment does not move it, which the art-unit test above addresses.[36]

Construction choices. The measure is built from title-and-abstract text rather than from claims, for the fixed-at-filing reason of Section 3, embedded with sentence-T5 (Ni et al., 2022), and expressed as art unit × publication year percentiles (Section 3.4). Table I.1 reports the Spearman rank correlation of the baseline measure with its un-percentiled version: the cell-wise normalization reorders the measure only modestly (Spearman 0.89 with the raw measure on the estimation window), so the ranking reflects genuine within-field variation rather than an artifact of cross-field level differences. Two further families of construction choices are varied, both of which leave the

[36] Neither test reaches a channel operating within the application rather than across the art unit. Continuation and divisional filings enter the forward window carrying text nearly identical to their parent, so a grant that generates them could raise FS_p for that application alone, which an art-unit average cannot detect. Section 6.2 re-estimates the composition excluding continuations. Appendix N collects the design's other limits.

underlying embeddings intact and change only what each application is compared against: the comparison pool's reference class and anchor (coarsening the focal art unit to the technology center or the workgroup, anchoring on the publication date in place of the filing date, or drawing the pool as a fixed count of nearest-in-time documents rather than a fixed-duration window), and the comparison window's length and placement (six months to ten years, with and without a gap at the anchor). The rank correlations span 0.14 to 0.94, and the range tracks what a variant changes: anchoring, placement, and the normalization leave the ranking nearly intact (0.89–0.94), while it moves with how the comparison pool itself is drawn, its reference class, its membership, and how short its window is – that is, with *who* an application is compared against. That dimension is an object of study rather than a nuisance: reference-class dependence is measured directly, its complier loading the +0.55% residual of Table I.5, and under every variant in the family, the grant-conditioned pool excluded as post-treatment, the pooled complier mean on the variant's own rank stays within one percentile point of the population mean, computed in the same sweep behind that table.

Table I.1: *Quality measure: sensitivity to construction choices*

Variant	Alters	Rank corr. with baseline
Raw (un-percentiled)	Within-cell normalization	0.886
Publication-date anchor	Comparison-pool anchor date	0.939
Technology-center pool	Comparison-pool reference class	0.334
Workgroup pool	Comparison-pool reference class	0.504
Six-month windows	Comparison-window length	0.243
One-year windows	Comparison-window length	0.443
Three-year windows	Comparison-window length	0.837
Ten-year windows	Comparison-window length	0.817
180-day gap at the anchor	Comparison-window placement	0.936
365-day gap at the anchor	Comparison-window placement	0.926
Nearest-in-time pool (k=200)	Pool construction (fixed count)	0.140

Notes: Spearman rank correlation of each construction variant with the baseline measure (filing-date-anchored comparison pool within the focal art unit, expressed as the within-cell percentile), on the estimation window 2005–2020, pairwise-complete per row (N between 5,352,310 and 5,364,971). Pool and window variants leave the underlying embeddings intact and change only what each application is compared against. Each variant is a one-dimension change from the publication-anchored construction, whose own row is the anchor contrast; rows below it therefore differ from the filing-anchored baseline in their stated dimension and the anchor date jointly. The nearest-in-time pool draws the 200 documents closest in time on each side of the anchor in place of the fixed-duration window. The raw row removes the within-cell percentile normalization. A variant whose comparison pool conditions on the grant outcome is excluded as post-treatment.

Convergent validity. Section 3.4 reports the two steps of the validation, and this appendix holds their tables. Table I.2 covers applications that were never granted. Among them, the top decile of Q_p receives 41.8% more citations from patents examined in other art units than the 40th to 60th percentiles, with a 95% confidence interval from 38.7% to 44.9% and standard errors clustered

by art unit × year. This comparison does not condition on the grant. Table I.3 covers granted patents. Relative to the 40th to 60th percentiles, the top decile, the top 5% and the top 1% of Q_p receive 32.2%, 39.8% and 54.9% more citations, and their stock market reaction (Kogan et al., 2017) is 7.6%, 10.5% and 12.8% higher. These proxies exist only for granted patents, so they condition on the grant. The same proxies by revision round appear in Figure D.3.

Table I.2: *Citations to refused applications by invention-quality bin*

Invention quality	n	All citations	Other art unit	Same art unit	Added by examiner	Added by applicant
Median (p40–60)	310,302	0.96	0.96	0.95	0.97	0.96
Top 10%	154,060	1.39 (+44%)	1.36 (+42%)	1.56 (+64%)	1.32 (+36%)	1.41 (+47%)
Top 5%	78,063	1.49 (+55%)	1.47 (+52%)	1.71 (+80%)	1.40 (+44%)	1.52 (+58%)
Top 1%	16,954	1.62 (+69%)	1.61 (+67%)	1.84 (+94%)	1.51 (+55%)	1.67 (+74%)

Notes: Applications never granted, publication years 2005–2020. Invention quality is the within-art-unit×year percentile rank of the text-based measure, re-ranked among these applications. Each cell is the mean, over the applications in the quality bin, of the number of later US patents citing the application's publication, divided by its mean within the application's art unit×publication year cell. Parentheses give the percent difference from the median (p40–60) bin. The columns count all citing patents, those examined in another art unit or in the same art unit as the cited application, and those whose citation was added by the examiner or by the applicant. Applications in cells without any citation of the counted kind are dropped, so the bin size n, shown for the other-art-unit column, differs across columns by at most 0.9%.

Table I.3: *Convergent validity: invention-quality value premium by tail bin*

Invention quality	n	Norm. fwd. citations	Market value (Kogan et al.)
Median (p40–60)	762,543	0.566	27.8
Top 10%	386,801	0.748 (+32%)	29.9 (+8%)
Top 5%	195,282	0.791 (+40%)	30.7 (+10%)
Top 1%	41,942	0.877 (+55%)	31.3 (+13%)

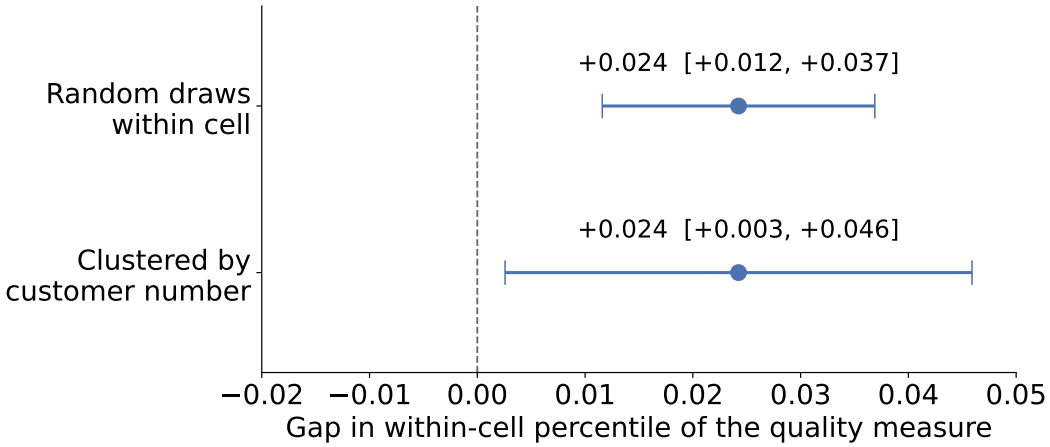
Notes: Granted applications, publication years 2005–2020. Invention quality is the within-art-unit×year percentile rank of the text-based measure, re-ranked among granted applications. Each cell is the mean over non-missing observations in the quality bin, and n is the bin size. Parentheses give the percent difference from the median (p40–60) bin. Forward citations are the citations a patent receives from later US patents, divided by the mean among cited patents issued in the same year and art unit, with uncited patents counted as zero (41.2% of granted applications). Winsorizing the ratio at its 99th percentile leaves premia of +30%, +36% and +49% across the three tail bins. Market value is the Kogan et al. (2017) estimate ($m), matched on the patent number and available for the publicly listed subset of grants.

Inventions recognized by outside parties. The proxies above are citations, market values and patent text. A further check asks whether inventions that an outside party judged to be new, found to be required by a technical standard, or selected for an award rank high on Q_p within their cells. I use three such judgements. The first is the FDA's classification of a drug as a new molecular entity, an active moiety never approved before, and I take the drug-substance patents

listed for these drugs in the Orange Book. The second is an independent evaluator's finding that a patent is essential to a technical standard, and I take the patents in standard pools that check essentiality, such as the H.264/AVC, HEVC and 4G/5G pools. The third is the jury's selection of finalists for the European Inventor Award, and I take the US family members of the finalists' European patents. The year of recognition is the approval year for drugs, the first year a patent appears in an archived pool list, and the award year for finalists. Only applications whose issue year lies more than three years before the year of recognition enter. This leaves 1,576 inventions with 2,683 granted applications in the estimation sample.

For each application, the gap is its percentile of Q_p minus the mean percentile of its art unit $\times$ publication year cell. Each invention receives weight one, split equally across its applications. Figure I.1 compares the mean gap of the recognized inventions with the mean gap obtained when each application is replaced by a random granted application from its own cell. The recognized inventions rank 2.4 percentile points higher, with a 95% confidence interval of 1.2 to 3.7 points from 10,000 random draws. Inventions filed under the same USPTO customer number, which identifies the correspondence address and often the law firm, need not be independent. Resampling customer numbers with replacement gives a 95% confidence interval of 0.3 to 4.6 points.

Figure I.1: *Quality of inventions recognized by outside parties*



Notes: The figure shows the mean gap in the within-cell percentile of the quality measure between 1,576 recognized inventions and random granted applications from the same art unit $\times$ publication year cell. The recognized inventions are drug-substance patents of new molecular entities listed in the FDA Orange Book, patents in standard pools that check essentiality, and finalists of the European Inventor Award, with 2,683 granted applications in all. Only applications issued more than three years before the recognition enter. Each invention has weight one, split equally across its applications. The dot is the mean gap of the recognized inventions minus the mean gap when each application is replaced by a random granted application from its cell. In the upper row, the bar is the 95% confidence interval from 10,000 such draws. In the lower row, it comes from 10,000 resamples of USPTO customer numbers, with each invention assigned its most frequent customer number. A gap of 0.01 is one percentile point.

Novelty-impact decomposition. The composite measure combines two components of Section 3: impact i_p, the similarity of the application to the average of its art unit's applications in the five following years, and novelty n_p, one minus its similarity to the average of the five preceding years. Both compare the application with an art-unit average, so impact measures how typical the application is of the work that follows it. The two components load on the grant decision with opposite signs and similar size. Table I.4 re-estimates the complier profile with each component in place of the composite. On novelty alone the share rises by 15.5% between the 10th and the 90th percentile, and on impact alone it falls by 16.8%, against a fall of 7.6% on the composite and a rise of 46% along filed scope. An application far from the preceding work of its art unit is often also far from the work that follows it. I rank this distance from both averages, atypicality, within art unit × year cells as the mean of standardized novelty and the negative of standardized impact. Along atypicality the share rises by 18.0%. Because atypical applications could also be the ones that request broad scope, Table I.4 also checks whether atypicality accounts for the rise along filed scope. Holding the mix of atypicality fixed across filed scope leaves 95% of that rise in place, and the share rises from the narrowest to the broadest filings within each of eight atypicality groups.

Table I.4: *Decomposition: complier share by composite quality, its two components, atypicality, and filed scope*

Axis of the complier profile	10th pctile	90th pctile	Change
Composite quality measure	7.20%	6.65%	−7.6%
Novelty component only	6.63%	7.66%	+15.5%
Impact component only	7.81%	6.50%	−16.8%
Atypicality (far from both averages)	6.57%	7.75%	+18.0%
Filed scope *(reference)*	5.91%	8.64%	+46.2%

Notes: Share of grants owed to the examiner assignment, in percent, at the 10th and the 90th percentile of each axis, on the smoothed curve of the estimator of Section 5.3. The quality rows rank applications within art unit × year cells on the stated measure. Each row is scaled so that its grant-weighted average equals the aggregate share. The simultaneous test rejects flatness in every row, with $p < 0.001$. The atypicality row ranks applications on the mean of standardized novelty and the negative of standardized impact. On the 8×8 surface of atypicality and filed-scope octiles, the share rises by 3.19 percentage points from the narrowest to the broadest filed-scope octile, and by 3.03 points when every filed-scope octile has the same mix of atypicality, a ratio of 0.95 (standard error 0.008). Within each atypicality octile the broadest filings exceed the narrowest by between 2.5 and 3.3 points. Regressing grant on within-cell percentile ranks gives +0.061 on impact and −0.066 on novelty, against +0.013 on the composite. The component loadings do not sum to the composite's, because the composite ranks their ratio rather than adding them. Part (i) of Proposition 1 predicts a rising complier profile along a component that lowers the acceptance index and a falling one along a component that raises it. The novelty and impact rows have these signs. The composite's net grant loading is small, and the grant rate accounts for only a sixth of its decline, as Section 6.3 shows.

Ambiguity of the quality assessment. Two senses of "hard to judge" concern the quality measure rather than the examiner, whose assessment noise Appendix B.1 takes up separately. Table I.5 reports both indices in all four conditioning forms. The measurement-noise score loads on compliers

negatively or not at all in every form, so compliers are not applications whose quality the measure pins down poorly. That reading carries a caveat the table itself reports: the noise score fails the treated-versus-untreated balance check in every conditioning form, including the within-cell one, so its loading is not as clean a complier characteristic as the other rows. The direction of the conclusion would survive an imbalance that moves the loading's size, but the table does not establish the imbalance immaterial. The reference-class score is positive in every form and halves once ranked within art unit × year, so roughly half of the raw association is between-cell composition. What survives is the small application-level residual the text quotes, and the flatness of its profile is rejected at the five percent level, with $p = 0.027$.

Table I.5: *Complier loadings on the ambiguity scores*

Complier − population (%)	z	Balance z	Flat p
Measurement noise: |half 1 − half 2|, fixed construction			
raw	+0.36	+0.65	+3.67	0.013
within quality deciles	+0.59	+1.02	+4.07	0.004
within filed-scope deciles	+0.15	+0.28	+3.96	0.025
within art unit × year	−0.36	−2.36	+4.14	0.000
Reference-class dependence: |art-unit − technology-center pool|			
raw	+1.29	+3.70	+0.99	0.000
within quality deciles	+1.26	+3.76	+1.25	0.001
within filed-scope deciles	+1.18	+3.40	+1.10	0.002
within art unit × year	+0.55	+3.65	+0.79	0.027

Notes: The Table reports, for two ambiguity indices in four forms each, the gap between the complier mean and the population mean of the score rank, in percent of the population mean, from the examiner-dummy UJIVE modified-outcome estimator of Appendix E on all 5,367,639 applications of the estimation window. Unlike the headline sample, this count keeps the applications without a headline quality score, because each index has its own coverage. Standard errors are clustered on art unit × publication year. The measurement-noise score rates every application on two independent halves of the text corpus at fixed construction and takes the absolute disagreement; the reference-class score holds the corpus fixed and takes the disagreement between the quality rank built on the art-unit comparison pool and the same rank rebuilt on the coarser technology-center pool (comparison pools selected on grant outcomes are excluded as post-treatment). Each score is percentile-ranked and, in the conditional forms, re-ranked within deciles of quality, within deciles of filed scope, or within art unit × publication year cells; the last form separates an application-level property from art-unit-level composition and is the form quoted in the text. Balance z is the treated-versus-untreated arm difference of the modified-outcome construction; Flat p is the simultaneous sup-t test that the complier representation profile over the score is flat.

Reliability of the quality rank. Table I.6 bounds what the noise from the comparison documents could do to the complier-quality profile. The split-half correlation scores every application on two independent halves of the text corpus at fixed construction. Each application's own text is the same in both halves, so the split does not measure how well that text records the invention's quality. At 0.88, the Spearman-Brown reliability of the full-corpus rank is 0.94, so this noise attenuates a binned profile over the rank by at most seven percent of its amplitude. The dispersion across

construction variants is a separate object, a choice rather than noise, and is measured by the reference-class score above.

Table I.6: *Reliability of the quality rank*

Split-half correlation of the quality rank (Pearson)	0.880
Split-half correlation of the quality rank (Spearman)	0.880
Full-corpus reliability (Spearman-Brown)	0.936
Implied ceiling on attenuation of a profile amplitude	$\times$ 1.068
Pairwise correlation across construction variants (min)	0.076
Pairwise correlation across construction variants (median)	0.504
Pairwise correlation across construction variants (max)	0.988

Notes: The Table reports the measurement reliability of the quality rank on the estimation sample (5,344,793 applications complete on all rank columns). The split-half correlation scores every application on two independent halves of the text corpus at fixed construction and correlates the two within-cell percentile ranks; the Spearman-Brown adjustment $2r/(1+r)$ converts it to the reliability of the full-corpus rank. Under classical measurement error, a binned profile over percentiles of the rank is attenuated by the square root of the reliability, so the stated factor, the inverse of the reliability, is a conservative ceiling on how far the true amplitude of the complier-quality profile exceeds the observed one. The split varies only the comparison documents, so the reliability covers noise from them and not how well an application's own text records its quality. The variant correlations are pairwise Pearson correlations among the ten retained construction variants of the quality measure; their dispersion is a construction choice rather than noise, and is the object priced by the reference-class score of Table I.5.

Table I.7: *Forward similarity and examiner leniency at the art-unit level*

	β	s.e.	p	N
Cross-section of art-unit means (raw units)	+0.038	0.004	0.000	701
Cross-section of art-unit means (SD units)	+0.321	0.031	0.000	701
Within unit: art unit $\times$ year, unit and year FE (SD units)	−0.054	0.037	0.153	9,896

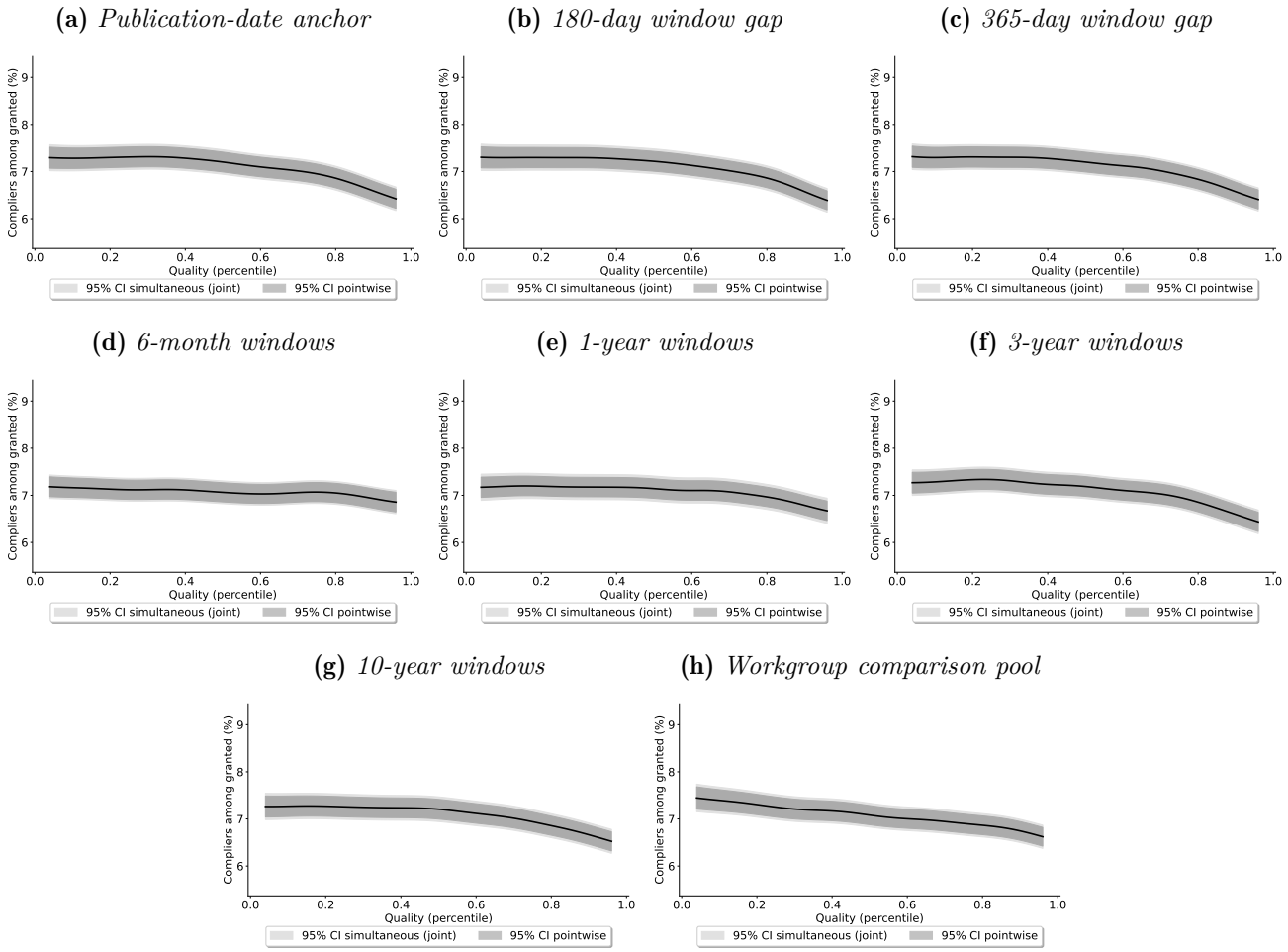
Notes: The Table reports regressions of mean forward similarity (the impact component of the quality measure) on mean examiner leniency. The first two rows collapse to art-unit means over the 2005–2020 estimation window and regress the level of one on the other, weighted by art-unit size, with heteroskedasticity-robust standard errors; the positive association reflects technology composition across fields. The third row is the exclusion-relevant design quoted in the text: art unit $\times$ publication year means with art-unit and year fixed effects, weighted by cell size, standard errors clustered on art unit, so the coefficient measures whether a unit's forward trajectory moves with the leniency of the examiners deciding in it. SD units standardize both variables.

Construction robustness of the complier profile. The decline of the complier share in quality is not specific to how the measure is built. Figure I.2 re-estimates the profile of Figure 4 with the quality measure rebuilt under eight alternative construction choices. The first panel anchors the text comparison on the publication date instead of the filing date. Each of the other seven changes one further dimension of that publication-anchored construction: the length of the text-comparison windows, the gap between the backward and forward windows, or the level of the comparison pool. The estimator and the instrument are identical across panels. The samples

differ only by the applications whose comparison windows are empty under a construction, at most a quarter of a percent of the sample. The share falls between the 10th and the 90th percentile under every construction, by between 3.2% with six-month windows and 10.0% with a 365-day gap, against 7.6% for the headline measure. The simultaneous test rejects a flat profile in all eight, with $p < 0.001$. Two further constructions compare each application only with its closest documents instead of the average of its art unit. The first divides the mean similarity to the ten nearest documents in the forward window by that to the ten nearest in the backward window. The second follows Kelly et al. (2021) and sums, in each window, only the similarities above a cutoff, then takes the ratio of the two sums. The cutoff is what makes these sums local. Without it, the sum of similarities over a window is proportional to the similarity to the window's average text, which is the comparison the headline measure makes. Kelly et al. (2021) set similarities below 0.05 to zero, which affects 93.4% of the pairs in their data.[37] Similarities between embeddings lie on a different scale, so I match the share of pairs instead of the value. The cutoff is a similarity of 0.81, which 6.6% of two million randomly drawn pairs of documents exceed. Both constructions use the art units and filing-date windows of the headline measure and exclude the application's own continuation family. The share falls between the 10th and the 90th percentile by 8.1% under the nearest neighbours and by 6.1% under the thresholded sums, and the simultaneous test rejects a flat profile in both, with $p < 0.001$.

[37] Kelly et al. (2021, Online Appendix, p. 5). They add that cutoffs between 0.05 and 0.10 give quantitatively indistinguishable results.

Figure I.2: *Construction robustness: compliers among granted patents by quality, across constructions of the quality measure*



Notes: Each panel plots the share of compliers among granted patents by quality percentile, from the estimator of Figure 4, with the quality measure rebuilt under one alternative construction choice per panel. The instrument is built once and shared across panels, so only the quality ranking varies. Each panel is scaled so that its grant-weighted average equals the aggregate share. Shaded areas are 95% simultaneous (light) and pointwise (dark) confidence bands, with the scaling factor held fixed. All panels share the vertical axis of Figure 4. The simultaneous test rejects a flat profile in every panel, with $p < 0.001$.

J Scope measure: construction robustness and normalization dependence

Filed scope is measured in one specific way: the average number of words per claim (Section 3.2), expressed as a within-year percentile rank. This appendix asks two questions about that choice. The first is construction: whether the results depend on which part of the claims text the measure reads. The second is normalization: whether quantities that sum scope across applications depend on the cardinal scale behind the rank.

Alternative claim measures. Two in-family alternatives are available for every application: the length of the first claim, the object that Kuhn and Thompson (2019) measure, and the total length of the claims. They correlate with the headline measure at Spearman 0.63 and 0.47, so they are different orderings of the same concept rather than relabelings, and where the samples overlap the corpus first-claim length agrees with the independently parsed first-claim word counts of Kuhn and Thompson (2019) at a rank correlation of 0.997. Table J.1 rebuilds the broadness axis from each alternative on the estimation sample. The decline in the grant rate from the narrowest to the broadest ventile is 19–22 percentage points under all three measures, with near-identical profiles (Panel J.1a).

Panel J.1b and Table J.2 carry the test to the paper's headline object. The share of grants decided by the examiner assignment is re-estimated with the estimator of Section 5.3 and one instrument, built once and shared by every row, so the rows differ only in the scope measure and, for the first-claim rows, in the sample. Under every measure the share rises from the narrowest to the broadest filings. Between the 10th and the 90th percentile of each measure it rises from 5.9% to 8.6% under the headline axis, from 6.0% to 8.5% under first-claim length, and from 5.7% to 8.5% under total claim length. No alternative curve departs from the headline curve by more than 0.31 percentage points, about a tenth of the rise the measures agree on. The first-claim curve stays inside the headline curve's simultaneous band everywhere, and the total-length curve at 81% of the grid points. Claim count is deliberately not in this family: applications with more claims write shorter ones, the grant rate is nearly flat in the count, and it enters the decomposition regressions as a control rather than as a scope measure.

Table J.1: *Scope measure: alternative claim measures, agreement and grant rate*

	Spearman vs. headline	Grant rate, narrowest 5%	Grant rate, broadest 5%	Decline (pp)	N
Average claim length (headline)	1.00	76.9	57.7	19.2	5,363,832
same, on the first-claim sample	1.00	76.7	56.9	19.9	4,880,897
First-claim length	0.63	79.0	57.4	21.6	4,880,897
Total claim length	0.47	78.8	56.9	21.9	5,363,832

Notes: Estimation sample (design-complete rows, publication years 2005–2020), $N = 5,363,832$. Each row measures filed scope from a different part of the claims text and rebuilds the broadness axis with the estimation convention: the within-publication-year percentile of the negated measure, so higher means broader. Grant-rate columns are mean grant rates in the bottom and top broadness ventiles. The first-claim measure excludes the 9.0% of applications whose published first claim holds at most 10 words, dominated by the "(canceled)" notation of continuations whose original first claim was withdrawn during prosecution; the headline measure re-estimated on those same rows is the second row. Claim count is deliberately not a member of this family: applications with more claims write shorter ones (Spearman +0.26 between claim count and word-per-claim broadness), yet the grant rate is nearly flat in claim count (Spearman +0.04) – it counts how many boundaries an application draws rather than how much language each one stakes out, and it enters the decomposition regressions as a control instead. At-issuance measures are excluded as post-decision text. Where the samples overlap ($n = 671,268$ granted, patent-matched), the corpus first-claim length agrees with the independent Kuhn–Thompson first-claim word counts at filing at a rank correlation of 1.00.

Figure J.1: *Grant rate and compliers among granted patents by filed scope, under alternative claim measures*

(a) *Grant rate (%)* **(b)** *Compliers among granted patents (%)*

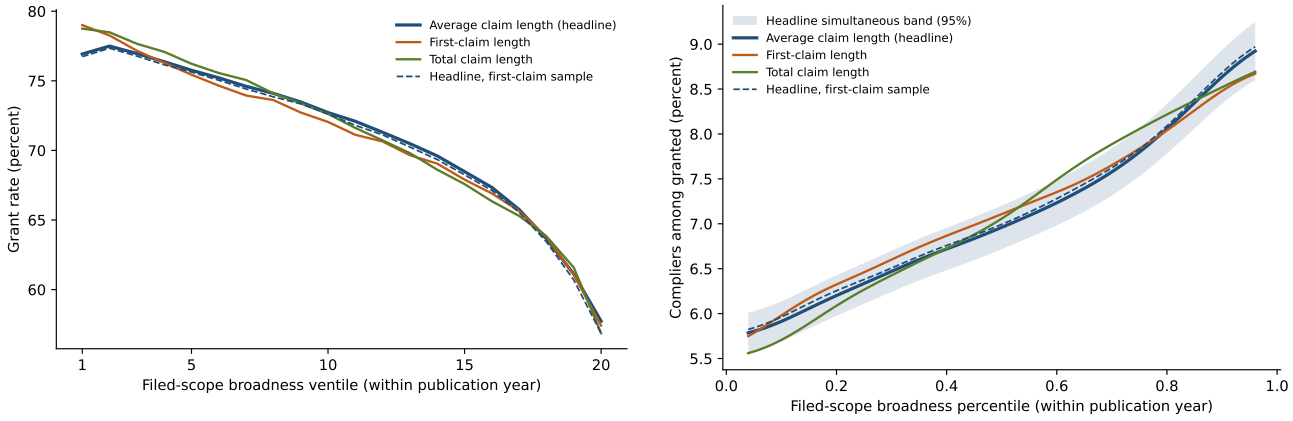


Table J.2: *Complier-scope gradient under alternative claim measures*

	Grants decided by the assignment (%)					
	10th percentile	90th percentile	Rise (pp)	All grants (%)	Inside headline band (%)	N
Average claim length (headline)	5.9	8.6	+2.7	7.1	100	5,360,465
same, on the first-claim sample	6.0	8.7	+2.7	7.1	100	4,876,823
First-claim length	6.0	8.5	+2.5	7.1	100	4,876,823
Total claim length	5.7	8.5	+2.8	7.1	81	5,359,453

Notes: Share of granted patents that the average examiner the application could have been assigned to would have refused, on the smoothed curve at the 10th and the 90th percentile of each scope measure, estimated as in Section 5.3. The instrument is built once and shared by every row, so the rows differ only in the scope measure and, for the first-claim rows, in the sample. Each curve is scaled so that its grant-weighted average equals the share among all grants on its own rows, reported in the fifth column. The first-claim rows exclude applications whose published first claim holds at most ten words, dominated by the "(canceled)" notation of continuations, and the second row shows this exclusion alone. "Inside headline band" is the share of the grid points at which the curve lies within the headline row's simultaneous band. The joint test of a flat profile rejects at $p < 0.001$ for every row.

Normalization dependence. Filed and granted scope enter as within-year percentile ranks of average claim length. Appendix N notes that quantities which *sum* scope across applications inherit a normalization that the ordering itself does not pin down. The rest of this appendix separates the part of the paper that is immune from the part that is exposed, and measures the exposure.

The immune part is larger than the objection suggests. A rank is unchanged by any strictly monotone transformation of claim length, so any result that uses only the ordering is exactly invariant to the choice of scale. That covers which applications fall in which equal-mass scope bin, the complier share estimated at each bin and therefore the shape of the profile, the ordinal reclassification counts, and any ratio taken across bins. It is not a robustness finding but an

algebraic one, and it requires no re-estimation. The exposed part is the class of quantities that aggregate scope in levels: the granted-scope budget and the changes in mean requested and granted scope under the counterfactuals of Section 7.7. A sum of ranks carries no cardinal content, so these are conditional on the scale in a way the composition results are not.

Table J.3 quantifies that conditionality. It re-expresses the same scope ordering on a family of positive scales, $1/w^{\gamma}$ for $\gamma \in \{\frac{1}{2}, 1, 2\}$ in the average claim length w, every member of which agrees exactly on which patent is broader and differs only in by how much, and reports the share of the granted-scope budget carried by each rank decile. Deciles are formed on the rank, so the groups are held fixed across columns and only the weighting moves, which isolates normalization from composition.

The shares move substantially. The broadest decile carries between 14.5% and 34.9% of the budget depending only on the scale, the largest swing in any single decile's share is 20.4 percentage points, and the implied concentration, the broadest decile's share relative to the narrowest, ranges from 2.4 to 32.7. The percentile rank is not an extreme member of the family: its decile profile lies closest to $1/w$, with a maximum deviation of 3.3 percentage points, so the reported magnitudes do not appear to be inflated by the normalization. They are nonetheless conditional on it, and the changes in mean requested and granted scope reported in Section 7.7 should be read that way.

Two limits of this exercise. It holds the allocation fixed and varies only the scale, so it bounds the accounting component of normalization dependence and not the behavioural one: filed scope responds to a revision cost defined on the rank scale, and a complete answer would re-estimate the model under the alternative scale. And the rank embeds a second normalization that the raw scales do not, since it is taken within publication year, so two patents with identical claim length in different years receive different ranks. That choice is held fixed throughout the table, and what varies is the cardinal spacing alone.

Table J.3: *Normalization dependence of the granted-scope budget*

Rank decile (1 = narrowest)	rank	$1/w^{1/2}$	$1/w$	$1/w^2$
	share of the granted-scope budget			
1	0.9%	6.0%	3.5%	1.1%
2	2.7%	7.6%	5.5%	2.4%
3	4.5%	8.4%	6.7%	3.6%
4	6.4%	9.0%	7.7%	4.8%
5	8.4%	9.6%	8.7%	6.1%
6	10.5%	10.2%	9.8%	7.7%
7	12.8%	10.8%	11.0%	9.7%
8	15.2%	11.5%	12.4%	12.5%
9	17.8%	12.4%	14.5%	17.1%
10	20.7%	14.5%	20.2%	34.9%
Broadest/narrowest	22.1	2.4	5.7	32.7

Notes: Each column re-expresses the same scope ordering on a different positive scale and reports what share of the total granted-scope budget each decile then carries. w is the average claim length in words. The three inverse-power scales are strictly decreasing in w and agree exactly on which patent is broader. The rank is computed within publication year, so it agrees with them within a year but can reorder cross-year pairs, and the four columns share an ordering up to that cross-year component rather than exactly. Deciles are formed on the rank, so the groups are held fixed across columns and only the weighting moves, which isolates normalization from composition up to the same cross-year reordering. Budget shares are invariant to multiplying a scale by a positive constant, so the inverse-power columns need no further normalization. Sample: $3,771,826$ granted patents in the core sample. The table bounds the accounting component of normalization dependence, holding the allocation fixed, and it does not bound the behavioural component, since filed scope responds to a cost defined on the rank scale.

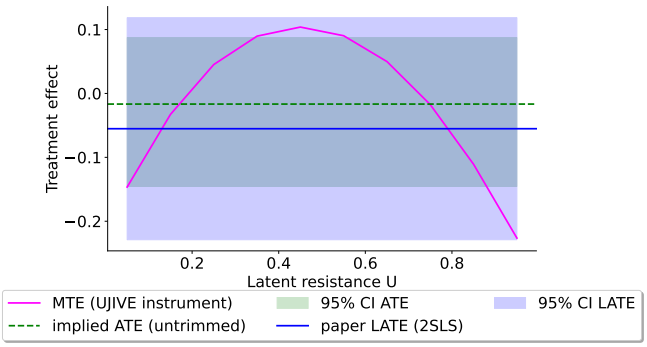
K What the examiner assignment identifies: marginal treatment effect curves

The composition of Section 6 and the mechanism of Section 6.3 are statements about which applications owe their grant to the examiner assignment. This appendix draws out the consequence for the estimates the examiner assignment has produced: if the complier pool is not representative, those estimates differ systematically from the average effects.

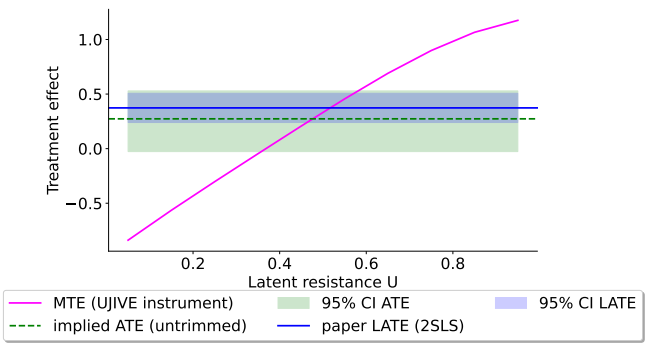
In a judge fixed-effects IV design, the LATE is a weighted average of treatment effects over the complier subpopulation, so when that subpopulation tilts toward broad-claiming applications of unremarkable quality, the LATE weights treatment effects differently from the population average. Which way the ATE and LATE differ depends on how the treatment effect varies across the resistance distribution, and that is an object to be estimated, not a corollary of the composition. I estimate it with marginal treatment effect (MTE) curves (Heckman and Vytlacil, 2007), whose propensity score $P(X, Q, Z) = \Pr(D = 1 \mid X, Q, Z)$ incorporates quality, filed scope, and examiner leniency jointly, so the correction accounts for all three dimensions at once rather than one at a time. The estimating equation, the implementation, the sample matching, and the joint test of

Figure K.2: *Marginal treatment effect curves*

(a) *log(Scientific publications) – Sampat and Williams (2019)*



(b) *log(Follow-on patent applications) – Farre-Mensa et al. (2020)*



Notes: Each panel plots the MTE curve (Equation 33) for the primary outcome of the respective paper, estimated on the variation of the examiner-dummy UJIVE instrument of the pass-through estimator of Appendix E: the leave-out deviation $\hat{D}^A_{-i} - \hat{D}^W_{-i}$ enters the selection probit as the instrument, and the sample is the connected set on which the leave-out fit is defined. For Sampat and Williams (2019) the examiner-leniency design is shown. The horizontal axis is the unobserved resistance term $U \in [0, 1]$. Higher values correspond to applications that require a more lenient examiner to be granted. The magenta curve is the MTE. The dashed green line is the ATE obtained by integrating that curve over the full support of U, with its 95% bootstrap confidence interval shaded in green. The solid blue line is the LATE delivered by each paper's own estimator, its published 2SLS specification re-estimated on exactly this sample so that the comparison is not a sample-composition artifact, with its 95% cluster-robust confidence interval shaded in blue. The two objects are therefore an average effect and the complier-weighted effect the original design reports, on common ground. The ATE integrated over the common-support window only is reported in the text rather than plotted. The two shaded bands are constructed by different routines and share no bootstrap draws, so their overlap is not a test that the two estimands are equal. The formal test, which re-estimates both on common resamples, is reported in Appendix Tables L.9 and L.10. Additional MTE outcomes are reported in Appendix Figure K.4.

the identifying assumptions follow below.

One caveat governs how far the results can be read. The design-based sections of the paper use the random assignment of examiners and little else. Recovering a marginal treatment effect requires a selection model, whose substantive assumptions are stated below before the estimates are read. I therefore state the conclusion as a claim about *direction* and about consistency across settings, not about magnitude.

Figure K.2 reports the curves for an outcome of each paper, with the underlying regressions in Appendix L.3. In Sampat and Williams (2019), the curve is hump-shaped in resistance, while in Farre-Mensa et al. (2020) it rises monotonically, implying negative selection on potential gains. What the two share is the sign of the always-taker region. Applications granted under nearly any examiner carry negative estimated effects in both, so in each setting the LATE is identified on the part of the resistance distribution where a grant is worth most. In the Farre-Mensa et al. (2020) setting the integrated ATE lies below the LATE in every outcome and under both integration windows, a consequence of where the complier mass sits rather than of the slope of the curve. In the Sampat and Williams (2019) setting the two estimands are statistically indistinguishable, with

small gaps of unstable sign, so the ordering claim rests on the first setting alone.

The strength of that pattern should not be overstated, and three ways of counting it give three different answers. Re-estimating both estimands on 5,000 common bootstrap resamples to test the difference between LATE and ATE, the gap is individually significant in two of the eight outcomes under bias-corrected and accelerated intervals, both of them citation outcomes of Farre-Mensa et al. (2020), and one survives (Romano and Wolf, 2005, 2016) familywise control. For both outcomes reported here, the difference is not statistically distinguishable from zero, so the ordering rests on point estimates alone. A uniform sign across one setting's outcomes is what a systematically non-representative complier pool predicts and what an unsystematic one does not. The Sampat and Williams (2019) gaps are small and of unstable sign, which is what a sample an order of magnitude smaller supports. The pattern suggests systematic overstatement where there is power to detect it, without establishing that any one published LATE is significantly overstated. I therefore rest the argument on the pattern rather than on any single test. The remainder of this appendix describes the two curves outcome by outcome, reports the full test and the diagnostics behind it, and sets out how far the finding generalizes beyond these two papers.

Estimating treatment effects beyond the LATE requires a selection model of the patent review process. For the settings of Sampat and Williams (2019) and Farre-Mensa et al. (2020), I estimate marginal treatment effect curves via local instrumental variables, as implemented in the `localIV` package of Zhou and Xie (2019). Following the standard treatment of the generalized Roy model and instrumental variables (for more details see Heckman and Vytlacil (2007)), the marginal treatment effect (MTE) curve is defined, adjusted to my previously used notation, as:

$$MTE(x,u) = \left.\frac{\partial\, \mathbb{E}[Y|X=x, Q=q, P(X,Q,Z)=p]}{\partial p}\right|_{p=u} \tag{33}$$

Three terms are new relative to the main text. The first is the propensity score $P(X,Q,Z) = \Pr(D=1 \mid X,Q,Z)$, which gives the probability of patent grant conditional on controls X, quality Q, and leniency Z. The second is the latent resistance term u, and the third is an outcome variable Y that is affected by the patent grant D. The MTE at u is the treatment effect for applications that are exactly indifferent at propensity score $p = u$: applications with low resistance are granted by nearly any examiner, while applications with high resistance are granted only by an unusually lenient examiner.

Equation (33) carries assumptions that quasi-random assignment does not supply, and they should be stated first. Three are substantive. First, a scalar latent index: resistance is a single unobserved u that orders all applications, so that any two applications with the same propensity score are exchangeable in their selection behavior. The threshold rule of Section 5.1 delivers exactly such an index, with $a(q,s)+l$ crossing a threshold, which is why the specification is a natural one

here, but it is imposed rather than tested, and it rules out selection on dimensions orthogonal to that index. Second, a parametric propensity specification: the score is estimated from a probit in quality, controls, and the instrument, so the MTE inherits any misspecification of that first step. Third, and most consequentially, *support*. Equation (33) is defined only where the propensity score has mass, and the estimated score does not reach either extreme in these samples. The full-support integral therefore extrapolates the fitted curve into regions the design never populates, and no quantity of additional data of this kind repairs it, since the examiner assignment simply does not move applications that far. I report the common-support integral alongside the full-support one throughout for this reason, and read the tails of every MTE figure as extrapolation.

This is the basis for the direction-only reading stated at the outset: the common-support comparison leans far less on the tails than the full-support level does. The direction is still an empirical regularity under the maintained specification rather than a support-robust implication, because both estimands weight the whole fitted curve and unsupported tails can in principle move the sign as well as the level.

One further feature of the specification makes the correction a joint one. The propensity score $P(X, Q, Z)$ incorporates quality, filed scope, and examiner leniency together, all three determinants of the grant decision named in Section 5.1, so the observed selection those determinants carry is accounted for jointly. The resistance index u ranks applications by their unobserved resistance conditional on the score, not by a position in the (q, s, l) space itself. The complier composition over quality and filed scope implies non-representativeness along the observed dimensions, which the MTE weights jointly rather than one at a time.

The two settings in detail. Sampat and Williams (2019) study whether patents on human genes blocked follow-on research, finding no significant effect on subsequent scientific publications. Panel K.2a of Figure K.2 shows that the MTE curve for their primary outcome is hump-shaped in resistance: the applications granted under nearly any examiner and those granted only under an unusually lenient one both show negative estimated effects (-0.31 and -0.56 at the fifth and 95th percentiles of u), while the intermediate-resistance applications that carry the most weight in the LATE show positive ones, peaking at 0.10. Their own specification, re-estimated on this sample, returns a LATE of -0.055 (cluster-robust s.e. 0.089), a null consistent with the original finding. Integrating the MTE over the full support of u yields an ATE of -0.017, and over the common-support window $+0.015$. On 5,000 common bootstrap resamples the ATE-minus-LATE difference is $+0.039$ (BCa 95% CI $[-0.109, 0.320]$) over the full support and $+0.070$ ($[-0.051, 0.380]$) over the common-support window (Table L.9). This setting therefore delivers no ordering: the gap sits inside its interval under either integral and its sign is not stable across the setting's outcomes. What it contributes is the shape, the hump in resistance, together with the caveat that both tails of

the fitted curve lie where the propensity score has little support, so I read neither the full-support level nor its sign as robust.

Farre-Mensa et al. (2020) find that startups assigned a lenient examiner experience 55% larger employment and 80% higher sales growth. Panel K.2b shows that the MTE curve for follow-on patent applications rises monotonically in resistance, from -0.83 at the 5th percentile of u to 1.18 at the ninety-fifth: *negative selection on potential gains* in its textbook form, in which the applications granted most readily are those that benefit least from the grant. Their own specification, re-estimated on this sample, returns a LATE of 0.369 ([0.236, 0.502]), against an integrated ATE of 0.269 ([0.021, 0.524]), or 0.318 ([0.082, 0.542]) over the common-support window. Here the ordering does not follow from the slope of the curve, since an upward-sloping MTE is consistent with an ATE on either side of the LATE, but from where the complier mass sits relative to it: the large negative values belong to the always-taker region at low resistance, which the complier-weighted LATE does not reflect. The ordering conclusion therefore rests on the Farre-Mensa et al. (2020) setting. Sampat and Williams (2019) contributes the curve rather than the ordering.

The joint bootstrap. All p-values reported in this appendix are from the bias-corrected and accelerated (BCa) interval of the ATE-LATE difference, computed on 5,000 common resamples, with both estimands conditioned on the same covariates, the control set of the MTE outcome equation, so that their difference is a well-defined object. For Farre-Mensa et al. (2020) that matched conditioning shifts the LATE by at most 0.005 relative to the published specification the figures plot, and for Sampat and Williams (2019) the two coincide. For the two primary outcomes the difference is not statistically significant: -0.104 ($p = 0.521$) for Farre-Mensa et al. (2020) and $+0.039$ ($p = 0.568$) for Sampat and Williams (2019). It is significant, at the five percent level and one outcome at a time, in two of the eight, both citation outcomes of Farre-Mensa et al. (2020): total citations within five years (-0.344, $p = 0.029$) and mean citations per application (-0.235, $p = 0.003$, and -0.203, $p = 0.008$, after common-support trimming). Only one of the two, mean citations per application, survives the Romano-Wolf familywise adjustment across the outcomes of its own table ($p = 0.017$). The other sits just above the threshold once the family is in view ($p = 0.088$). Tables L.9 and L.10 report the test for every outcome together with the diagnostics behind it. Those resamples are drawn over applications rather than over the art unit $\times$ year cells the standard errors elsewhere cluster on, and Table L.11 shows that resampling those cells instead reproduces the analytic cluster-robust standard error of the LATE and leaves every verdict unchanged. That check covers the Farre-Mensa et al. (2020) family, which is where every individually significant difference sits. No Sampat and Williams (2019) outcome is individually significant under either integral, so no verdict in that setting rests on the resampling unit.

How far this travels. The findings plausibly extend beyond the two papers studied here, though the evidence is not equally strong in both dimensions. The composition documented in Section 6 is a property of a threshold grant rule in which scope is penalized, so any examiner-leniency instrument applied to USPTO prosecution should inherit some version of it. The *scope* tilt is the well-supported half: it is the steeper gradient, it survives the caseload-composition purge, and it is what the mechanism of Section 6.3 predicts most sharply. The quality tilt is milder and its complier mean is close to the population's, so I would not assert that every examiner-leniency LATE is materially mis-weighted on quality. This stands in contrast to judge fixed-effects designs in the economics of crime (e.g., Agan et al., 2023), where defendants do not make strategic pre-treatment choices that select a non-representative complier pool. In patent prosecution, applicants' strategic filing decisions generate the non-representative complier pool, and the MTE framework is the natural tool for recovering treatment effects informative about the average applicant rather than the marginal one.

Estimation samples. Implementation follows the original papers' specifications, augmented with my quality measure. Because the samples that reach estimation are materially smaller than the samples I match, I report the full ladder rather than its first rung. For Farre-Mensa et al. (2020), the replication file contains 34,435 rows, 34,418 of them distinct applications. Matching to my data without the publication-year restriction of Section 3.1 recovers 28,395. Imposing that restriction, which the rest of the paper applies throughout, leaves 16,683 and drops 11,712 applications (41.2%), all of them published before 2005. The subsequent outcome and covariate requirements drop essentially nothing. The binding step is instead the estimator's: the leave-one-out examiner instrument is defined only on the largest connected component of the examiner-by-cell bipartite graph, which retains 10,289 applications (38.4% dropped) across 2,196 examiners and 873 art unit $\times$ year clusters. The leverage and scope filters remove no further observations. It is 10,289, not 28,395, that the estimates rest on. Quoting the larger figure as the sample would start the ladder in one population and finish it in another.

For Sampat and Williams (2019) the ladder is shorter and cut in a different place: 1,545 distinct applications, 1,450 matched, and 1,196 reaching the examiner-design estimates (1,108 for the gene-level design), with the loss falling on the covariate requirement rather than on the connected set. This setting is not subject to the publication-year restriction, deliberately: its gene patents are filed between 2002 and 2006, and imposing the restriction would discard most of the sample. One consequence bears on how the two replications are compared: my quality measure is a percentile rank, and it is therefore computed over different populations in the two settings. Columns from the two designs are indexed by deciles of different underlying distributions and should not be read as though they shared a scale. For Sampat and Williams (2019), the primary outcome is follow-on innovation in 2011-2012, measured by the logarithm of scientific publications

(Table 2, Panel A in the original paper). For Farre-Mensa et al. (2020), it is the logarithm of follow-on patent applications. In each setting, I re-estimate the original specification including my measure of quality as an additional control variable, estimate the propensity score from quality, controls, and examiner leniency, and trace out the MTE curve via Equation 33. The resulting curves for the primary outcomes are reported in Figure K.2.

The subsamples corroborate the *level* of the full-sample complier share but are too small to resolve its gradients. Figure K.3 reports the share of compliers among granted patents for both papers, estimated with the same examiner-dummy modified-outcome estimator used for the full sample in Figure 4a. In the Farre-Mensa et al. (2020) subsample the profile hovers near 5% along both the quality and the filed-scope axis, close to the 4.7% full-sample share, but is flat to within its confidence bands in each. In the Sampat and Williams (2019) subsample the quality profile rises from roughly 4.5% to an interior peak near 11% before settling near 7%, although the joint test of flatness does not reject it ($p = 0.071$). The scope profile does reject flatness ($p = 0.000$), but its shape is dominated by a region of the scope distribution in which the gene-patent subsample contains almost no applications, visible as the break in the fitted curve. The bands are wide enough at these sample sizes that the point estimates should not be read at face value. What these subsamples establish is that the magnitude of the share owed to the examiner assignment documented in Section 6 is not an artifact of the full-sample construction. They neither confirm nor contradict its quality and scope gradients, which are identified off the full sample.

The subsamples also permit a formal test of the identifying assumptions that is infeasible in my main first-stage-only analysis. Because both settings provide a second-stage outcome, I can apply the joint test for the exclusion restriction and monotonicity of Frandsen et al. (2023). The respective p-values of the test are 1.000 for the setting of Sampat and Williams (2019) and 0.570 for Farre-Mensa et al. (2020). In both settings, the tested implications of exclusion and monotonicity thus survive. The scalar latent-index structure, the parametric propensity specification, and the support conditions the MTE integration additionally requires are not what this test probes.

The MTE findings are robust to alternative outcome measures. Figure K.4 reports MTE curves for any follow-on scientific publication in the subsample of Sampat and Williams (2019), and for the logarithm of application and patent citations after five years, application citations after five years, and the logarithm of granted patents in the subsample of Farre-Mensa et al. (2020). Across these alternative outcomes, the qualitative pattern of the primary estimates carries over: the shapes of the curves, monotonically upward-sloping in the Farre-Mensa et al. (2020) setting and hump-shaped in Sampat and Williams (2019), mirror the primary outcomes in Figure K.2, and the integrated ATEs lie below the LATEs delivered by each paper's own published specification in every Farre-Mensa et al. (2020) outcome and for both the full-support and the common-support integral, while the Sampat and Williams (2019) gaps remain inside their intervals with unstable

sign. Appendix Tables L.9 and L.10 test the comparison formally, outcome by outcome, by re-estimating the ATE and the LATE on common bootstrap resamples and under a common control set. The difference is individually significant for the two citation outcomes of Farre-Mensa et al. (2020) and nowhere else. The substantive interpretation of these results is developed at the start of this appendix.

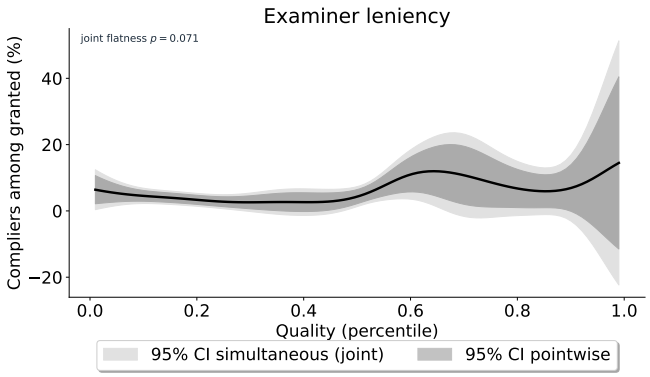
Which Sampat and Williams estimand is on display. The LATE I report for that setting is their specification re-estimated *unweighted*. Their published estimand is not. Their frame is gene-level – 293,652 gene rows over 1,545 applications, a mean of 190 genes per application and a maximum of 3,103, and the published estimates weight by that count. With dispersion of that size the weights bite. Re-estimating the identical specification with them moves every LATE by up to one and a half unweighted standard errors and carries four of the six across zero:

Design	Outcome	Unweighted	Gene-weighted	Shift (SE)
Examiner	log follow-on publications	−0.0552 (0.0887)	−0.0004 (0.0305)	+0.62
Examiner	Any follow-on publication	−0.0508 (0.0868)	+0.0131 (0.0224)	+0.74
Examiner	Any follow-on test	+0.0254 (0.0411)	−0.0164 (0.0342)	−1.02
Gene	log follow-on publications	+0.0243 (0.0252)	−0.0135 (0.0205)	−1.50
Gene	Any follow-on publication	+0.0263 (0.0225)	−0.0070 (0.0159)	−1.48
Gene	Any follow-on test	−0.0026 (0.0237)	−0.0159 (0.0217)	−0.56

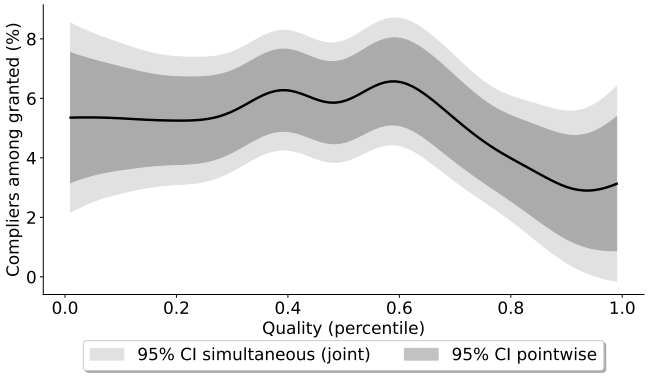
Every estimate, weighted or not, is statistically indistinguishable from zero, and the shifts run to between 0.6 and 1.5 unweighted standard errors. Two conclusions follow, and they point in opposite directions. The substantive reading is untouched: this setting delivers a null on follow-on innovation whichever estimand is used, which is what the comparison above rests on. But the *signs* of the point estimates are not robust to the weighting choice, so no argument here should lean on them, and the tables label which estimand they display rather than leaving it implicit.

Figure K.3: *Conditional share of patents granted due to the examiner*

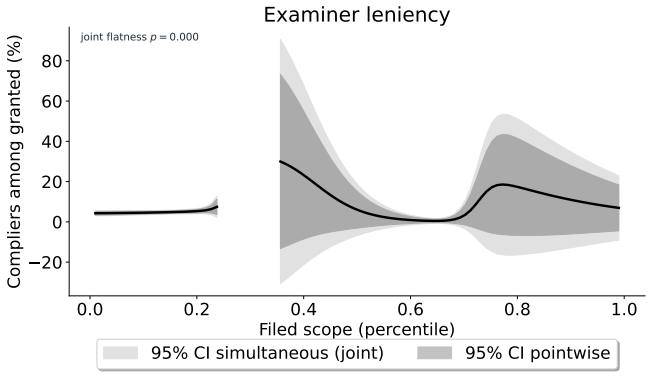
(a) *Sampat and Williams (2019) – by quality*



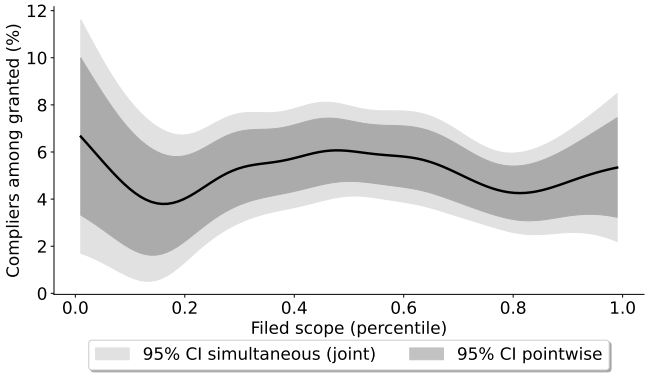
(b) *Farre-Mensa et al. (2020) – by quality*



(c) *Sampat and Williams (2019) – by filed scope*

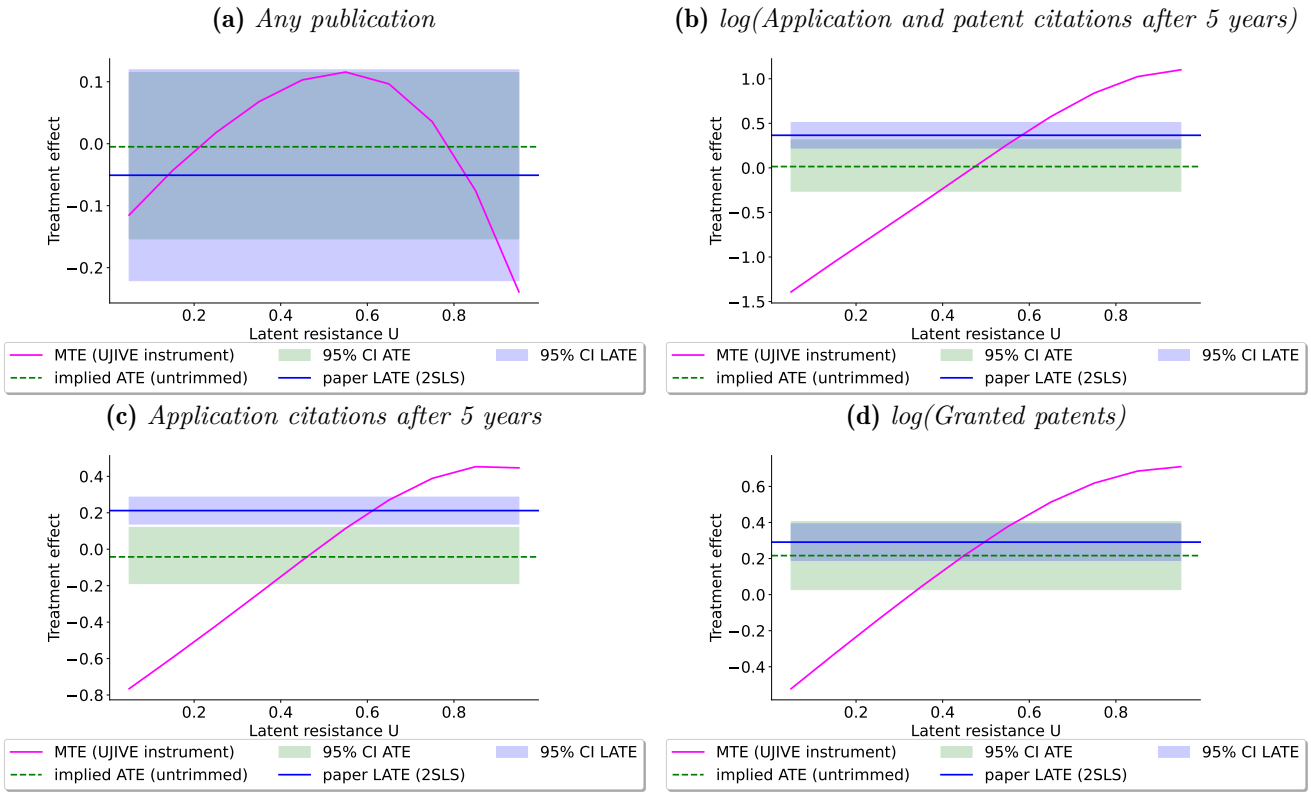


(d) *Farre-Mensa et al. (2020) – by filed scope*



Notes: The Figure plots the share of compliers among granted patents, $\phi_c^{D=1}$, across the percentiles of quality (Panels K.3a and K.3b) and of filed scope (Panels K.3c and K.3d), for the sample of Sampat and Williams (2019) in the left column and for the sample of Farre-Mensa et al. (2020) in the right column. The estimator is the same examiner-dummy UJIVE modified-outcome estimator used for the full sample in Figure 4, applied here within each replication subsample. For Sampat and Williams (2019) the examiner-leniency design is shown. The shaded areas show 95% pointwise and 95% simultaneous (multiplier bootstrap) confidence bands. Where reported, the annotated p-value is that of the joint sup-test of the null that the curve is constant. Note the difference in vertical scale across panels: the replication subsamples are two to three orders of magnitude smaller than the full sample, so the bands are correspondingly wide and the point estimates should not be read at face value.

Figure K.4: *Additional MTE curves*



Notes: The Figure plots the results of estimating Equation 33 for any follow-on scientific publication in the subsample of Sampat and Williams (2019) in Panel K.4a and the results on follow-on innovation in Farre-Mensa et al. (2020) in Panel K.4b, K.4c, and K.4d. Estimation, sample, and the meaning of each plotted series are as in Figure K.2: the selection probit is instrumented with the leave-out examiner-dummy UJIVE deviation on the connected-set sample, the magenta curve is the MTE, the dashed green line the ATE integrated over the full support of U, and the solid blue line the LATE from each paper's own published 2SLS specification re-estimated on this same sample. Shaded bands are the corresponding 95% confidence intervals. In every Farre-Mensa et al. (2020) outcome shown, as in Figure K.2, the integrated ATE lies below that LATE. The bands are not comparably constructed and their overlap does not test equality of the two estimands. Appendix Tables L.9 and L.10 report the joint-bootstrap test of that difference for each outcome shown here.

L Regression tables underlying the figures

Several exhibits in the paper plot the output of a regression rather than a raw moment of the data. This appendix reports the underlying estimates in table form, so that every plotted curve can be read as coefficients with standard errors. The correspondence is one-to-one and stated in each table's note.

Two families of exhibit are deliberately absent. The by-revision panels (Figures 3 and D.3) and the leniency density (Figure D.4) plot sample means and a kernel density, not fitted coefficients. The model's fit and counterfactual exhibits compare simulated to empirical moments and are reported in Appendix M.

L.1 Conditional correlations

The binned scatter plots of Section 3.4 and Section 4 overlay a quadratic fit of the residualized outcome on the quality measure. Tables L.1 and L.2 report that fit: the linear and squared quality terms separately, and the net predicted change across the plotted window, which is the statistic annotated on each panel.

Table L.1: *Conditional correlations: validation against external quality proxies*

	(1) Citations	(2) Renewals	(3) Kogan et al.	(4) Kelly et al.
Quality	0.0532***	0.0204***	0.0006	0.0298***
	(0.0016)	(0.0012)	(0.0014)	(0.0013)
Quality^2	0.0173***	−0.0054***	0.0050***	0.0064***
	(0.0011)	(0.0010)	(0.0012)	(0.0012)
Net effect ($q: 0 \to +1$)	0.0705***	0.0149***	0.0056***	0.0363***
	(0.0023)	(0.0014)	(0.0019)	(0.0017)
Controls	✓	✓	✓	✓
Art unit × year demeaning	✓	✓	✓	✓
Observations	3,812,246	3,805,365	1,216,203	1,740,472
R^2	0.030	0.010	0.013	0.062

Notes: Quadratic specification $y = \beta_1 q + \beta_2 q^2 + \gamma' X + \varepsilon$ estimated by OLS with CRV1 (ind_year_fe) standard errors in parentheses. *Net effect* is $\hat{\beta}_1(q_{hi} - q_{lo}) + \hat{\beta}_2(q_{hi}^2 - q_{lo}^2)$ over $q_{lo} = 0$ to $q_{hi} = 1$, with a delta-method standard error from the CRV1 (ind_year_fe) covariance of $(\hat{\beta}_1, \hat{\beta}_2)$. Citations count the citations a granted patent receives from later US patents, matched on the patent number. Renewals are the maintenance fees paid minus the fee decisions observable by June 2023. The measures of Kogan et al. (2017) and Kelly et al. (2021) are matched on the patent number. Outcomes and the quality measure are standardized and demeaned within art unit × year cells before estimation, so the cell fixed effects enter as a pre-transformation rather than as absorbed dummies and R^2 is a within-cell R^2. Controls are a new-examiner indicator, continuation-type dummies, dummies for the first five years of examiner experience, and small- and micro-entity indicators. Stars: $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Columns correspond to Panels 1a, 1b, 1c and 1d of Figure 1.

Table L.2: *Conditional correlations: grant rate and scope, free-standing applications*

Regressor x	(1) Grant rate Quality	(2) Grant rate Filed scope	(3) Filed scope Quality	(4) Granted scope Quality	(5) Difference in scope Quality
x	0.0079***	−0.0434***	−0.0001	−0.0011	0.0075***
	(0.0007)	(0.0004)	(0.0017)	(0.0016)	(0.0011)
x^2	−0.0008*	−0.0242***	0.0058***	0.0029**	0.0026***
	(0.0005)	(0.0005)	(0.0011)	(0.0012)	(0.0009)
Net effect ($x: 0 \to +1$)	0.0071***	−0.0675***	0.0057***	0.0018	0.0101***
	(0.0009)	(0.0007)	(0.0021)	(0.0020)	(0.0014)
Net effect, NST minus free-standing	−0.0041***	0.0170***	−0.0168***	−0.0177***	−0.0026
	(0.0011)	(0.0011)	(0.0027)	(0.0028)	(0.0024)
Net effect, continuations minus free-standing	−0.0028***	0.0305***	−0.0123***	−0.0137***	−0.0014
	(0.0009)	(0.0009)	(0.0022)	(0.0023)	(0.0020)
Controls	✓	✓	✓	✓	✓
Art unit × year demeaning	✓	✓	✓	✓	✓
Observations	2,900,123	2,901,417	2,900,123	2,024,630	2,024,630
R^2	0.084	0.094	0.017	0.018	0.007

Notes: Quadratic specification $y = \beta_1 x + \beta_2 x^2 + \gamma' X + \varepsilon$ estimated by OLS on all applications with every term interacted with the filing group (free-standing applications, national-stage entries, and continuations, divisionals and continuations-in-part), with CRV1 (ind_year_fe) standard errors in parentheses. The rows above the differences report the free-standing coefficients. *Net effect* is $\hat{\beta}_1(x_{hi} - x_{lo}) + \hat{\beta}_2(x_{hi}^2 - x_{lo}^2)$ over $x_{lo} = 0$ to $x_{hi} = 1$, with a delta-method standard error. The two difference rows give the net effect of each other group minus the free-standing one, with standard errors from the joint covariance of the stacked model. The regressor is the standardized percentile rank of quality, except in column (2), where it is the standardized within-year percentile rank of filed scope. Filed scope in column (3) is that rank, granted scope in column (4) the percentile among the filed and granted claims of the same year, and column (5) their difference, all standardized. Every variable is demeaned within art unit × year × filing group cells before estimation, the grant rate is kept in levels, and R^2 is a within-cell R^2. Controls are a new-examiner indicator, continuation-type dummies, dummies for the first five years of examiner experience, and small- and micro-entity indicators. Stars: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Columns (1)–(4) correspond to Panels 2a, 2b, 2c and 2d of Figure 2. Column (5) corresponds to Appendix Figure D.1.

L.2 Complier shares

The complier-share figures plot a kernel-smoothed curve in the percentile of the conditioning covariate. Tables L.3 to L.6 report that curve at deciles, together with the objects that determine the plotted bands: the simultaneous-band multiplier, the bootstrap sup-test of flatness, and the first-stage share and dose that scale the estimand.

Table L.3: *Complier share among granted patents, by quality and filed scope*

Percentile	(1) Quality	(2) Filed scope
0.10	7.20 (0.11)	5.91 (0.09)
0.20	7.27 (0.11)	6.20 (0.09)
0.30	7.33 (0.11)	6.46 (0.10)
0.40	7.28 (0.11)	6.71 (0.10)
0.50	7.21 (0.11)	6.97 (0.10)
0.60	7.10 (0.11)	7.25 (0.11)
0.70	7.00 (0.10)	7.59 (0.11)
0.80	6.87 (0.10)	8.07 (0.12)
0.90	6.65 (0.10)	8.64 (0.13)
Joint flatness p	< 0.001	< 0.001
Sup-t critical value	2.39	2.46
Aggregate share (%)	7.06	7.06
Bandwidth	0.07	0.07
Bootstrap draws	2,000	2,000

Notes: Share of compliers among granted patents, $\phi_c^{D=1}$, in percent, from the estimator of Section 5.3, evaluated at deciles of the conditioning characteristic. The profile is scaled so that its grant-weighted average equals the aggregate share. Standard errors from the joint influence function, clustered on the art unit × year cell, in parentheses. *Joint flatness p* is the multiplier-bootstrap sup-test of the null that the curve is constant, and the sup-t critical value is the simultaneous-band multiplier plotted in the figure. Curves are kernel-smoothed at the reported bandwidth. Estimation sample: 5,360,465 applications. Columns correspond to Panels 4a and 4b of Figure 4.

Table L.4: *Decomposition of the complier share: representation ratio and grant rate*

	Quality				Filed scope			
Percentile	$\psi(x)$	$d(x)$	$P(D{=}1 \mid x)$	$\phi_c^{D=1}$	$\psi(x)$	$d(x)$	$P(D{=}1 \mid x)$	$\phi_c^{D=1}$
0.10	0.997	0.0510	0.706	7.20	0.924	0.0492	0.770	5.91
0.20	1.017	0.0506	0.707	7.27	0.955	0.0494	0.762	6.20
0.30	1.031	0.0503	0.708	7.33	0.978	0.0496	0.750	6.46
0.40	1.028	0.0503	0.710	7.28	0.997	0.0497	0.739	6.71
0.50	1.021	0.0502	0.711	7.21	1.011	0.0499	0.724	6.97
0.60	1.007	0.0502	0.712	7.10	1.022	0.0503	0.709	7.25
0.70	0.995	0.0502	0.713	7.00	1.035	0.0506	0.690	7.59
0.80	0.983	0.0500	0.715	6.87	1.050	0.0510	0.663	8.07
0.90	0.964	0.0494	0.716	6.65	1.051	0.0516	0.627	8.64
$p90/p10$	0.966	0.969	0.987	0.924	1.137	1.048	1.227	1.462
Share of log gradient (%)	44	39	17	100	34	12	54	100

Notes: Decomposition of the complier share of Table L.3 into its three moving parts. The identity $\phi_c^{D=1}(x) = \psi(x)\, d(x) \,/\, P(D{=}1 \mid x)$ holds exactly at every grid point, so the columns multiply to the reported share. $\psi(x)$ is the sensitivity of the grant to leniency, scaled so that the profile averages to the aggregate share, $d(x)$ the one-sided dose $\mathbb{E}[(\ell)_+ \mid x]$ in grant-probability units, and $P(D{=}1 \mid x)$ the conditional grant rate. The final rows report the $p90/p10$ ratio of each factor, with the grant rate entered as $P(D{=}1 \mid p10)/P(D{=}1 \mid p90)$ so that a ratio above one raises the share across the range, and the share of the log gradient in $\phi_c^{D=1}$ that each factor carries. Estimator, bandwidth and sample are those of Table L.3.

Table L.5: *Complier share: sample-restriction robustness*

Panel A: by quality

Percentile	(1) Baseline	(2) Excl. continuations	(3) Random-digit cells
0.10	7.20 (0.11)	7.39 (0.14)	7.66 (0.28)
0.20	7.27 (0.11)	7.42 (0.14)	7.72 (0.27)
0.30	7.33 (0.11)	7.42 (0.13)	7.73 (0.27)
0.40	7.28 (0.11)	7.39 (0.13)	7.60 (0.26)
0.50	7.21 (0.11)	7.33 (0.13)	7.31 (0.25)
0.60	7.10 (0.11)	7.16 (0.13)	7.11 (0.23)
0.70	7.00 (0.10)	6.95 (0.12)	6.93 (0.22)
0.80	6.87 (0.10)	6.82 (0.12)	6.68 (0.21)
0.90	6.65 (0.10)	6.55 (0.12)	6.31 (0.20)
Joint flatness p	< 0.001	< 0.001	< 0.001
Sup-t critical value	2.39	2.52	2.41
Aggregate share (%)	7.06	7.09	7.17
Bandwidth	0.07	0.07	0.07
Bootstrap draws	2,000	2,000	2,000

Panel B: by filed scope

Percentile	(1) Baseline	(2) Excl. continuations	(3) Random-digit cells
0.10	5.91 (0.09)	6.03 (0.12)	5.89 (0.22)
0.20	6.20 (0.09)	6.36 (0.12)	6.16 (0.22)
0.30	6.46 (0.10)	6.67 (0.12)	6.48 (0.22)
0.40	6.71 (0.10)	6.91 (0.12)	6.82 (0.23)
0.50	6.97 (0.10)	7.19 (0.12)	7.10 (0.23)
0.60	7.25 (0.11)	7.50 (0.13)	7.46 (0.24)
0.70	7.59 (0.11)	7.84 (0.13)	7.86 (0.24)
0.80	8.07 (0.12)	8.31 (0.14)	8.29 (0.25)
0.90	8.64 (0.13)	8.73 (0.15)	8.78 (0.26)
Joint flatness p	< 0.001	< 0.001	< 0.001
Sup-t critical value	2.46	2.56	2.37
Aggregate share (%)	7.06	7.09	7.17
Bandwidth	0.07	0.07	0.07
Bootstrap draws	2,000	2,000	2,000

Notes: Each column estimates the share of compliers among granted patents of Table L.3 on a different sample, with the estimator of Section 5.3 and the instrument built on the full sample. Only the rows the estimator uses change, and each column is scaled so that its grant-weighted average equals its own aggregate share. Standard errors clustered on the art unit × year cell in parentheses. Applications: Baseline 5,360,465; Excl. continuations 2,144,479; Random-digit cells 1,208,862. Reading conventions are as in Table L.3. Column two corresponds to Figure G.1. Column three reports the random-digit cells with their own instrument imbalance, and Figure G.2 draws them with the baseline's.

Table L.6: *Complier share in the replication subsamples*

Panel A: by quality

Percentile	(1) Sampat-Williams	(2) Farre-Mensa et al.
0.10	4.55 (0.84)	5.33 (0.88)
0.20	3.37 (0.70)	5.25 (0.75)
0.30	2.60 (0.99)	5.53 (0.70)
0.40	2.65 (1.45)	6.27 (0.70)
0.50	4.55 (1.08)	5.93 (0.71)
0.60	11.16 (2.77)	6.54 (0.75)
0.70	10.64 (4.47)	5.26 (0.73)
0.80	6.71 (2.92)	3.97 (0.74)
0.90	6.94 (3.62)	3.02 (0.89)
Joint flatness p	0.071	0.109
Sup-t critical value	2.78	2.85
First-stage $\bar{\psi}$	0.327	0.835
Mean favourable deviation $\bar{d}$	0.0362	0.0337
Bandwidth	0.07	0.07
Bootstrap draws	2,000	2,000

Panel B: by filed scope

Percentile	(1) Sampat-Williams	(2) Farre-Mensa et al.
0.10	4.56 (0.38)	4.44 (1.29)
0.20	5.29 (0.46)	4.00 (0.93)
0.30	–	5.29 (0.80)
0.40	23.98 (16.79)	5.72 (0.72)
0.50	5.53 (4.33)	6.03 (0.65)
0.60	0.85 (0.82)	5.80 (0.67)
0.70	3.61 (2.68)	5.05 (0.62)
0.80	17.26 (12.25)	4.26 (0.58)
0.90	10.92 (8.64)	4.73 (0.74)
Joint flatness p	0.000	0.704
Sup-t critical value	2.75	2.91
First-stage $\bar{\psi}$	0.327	0.835
Mean favourable deviation $\bar{d}$	0.0362	0.0337
Bandwidth	0.07	0.07
Bootstrap draws	2,000	2,000

Notes: The same examiner-dummy UJIVE modified-outcome estimator as Table L.3, applied within each replication subsample; for Sampat and Williams (2019) the examiner-leniency design is shown. Sample sizes: Sampat and Williams (2019) 233,442 gene-level observations across 81 examiners; Farre-Mensa et al. (2020) 10,286 applications across 2,196 examiners. The Sampat and Williams (2019) frame is gene-level, so its count is observations rather than applications; the underlying application count is smaller by roughly two orders of magnitude and is given with the full sample ladder in Appendix K. These subsamples are two to three orders of magnitude smaller than the full sample, so the standard errors are correspondingly wide and the point estimates should not be read at face value. Columns correspond to the two columns of Figure K.3.

L.3 Marginal treatment effects

The MTE figures plot three objects at once: the curve, the ATE obtained by integrating it, and the LATE from each paper's own published specification re-estimated on the same sample. Tables L.7 and L.8 report all three, so the ATE-versus-LATE comparison developed in Appendix K can be read numerically. The LATE row is the paper's own estimator and is a different estimand from the LATE implied by the MTE machinery. It is the former that the figures plot.

Tables L.9 and L.10 then test whether the two differ. The figures invite the wrong inference. Their ATE and LATE confidence bands overlap, and overlapping intervals are not a test of equality, not even for independent estimates, and these two are neither independent nor comparably constructed. The plotted ATE interval is a percentile bootstrap of the integrated curve. The plotted LATE interval is an analytic cluster-robust interval from a two-stage least squares fit. They come from different routines and share no draws, so nothing in them identifies the covariance between the two estimands, which is what the variance of their difference depends on. The tables therefore re-estimate both objects on a common set of bootstrap resamples and report the confidence interval of the difference, which handles that covariance by construction.

One convention in these tables differs from the figures. A test of two estimands is only interpretable if both are conditioned on the same covariates, so the LATE in the test is estimated under the control set that the marginal treatment effect's own outcome equation uses, on the identical estimation frame. For Sampat and Williams (2019) that set is the published one and the two coincide exactly. For Farre-Mensa et al. (2020) it adds the filed-scope control that the MTE outcome equation carries, so the LATE in Table L.10 is not numerically the LATE plotted in Figure K.2: it is 0.373 against 0.369 for follow-on applications, and at most 0.005 apart in any outcome, against a bootstrap standard error of roughly 0.07. The figures and Table L.8 report the published specification exactly, because there the LATE is a description of what the original design delivers rather than one side of a hypothesis test. Nothing in the verdicts below turns on the choice. Conditioning both estimands alike is what makes the difference a well-defined object.

Four refinements matter for how the resulting p-values should be read. First, the raw percentile interval is only first-order accurate, and the resample distribution here is neither symmetric nor median-unbiased, since the observed difference sits away from the median of its own resample distribution in every outcome. The headline interval is therefore bias-corrected and accelerated, which is second-order accurate. The other second-order-accurate option, a studentized interval, is not available: the integrated ATE has no closed-form variance estimator, so studentizing would require an inner bootstrap inside every outer replicate. Second, the outcomes within each setting are strongly dependent, being different follow-on measures constructed on the same applications, so the familywise adjustment is a Romano-Wolf stepdown (Romano and Wolf, 2005, 2016) computed

from the joint resample distribution rather than a Bonferroni or Holm correction, which would treat that dependence as absent and be badly conservative. Third, the number of replicates is itself an inferential choice rather than a computational detail: at 500 draws the Monte Carlo standard error of a p-value near 0.05 is roughly 0.010, wide enough that the borderline outcomes could not be assigned a side of the threshold, so the tables are computed at 5,000. Fourth, the resampling unit is itself a choice. The tables above resample applications independently, which matches the convention of the estimating cell but not the level the standard errors cluster on, examiner art unit $\times$ publication year. Table L.11 repeats the Farre-Mensa et al. (2020) test resampling those clusters instead, redrawing cluster labels on every replicate. The estimands are unchanged by construction, being full-sample quantities that do not depend on the bootstrap, and the clustered bootstrap reproduces the analytic cluster-robust standard error of the LATE marginally more closely than the iid one does, at a mean ratio of 0.99 against 0.98 across the outcomes, which is what one expects when the dependence the clustering absorbs is modest. The signs, the ordering of the outcomes, and which of them survives the familywise adjustment are all unchanged. That table is computed at $B = 1{,}000$, however, so by the Monte Carlo logic just stated its borderline entries cannot be assigned a side of the five percent threshold: it is evidence on the resampling scheme, not a replacement for the tables above.

The verdicts match Appendix K: the ATE lies below the LATE in every Farre-Mensa et al. (2020) outcome under both integration windows, the difference rejects individually only in its two citation outcomes, one of which survives the Romano-Wolf adjustment, and neither primary outcome rejects equality of the two estimands. One detail is specific to these tables: both citation rejections are BCa p-values, and in both the correction moves the verdict toward rejection relative to the percentile interval, which is why the familywise column is reported alongside.

One diagnostic in these tables is of independent interest. The correlation between the ATE and the LATE across resamples never exceeds 0.13, even though the two are computed on identical rows. The integrated ATE loads on the fitted curve across the whole support of U, whereas the two-stage least squares LATE loads on variation in the complier region. The two draw on nearly disjoint information. A consequence is that the naive procedure of adding the two variances, which the joint bootstrap was designed to avoid, would have been approximately right here.

Table L.7: *Marginal treatment effects: Sampat and Williams (2019)*

	(1) log(Publications)	(2) Any publication
ATE (full support)	−0.017 [−0.146, 0.087]	−0.005 [−0.154, 0.115]
ATE (common support)	0.015 [−0.081, 0.099]	0.028 [−0.060, 0.126]
LATE (paper 2SLS)	−0.055 (0.089)	−0.051 (0.087)
ATE − LATE (full support)	0.039	0.046
ATE − LATE (common support)	0.070	0.079
MTE at $U = 0.10$	−0.090	−0.079
MTE at $U = 0.50$	0.097	0.109
MTE at $U = 0.90$	−0.169	−0.157
Observations	1,196	1,196

Notes: Marginal treatment effect estimates for Sampat and Williams (2019), on the variation used by the examiner-dummy UJIVE estimator (arm A: the leave-out deviation enters the selection probit as the instrument). ATE integrates the MTE curve over the full support of U; the common-support ATE integrates over the region where the propensity score is identified. Bootstrap 95% confidence intervals in brackets. The LATE row is each paper's *own* published 2SLS specification re-estimated on exactly this sample, with its cluster-robust standard error in parentheses – it is the blue line in the corresponding figure and is a different object from the MTE-machinery LATE. *Caveat*: the LATE row is unweighted across applications, whereas the original estimand of Sampat and Williams (2019) is gene-weighted. MTE rows report the curve evaluated at the stated latent resistance, linearly interpolated on the estimation grid. Column (1) corresponds to Panel K.2a of Figure K.2 and column (2) to Panel K.4a of Figure K.4.

Table L.8: *Marginal treatment effects: Farre-Mensa et al. (2020)*

	(1) log(Follow-on appl.)	(2) log(All cites, 5y)	(3) log(Appl. cites, 5y)	(4) log(Granted patents)
ATE (full support)	0.273	0.015	−0.042	0.216
	[−0.027, 0.529]	[−0.262, 0.314]	[−0.189, 0.120]	[0.028, 0.405]
ATE (common support)	0.340	0.102	0.005	0.264
	[0.055, 0.559]	[−0.168, 0.365]	[−0.138, 0.147]	[0.070, 0.430]
LATE (paper 2SLS)	0.373	0.366	0.212	0.291
	(0.068)	(0.073)	(0.038)	(0.052)
ATE − LATE (full support)	−0.100	−0.351	−0.254	−0.075
ATE − LATE (common support)	−0.033	−0.264	−0.207	−0.027
MTE at $U = 0.10$	−0.701	−1.223	−0.681	−0.426
MTE at $U = 0.50$	0.331	0.099	0.027	0.297
MTE at $U = 0.90$	1.120	1.063	0.450	0.697
Observations	10,286	10,286	10,286	10,286

Notes: Marginal treatment effect estimates for Farre-Mensa et al. (2020), on the variation used by the examiner-dummy UJIVE estimator (arm A: the leave-out deviation enters the selection probit as the instrument). ATE integrates the MTE curve over the full support of U; the common-support ATE integrates over the region where the propensity score is identified. Bootstrap 95% confidence intervals in brackets. The LATE row is each paper's *own* published 2SLS specification re-estimated on exactly this sample, with its cluster-robust standard error in parentheses – it is the blue line in the corresponding figure and is a different object from the MTE-machinery LATE. MTE rows report the curve evaluated at the stated latent resistance, linearly interpolated on the estimation grid. Column (1) corresponds to Panel K.2b of Figure K.2, and columns (2)–(4) to Panels K.4b, K.4c and K.4d of Figure K.4.

Table L.9: *Testing the average against the local average treatment effect: Sampat and Williams (2019)*

	(1) log(Publications)	(2) Any publication	(3) Genetic test
Estimands			
ATE (full support)	−0.017	−0.005	0.013
ATE (common support)	0.015	0.028	0.008
LATE (paper 2SLS)	−0.055	−0.051	0.025
Test of ATE = LATE, full-support integral			
Difference	0.039	0.046	−0.012
95% BCa CI	[−0.109, 0.320]	[−0.119, 0.321]	[−0.143, 0.183]
p (BCa)	0.568	0.524	0.910
p (percentile)	0.667	0.606	0.788
p (Romano-Wolf, FWER)	0.830	0.830	0.855
Test of ATE = LATE, common-support integral			
Difference	0.070	0.079	−0.018
95% BCa CI	[−0.051, 0.380]	[−0.039, 0.439]	[−0.133, 0.143]
p (BCa)	0.253	0.197	0.792
p (percentile)	0.342	0.318	0.740
p (Romano-Wolf, FWER)	0.492	0.492	0.749
Diagnostics			
Corr(ATE, LATE) across resamples (full support)	0.063	0.018	0.015
Bootstrap s.e. of LATE	0.225	0.233	0.214
Cluster-robust s.e. of LATE	0.089	0.087	0.041
BCa bias correction $\hat{z}_0$ (full support)	0.069	0.060	0.077
BCa acceleration $\hat{a}$ (full support)	0.0147	0.0032	0.0365
Usable bootstrap replicates	5,000	5,000	5,000
Jackknife groups	50	50	50
Observations	1,196	1,196	1,196

Notes: Test of the null that the integrated ATE equals the LATE, for Sampat and Williams (2019). Both estimands are re-estimated on *the same* 5,000 iid row resamples, so the interval for their difference accounts for the covariance between them – which neither band in the corresponding figure contains, so that overlap is not a test of equality. ATE (full support) integrates the MTE curve over all of U, ATE (common support) only where the propensity score has support, and LATE is each paper's *own* published 2SLS estimator, re-estimated on this sample under the control set of the marginal-treatment-effect outcome equation so that both estimands are conditioned alike. The BCa p inverts a bias-corrected and accelerated interval and is the single-hypothesis test, so the stars follow it; the percentile p is its first-order-accurate counterpart; the Romano-Wolf p is a stepdown adjustment controlling the familywise error rate across the outcomes of this table. The examiner-leniency design is shown, matching the figures; the alternative design is estimated on the same sample. Diagnostics: the resample correlation of the two estimands is the quantity the naive add-the-variances test assumes away; the bootstrap and cluster-robust standard errors of the LATE are shown side by side as a check on whether iid resampling understates within-cell dependence; $\hat{z}_0$ and $\hat{a}$ are the BCa median-bias and skewness constants, both zero when BCa and the percentile method coincide. The LATE row is Sampat and Williams's specification re-estimated *unweighted* on this sample. Their published estimand weights by the number of genes per application; re-estimating with those weights reverses the sign of every LATE reported here while leaving each one statistically indistinguishable from zero, so the null result is robust to the choice of estimand but the signs are not. The weighted estimates are reported in Appendix K. Sample: 1,196 applications. Column (1) corresponds to Panel K.2a of Figure K.2 and column (2) to Panel K.4a of Figure K.4. Column (3) is not plotted. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table L.10: *Testing the average against the local average treatment effect: Farre-Mensa et al. (2020)*

	(1) log(Follow-on appl.)	(2) log(All cites, 5y)	(3) log(Appl. cites, 5y)	(4) log(Granted patents)	(5) Approval rate
Estimands					
ATE (full support)	0.269	0.018	−0.024	0.221	0.138
ATE (common support)	0.318	0.081	0.008	0.255	0.161
LATE (paper 2SLS)	0.373	0.362	0.211	0.291	0.200
Test of ATE = LATE, full-support integral					
Difference	−0.104	−0.344**	−0.235***	−0.070	−0.062
95% BCa CI	[−0.379, 0.209]	[−0.672, −0.041]	[−0.408, −0.083]	[−0.275, 0.163]	[−0.214, 0.097]
p (BCa)	0.521	0.029	0.003	0.607	0.454
p (percentile)	0.443	0.038	0.006	0.457	0.378
p (Romano-Wolf, FWER)	0.660	0.088	0.017	0.660	0.660
Test of ATE = LATE, common-support integral					
Difference	−0.056	−0.281*	−0.203***	−0.036	−0.039
95% BCa CI	[−0.302, 0.241]	[−0.567, 0.014]	[−0.356, −0.057]	[−0.219, 0.193]	[−0.175, 0.109]
p (BCa)	0.775	0.062	0.008	0.862	0.663
p (percentile)	0.584	0.056	0.010	0.590	0.512
p (Romano-Wolf, FWER)	0.834	0.152	0.028	0.834	0.834
Diagnostics					
Corr(ATE, LATE) across resamples (full support)	0.073	0.044	0.079	0.067	0.041
Bootstrap s.e. of LATE	0.068	0.072	0.039	0.050	0.036
Cluster-robust s.e. of LATE	0.069	0.074	0.038	0.052	0.038
BCa bias correction $\hat{z}_0$ (full support)	0.057	−0.095	−0.172	0.112	0.068
BCa acceleration $\hat{a}$ (full support)	0.0225	0.0190	0.0152	0.0134	−0.0029
Usable bootstrap replicates	5,000	5,000	5,000	5,000	5,000
Jackknife groups	50	50	50	50	50
Observations	10,289	10,289	10,289	10,289	10,289

Notes: Test of the null that the integrated ATE equals the LATE, for Farre-Mensa et al. (2020). Both estimands are re-estimated on *the same* 5,000 iid row resamples, so the interval for their difference accounts for the covariance between them – which neither band in the corresponding figure contains, so that overlap is not a test of equality. ATE (full support) integrates the MTE curve over all of U, ATE (common support) only where the propensity score has support, and LATE is each paper's *own* published 2SLS estimator, re-estimated on this sample under the control set of the marginal-treatment-effect outcome equation so that both estimands are conditioned alike. The BCa p inverts a bias-corrected and accelerated interval and is the single-hypothesis test, so the stars follow it; the percentile p is its first-order-accurate counterpart; the Romano-Wolf p is a stepdown adjustment controlling the familywise error rate across the outcomes of this table. Diagnostics: the resample correlation of the two estimands is the quantity the naive add-the-variances test assumes away; the bootstrap and cluster-robust standard errors of the LATE are shown side by side as a check on whether iid resampling understates within-cell dependence; $\hat{z}_0$ and $\hat{a}$ are the BCa median-bias and skewness constants, both zero when BCa and the percentile method coincide. That conditioning is why the LATE here differs from the one in Table L.8, by up to 0.004: that table and the figures report the published specification exactly, since there the LATE describes what the original design delivers rather than forming one side of a hypothesis test. Sample: 10,289 applications. Column (1) corresponds to Panel K.2b of Figure K.2, and columns (2)–(4) to Panels K.4b, K.4c and K.4d of Figure K.4. Column (5) is not plotted. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table L.11: *Testing the average against the local average treatment effect under a clustered bootstrap: Farre-Mensa et al. (2020)*

	(1) log(Follow-on appl.)	(2) log(All cites, 5y)	(3) log(Appl. cites, 5y)	(4) log(Granted patents)
Estimands				
ATE (full support)	0.269	0.018	−0.024	0.221
ATE (common support)	0.318	0.081	0.008	0.255
LATE (paper 2SLS)	0.373	0.362	0.211	0.291
Test of ATE = LATE, full-support integral				
Difference	−0.104	−0.344**	−0.235***	−0.070
95% BCa CI	[−0.380, 0.201]	[−0.680, −0.057]	[−0.418, −0.094]	[−0.282, 0.159]
p (BCa)	0.498	0.026	0.002	0.531
p (percentile)	0.438	0.062	0.010	0.508
p (Romano-Wolf, FWER)	0.687	0.084	0.021	0.687
Test of ATE = LATE, common-support integral				
Difference	−0.056	−0.281**	−0.203***	−0.036
95% BCa CI	[−0.299, 0.233]	[−0.592, −0.001]	[−0.366, −0.065]	[−0.229, 0.180]
p (BCa)	0.765	0.050	0.004	0.759
p (percentile)	0.586	0.080	0.012	0.644
p (Romano-Wolf, FWER)	0.853	0.145	0.031	0.853
Diagnostics				
Corr(ATE, LATE) across resamples	0.176	0.159	0.192	0.117
Bootstrap s.e. of LATE	0.069	0.072	0.037	0.053
Cluster-robust s.e. of LATE	0.069	0.074	0.038	0.052
BCa bias correction $\hat{z}_0$	0.050	−0.164	−0.264	0.018
BCa acceleration $\hat{a}$	−0.0044	−0.0063	0.0003	−0.0003
Usable bootstrap replicates	1,000	1,000	1,000	1,000
Jackknife groups	50	50	50	50
Observations	10,289	10,289	10,289	10,289

Notes: Companion to Table L.10, which reports the same test under an iid row resample. Here the bootstrap resamples *clusters* of examiner art unit × publication year, matching the unit the standard errors cluster on: 873 clusters, 1,000 usable replicates, with cluster labels redrawn on each replicate. The ATE and LATE rows are identical to Table L.10 by construction: they are full-sample quantities and do not depend on the resampling scheme. Only the intervals and p-values differ. *This table is not a corrected version of Table L.10*: it uses $B = 1{,}000$ replicates against 5,000 there, so the two differ in resampling scheme and in number of replicates simultaneously, and any difference between them cannot be attributed to the scheme alone. The Romano-Wolf stepdown controls the familywise error rate across the full 5-outcome bootstrap family, which includes 1 outcome not displayed here (the approval rate). Stars follow the BCa p-value. Columns correspond to those of Table L.10. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

M The model: specification, identification, and fit

This appendix collects the full specification of the dynamic prosecution model of Section 7, states which empirical moments discipline each parameter, and reports the fit of the estimated model. The model is estimated by indirect inference (Gouriéroux et al., 1993) (Section 7.5): the parameter vector θ is chosen to minimize the weighted distance between a set of binned empirical moments and their simulated counterparts, the latter evaluated on a panel of seven million applications drawn from the estimated primitives and solved by value-function iteration at each candidate θ. Common random numbers are held fixed across parameter evaluations, so differences in the criterion reflect θ rather than the draw. The estimate comes from a sequence of searches, each started from the previous estimate. The last ran a covariance-matrix-adaptation evolution strategy (Hansen, 2006) and stopped on its step-size criterion after 36,596 evaluations. By the convergence rule set before it ran, it had not converged: over its last ten hours the criterion fell by 38.5, against a tolerance of 13.6. A further search started from the estimate, under the same objective, found no improvement in 13,279 evaluations. The estimate is therefore the best point these searches found, not a certified minimum. The three subsections that follow list the parameters and their estimates (Appendix M.1), the identification argument and the inference procedure (Appendix M.2), and the fit (Appendix M.3). The last includes an untargeted comparison of the model's complier shares with the examiner-assignment estimates of Section 6.

M.1 Parameters

Table M.1 lists the parameters, their economic role, the admissible bounds imposed during estimation, and the point estimate. Curvature χ, the examiner's signal noise σ_η, the offer's leniency gradient γ_l and its floor $\underline{l}$ are calibrated and held fixed, together with the examiner's prior mean $\bar{q}$ and standard deviation σ_q and the applicant's belief bias q_{err}. The remaining 27 are estimated by indirect inference, and in the standard errors ν is held fixed as the numeraire of the payoff scale (Appendix M.2).

The signal noise. σ_η is set by a profile. Each of five runs holds it fixed, at 3, 1.5, 1, 0.7 or 0.5, and re-estimates the other parameters from the same starting point, adjusted only to hold the first-round quality coefficient fixed, under common random numbers. The criterion is about 904, 873, 895, 862 and 837. It is lowest at 0.5, the smallest value profiled, so the profile does not bracket σ_η. By the convergence rule set before the runs, three of the five searches had not converged (at 1.5, 0.7 and 0.5), and the run at 3 found no improvement on its starting point, so each value is an upper bound on the profiled criterion at that σ_η.

The scope penalty $\Pi(s) = \alpha_1 s + \alpha_2 s^2 + \alpha_3 s^3$ is a fitted polynomial rather than a shape imposed

a priori. At the estimates all three coefficients are positive, so Π is increasing on $[0,1]$ and the acceptance index falls with filed scope everywhere. The design-based results of Sections 5–6 use only the sign of $\partial a/\partial s$ where the filings actually are, which is negative and is measured directly from the grant rate by filed scope.

Two naming points, since both have caused confusion. The applicant's belief dispersion q_{unc} and the examiner's prior standard deviation $\sigma_q = 1/\sqrt{12}$ are *different objects* and are not alternative values of one quantity: q_{unc} governs the subjective density an applicant holds over own quality (Section 7.1) and is estimated, while σ_q enters only the examiner's precision weight f_r in (15) and is calibrated to the spread of the quality rank. They are grouped separately below for that reason. The round drift of the acceptance index is b_r throughout.

The shock variance and the quality pivot. The grant rule (16) depends on the examiner's posterior mean $\mathbb{E}_r[q]$, which moves with signals the applicant never sees. Given true quality, this posterior mean equals its average $f_r\, q + (1 - f_r)\, \bar{q}$ from Equation (15) plus an error that comes from the signal noise. The model is solved and simulated with the average in the acceptance index and the error added to the shock ε_r, so the shock of this averaged grant rule has two sources,

$$\sigma^2_{\text{tot}}(r) \;=\; \sigma_e(r)^2 + \sigma^2_{\text{grant}}, \qquad \sigma_e(r) \;=\; \alpha_q\, f_r\, \frac{\sigma_\eta}{\sqrt{r+1}}, \tag{34}$$

the first term being the dispersion of the examiner's posterior around this average, equal to $\alpha_q \sigma_q \sqrt{f_r(1-f_r)}$. At $\sigma_\eta = 0.5$ it rises from $0.125\alpha_q$ at $r = 0$ to its maximum of $0.144\alpha_q$ at $r = 2$, where $f_r = \frac{1}{2}$, and falls slightly thereafter, to $0.136\alpha_q$ at $r = 5$, while the posterior standard deviation itself falls from 0.250 to 0.167. The second, σ_{grant}, is an idiosyncratic case-specific component that does not vanish as $\alpha_q \to 0$ and so keeps the grant probability interior when quality carries little weight. Using the realized posterior mean together with this wider shock would count the signal noise twice. In the first round, $r = 0$, the averaged rule is the static grant rule of Equation (5), with $f = f_0$, $\sigma = \sigma_{\text{tot}}(0)$ and the direct quality term $-(b_q + b_{0q})(q - \frac{1}{2})$ added. Because the round drift $b_r r$ is zero in the first round, the threshold μ of the static rule corresponds to $\mu + b_0$ here. The simulation draws the round-r shock independently across rounds, although the σ_e component is not in fact independent: successive posteriors share signals, so $\text{Corr}(r, r') = \sqrt{(r+1)/(r'+1)}$, rising from 0.71 to 0.91 over the observed range. The approximation is immaterial by construction rather than by calibration: $\sqrt{f_r(1-f_r)} \le \frac{1}{2}$, so $\sigma_e \le \alpha_q \sigma_q / 2$, which at the estimated α_q is 2.0% of σ_{grant} for *any* value of σ_η. The bound binds through α_q, not through the signal noise.

Quality enters the acceptance index in two ways. The posterior mean is affine in true quality: by (15), its average given true quality enters the index as $\alpha_q\, \mathbb{E}[\mathbb{E}_r[q] \mid q] = \alpha_q \bar{q} + \alpha_q f_r\, (q - \bar{q})$. That part pivots about $q = \bar{q}$ as rounds accumulate, steepening in f_r while its value at the prior mean is

unchanged. The quality barriers subtract $b_q(q - \frac{1}{2})$ in every round and $b_{0q}(q - \frac{1}{2})$ in the first, and with $\bar{q} = \frac{1}{2}$ they pivot about the same point. The net slope, $\alpha_q f_r - b_q - b_{0q}\mathbf{1}[r = 0]$, is negative and rises toward zero, from -0.031 at $r = 0$ to -0.002 at $r = 5$. No amount of learning moves the profile uniformly, so a trend in generosity common to all applications has to come from b_r and b_0.

The offer is proportional to the claim. At the estimates, the offer ratio $\pi(l, s, q)\,\gamma_{\mathrm{E}}^{r}$ of Equation (17) lies between 0.75 and 0.84 in every state of the grid. The cap $\hat{s}_r \leq s_r$ therefore never binds, and granted scope is the round-r claim times the offer ratio. The one exception is the bottom of the scope grid. The simulation bounds granted scope below at 0.01, the smallest grid value, so an application filed at $s = 0.01$ is granted at its filed scope. In the simulated panel at the estimates, these applications make up 2.5% of grants, and each would otherwise concede about 0.002. Across all grants, mean granted scope is 0.78 times mean filed scope, and a regression of granted on filed scope has slope 0.75 and intercept 0.015. Proportionality serves two purposes. First, the base offer ratio γ_s sets the level of granted scope and the gap between filed and granted scope together, and the observed ratio of granted to filed scope identifies it without a correction for capped applications. Second, dividing granted scope by the offer ratio recovers the claim, so the offer can be re-priced without re-simulating the panel, as in Appendices M.3 and M.4.

What the applicant does not internalize in the offer. The applicant's value function (19) values a grant at νs, the claim, rather than at $\nu \hat{s}$, the granted scope. Because the offer is proportional to the claim, the average offer ratio enters the grant payoff only as a factor on ν and is absorbed by it. The value $\nu = 3.76$ is therefore a value per unit of filed scope, which at the mean offer ratio of 0.78 corresponds to about 4.8 per unit of granted scope, in the units the numeraire sets. What the restriction removes is any variation in the offer ratio that the applicant could anticipate. Three of the four sources of such variation carry almost none at the estimates. The round decay γ_{E} and the scope gradient $\gamma_{s\text{-sl}}$ sit at their bounds (0.999 and 0.000), and the quality gradient γ_q moves the ratio by about 0.008 across the quality range. The leniency gradient does carry variation, but only above the median examiner. With $\gamma_l = 0.141$ and the floor at $\underline{l} = \frac{1}{2}$, the offer ratio is about 0.77 for every examiner at or below the median and rises to 0.84 at the most lenient. The applicant's belief about l enters the grant probability $\bar{P}_r(s)$ but not the payoff, so an applicant who infers a strict examiner from repeated rejections expects a lower chance of a grant but not a narrower grant. The restriction therefore affects the applicant's choices of scope and of when to stop.

Table M.1: *Model parameters, roles, and estimates*

Parameter	Role	Bounds	Estimate
Examiner grant rule			
α_q	Quality term of the acceptance index	[0, 0.06]	0.056
b_q	Quality barrier	[−0.06, 0.06]	0.039
b_{0q}	First-round quality barrier	[−0.06, 0.06]	0.006
α_1	Linear scope penalty	[−0.5, 2]	0.291
α_2	Quadratic scope penalty	[0, 2]	0.062
α_3	Cubic scope penalty	[0, 0.02]	0.004
μ	Acceptance threshold	[0.1, 1.5]	0.468
σ_{grant}	Idiosyncratic grant noise	[0.25, 0.45]	0.409
b_r	Round drift of acceptance index	[0.01, 0.25]	0.033
b_0	First-round grant barrier	[0.5, 0.6]	0.587
Applicant beliefs			
q_{unc}	Dispersion of belief over own quality	[0, 1]	0.391
q_{err}	Bias of belief over own quality	calibrated	0.0
Applicant scope value and cost			
κ	Scope-cost scale	[0.05, 1.5]	0.458
χ	Scope-cost curvature	calibrated	2.0
ν	Filing-value multiplier (scope_mult)	[1, 4]	3.759
σ_s	Scope-policy dispersion (scope_e)	[0.2, 0.5]	0.288
c_0	Fixed revision cost	[0.009, 0.35]	0.009†
c_s	Scope revision cost	[0, 2]	0.295
c_r	Round revision cost	[0, 0.25]	0.249†
c_q	Quality factor of the stakes	[0, 0.3]	0.134
c_{q^2}	Quality factor of the stakes, top half (one-sided)	[0, 0.8]	0.212
σ_c	Revision-cost dispersion	[0.05, 0.65]	0.649†
Scope policy and examiner offer			
β_s	Leniency term of the filing displacement	[−0.05, 0.5]	−0.023
β_q	Quality term of the filing displacement	[−0.5, 0.5]	−0.019
γ_s	Base offer ratio (granted/filed scope)	[0.25, 0.95]	0.767
γ_l	Leniency gradient of the offer ratio	calibrated	0.141
$\underline{l}$	Leniency floor of the offer ratio	calibrated	0.5
$\gamma_{s\text{-sl}}$	Filed-scope gradient of the offer ratio	[0, 1]	0.000†
γ_q	Quality gradient of the offer ratio	[−0.05, 0.02]	−0.008
γ_{A}	Scope-concession decay	[0.5, 1]	0.715
γ_{E}	Per-round decay of the offer (scope_decay)	[0.7, 1]	0.999†
Calibrated (not estimated)			
σ_η	Examiner signal noise	fixed	0.5
$\bar{q}$	Examiner prior mean of quality	fixed	0.5
σ_q	Examiner prior SD of quality	fixed	$1/\sqrt{12}$

Notes: Point estimates from the warm-started CMA-ES run at $\sigma_\eta = 0.5$, the vector used throughout Sections 7.5 and 7.7. Full-sample weighted moment distance 836.7 over the sixteen targeted moment sets. Calibrated constants: curvature $\chi = 2.0$, $\sigma_\eta = 0.5$, the offer's leniency gradient $\gamma_l = 0.141$ (data-anchored to the measured round-zero conditional leniency slope) with the offer's leniency term floored at $\underline{l} = 0.5$, and $q_{\text{err}} = 0$, collinear with q_{unc}. † Estimate within 1% of the searched interval (5 parameters). Further estimates within 10% of a bound: α_2 at 3.1%, β_s at 4.9%, α_q at 94.1%, ν at 92.0%, b_r at 9.7%.

Horizon. The finite horizon $R = 12$ is a computational approximation rather than a statutory limit, so it has to be slack at the estimates for the round-cap counterfactuals of Section 7.7 to read as reductions in prosecution length rather than as artifacts of where the model was truncated. In the seven-million-application panel at $\hat{\theta}$, no application reaches the last round, and one reaches the round before it. Re-solving the model at $\hat{\theta}$ over horizons $R \in \{4, 6, \ldots, 16\}$ under common random numbers, the aggregate grant rate has converged by $R = 8$ and is the same at every horizon at or above the baseline, and lengthening the horizon to sixteen rounds moves the grant losses under the round caps by under 1.5%. The grant losses under the round caps are therefore caused by the caps being short, not by the baseline being close to an edge.

M.2 Identification and inference

Although all parameters are estimated jointly, each is disciplined primarily by a distinct feature of the data. Table M.7 states the leading source of identifying variation for each parameter, together with the weight the moment carries in the objective, and Table M.8 groups the same mapping into parameter blocks and gives the economic reasoning behind each assignment. The argument is heuristic, in the spirit of the sensitivity measure of Andrews, Gentzkow, and Shapiro (2017): each parameter is assigned to the moment set to which the simulated moments $m^{\text{sim}}(\theta)$ are most responsive at $\hat{\theta}$. The grant rate by leniency carries the dominant weight (200) because it pins the central mechanism, the slope of the grant probability in leniency.

The weighting matrix. Two choices in W depart from textbook practice and are made deliberately. First, each binned profile is standardized by the cross-bin standard deviation of its data reference curve rather than by per-bin sampling standard errors. With hundreds of thousands of applications per bin, per-bin standard errors shrink toward zero and a conventional efficient (inverse-sampling-variance) weighting would let residual bin-level noise dominate the criterion. Standardizing by the amplitude of each profile across bins instead targets the *shape* of the conditional-mean curve, the slope and curvature that identify the scope penalties, the filing gradients, and the concession and offer parameters. Second, W is diagonal, combining that within-profile standardization with a small set of across-profile weights that equalize the influence of shape-carrying moments that would otherwise be swamped. The grant rate by leniency is the clearest case, since its amplitude is small but its slope identifies σ_{grant}. Both choices trade asymptotic efficiency for identification stability, which is why the across-profile weights are part of the estimand and not only of its efficiency (Section 7.5).

The calibrated offer gradient. The offer's leniency gradient γ_l and its floor $\underline{l}$ are held at data-anchored values rather than estimated. At round zero (17) implies a granted-to-filed ratio of

exactly $\pi(l)$, flat in leniency below $\underline{l}$ and rising at rate γ_l above it. In the data the round-zero ratio is flat below the median examiner and rises above it, which sets $\underline{l} = \frac{1}{2}$, and γ_l is anchored on the ratio's leniency slope, free of the grant selection that flattens the unconditional granted-scope profile. The round-zero ratio is not itself an auxiliary statistic in (21), so the anchoring adds no moment to the objective: granted scope by leniency remains a weighted target that the estimated parameters have to fit. Appendix M.4 varies both constants.

Two assignments are easily confused. The concession rate γ_{A} and the offer decay γ_{E} both shape how scope narrows across rounds, the first through the claims applicants file after a rejection and the second through the scope the examiner offers. In the computed sensitivity of Table M.4, the estimate of γ_{A} is most sensitive to filed scope by leniency and that of γ_{E} to granted scope by leniency, the split between filed and granted scope. Separately, the leniency term β_s of the filing displacement noted in Section 7.3 is disciplined by the scope-by-leniency profile as a co-movement, and separately from the offer's leniency gradient γ_l, which moves *granted* rather than filed scope.

Local identification. The assignment above is heuristic, so I check it against the object that actually governs local identification: the curvature of the objective, $\Lambda' W \Lambda$, with $\Lambda = \partial m^{\mathrm{sim}}/\partial \theta'$ evaluated at $\hat{\theta}$ by finite differences under common random numbers, central where the step stays inside the searched interval and one-sided inward for the five estimates at a bound. Its columns are expressed in bound-range units, so a unit move is the same fraction of the searched interval for every parameter and the diagnostic is not an artifact of scaling. The stacked Jacobian is $1{,}320 \times 27$ over the 16 moment sets.

The spectrum over the 26 columns that remain with ν held fixed is sharply graded, with a condition number of 3.1×10^6 (Table M.6). Its three smallest eigenvalues, in bound-range units, load on the offer block. The smallest combines the offer's slope in filed scope $\gamma_{s\text{-sl}}$ with the offer decay γ_{E}, the base offer γ_s and the concession rate γ_{A}, the second is γ_{A} nearly alone, and the third sets γ_{E} against $\gamma_{s\text{-sl}}$ and the offer's quality gradient γ_q. These are the directions the objective measures worst: what the moment set measures least precisely is how the narrowing of scope splits between the applicant's concessions and the examiner's offer.

A near-flat pair sits in the offer block. The Jacobian columns of the base offer ratio γ_s and the offer decay γ_{E} are correlated at 0.999 at $\hat{\theta}$: because most grants occur early and the granted claim is a near-common multiple of the filed claim there, the level and the per-round decay of the offer move the moments along nearly the same direction, and the round-resolved endpoint variation separates them only weakly. This is consistent with γ_{E}'s estimate sitting against its bound, and the identification table's assignment of the decay to the narrowing of scope across rounds should be read with that near-collinearity in mind. The concession rate γ_{A} is close to the pair as well: its column correlates with either at about 0.77.

A further flat direction is economic rather than numerical. The applicant's problem is homogeneous of degree one in the payoff block, the value multiplier ν, the scope cost κ and the revision-cost coefficients c_0, c_s and c_r: scaling all five jointly leaves every policy and every simulated moment unchanged, and no auxiliary statistic observes a payoff level. The payoff scale is therefore a normalization, not an estimate: I read ν as the numeraire and the four cost coefficients as relative to it. Their standard errors are computed with ν held fixed, so they describe the cost coefficients relative to the numeraire.

The estimator also smooths the applicant's posterior after each rejection with a Gaussian kernel across neighbouring cells of the belief grid. Its bandwidth is estimated at 0.078 grid cells, where each neighbouring cell receives a weight of about 4×10^{-36}, so the smoothing is the identity to machine precision and the applicant updates by Bayes' rule on the grid. The model is therefore presented without it.

The per-round scope decay γ_{E} sits at its upper bound, and a weakly measured direction and a corner are the same finding seen twice: the objective gives the estimator little reason to hold γ_{E} away from unity, so it does not. The estimate $\hat{\gamma}_{\mathrm{E}} \approx 1$ means the offer's scope concession does not decay across rounds, and Section 7.7's round-cap counterfactual should be read with that in mind.

Inference. Under standard indirect-inference regularity conditions (Gouriéroux et al., 1993), $\hat{\theta}$ is consistent and asymptotically normal with sandwich variance

$$\widehat{\mathrm{Var}}(\hat{\theta}) \;=\; \left(1 + \tfrac{1}{S}\right) \left(\Lambda' W \Lambda\right)^{-1} \Lambda' W \,\widehat{\Omega}\, W \Lambda \left(\Lambda' W \Lambda\right)^{-1}, \tag{35}$$

where Λ is the binding-function Jacobian, $\widehat{\Omega}$ the asymptotic variance of the empirical auxiliary statistics, W the weighting matrix of Section 7.5, and $S = N_{\mathrm{sim}}/N$ the simulation ratio, whose finite value inflates the variance by $1+1/S$. The bread is the Gauss–Newton form: under misspecification the population criterion's Hessian carries an additional residual-weighted curvature term of the binding function, which (35) omits, so the variance is the standard local approximation at the pseudo-true parameter rather than a misspecification-robust one.

Two features of $\hat{\theta}$ govern what (35) can deliver. The first is that the flat directions above are joint combinations rather than columns, so with ν held fixed as the numeraire the inversion runs over the other 26 parameters exactly rather than through a pseudo-inverse. The second is the position of five estimates on a bound, taken up below.

Standard errors for the 26 follow from (35) with $\widehat{\Omega}$ estimated by a block bootstrap over examiners, the level at which the instrument varies, on 499 resamples, with $S = 1.305$. Twenty-five of the twenty-six are more than two standard errors from zero. The one that is not is the offer's slope in filed scope $\gamma_{s\text{-sl}}$, at its lower bound of zero. The Jacobian uses a step of 2% of each

parameter's searched interval, and for nine of the 26 the column is rough across adjacent bins, with a correlation below 0.9: q_{unc}, b_{0q}, c_q, c_{q^2}, b_0, α_3, c_r, b_q and σ_c. Noise in a finite-difference column inflates it, so the standard errors of these nine may be understated. For the coordinates on bounds the sandwich is at best descriptive: a constrained estimator's limiting distribution is a projection onto the feasible cone rather than the interior normal. Table M.2 reports their standard errors from one-sided differences and flags them. Where this section states such a coordinate in standard-error units, the unit is the unrestricted sandwich scale, and no comparison with zero for those coordinates is a test.

The second requirement is that $\hat{\theta}$ lie in the interior, and five of the twenty-seven do not: c_0 and $\gamma_{s\text{-sl}}$ at their lower bounds, and c_r, σ_c and γ_{E} at their upper bounds. Two more, α_2 and β_s, lie within five percent of the searched interval from their lower bounds, about six standard errors away. At a bound the asymptotic distribution is not normal and a symmetric interval misstates it in a known direction, so the reported errors for those five should be read as describing curvature rather than as confidence intervals. The case worth stating explicitly is γ_{E}, whose estimate of 0.9990 lies 0.07 standard errors from its upper bound of one: the data are consistent with no per-round decay in the offer at all, which is the reading Section 7.7 attaches to the round-cap counterfactual. Table M.6 reports the position of every estimate within its box, together with the condition number.

An over-identification test rejects, and the useful thing is to say so and explain what it does and does not establish. The full-dimensional statistic is not available: forming it requires inverting $\widehat{\Omega}$ across all 1,320 bins, and a bootstrap covariance built from 499 resamples has rank at most 498, so any such quantity would live on a subspace and its degrees of freedom would not be what they appear. Averaging each moment set to five bins reduces the problem to 79 dimensions, where the bootstrap supports an inverse. The statistic is then 721,734 against 53 nominal degrees of freedom. I read it as a descriptive magnitude rather than a calibrated test: the estimates minimize a differently weighted criterion over the original bins, so the aggregated statistic does not inherit a χ^2 reference distribution. It rejects at any conventional level.

That is in part what an over-identification test does to an estimated model at $n = 5.4$ million. The auxiliary statistics are estimated so precisely that a discrepancy far too small to matter economically is still many standard errors wide, and a consistent test rejects any fixed approximation error once the sample is large enough. But that argument is available to every paper that estimates a model and establishes only the unsurprising part, that the model is not literally true. The informative question is where the statistic comes from and whether it reaches the estimates, and both can be answered directly.

Table M.3 answers the first. It reports each profile's residual in its own units alongside its share of J, the two being the same misfit priced on different scales. The distribution is concentrated

in two profiles. The abandonment hazard by round supplies 76.6% of the statistic on its own. The model abandons too few applications early in prosecution and too many late, placing the hazard at 0.14 against 0.22 at $r = 1$ and at 0.48 against 0.29 at $r = 4$. Filed scope by the number of rounds an application takes supplies 15.5%: it rises with the number of rounds in both series, almost twice as steeply in the model. Of the remainder, the grant rate by filed scope contributes 4.2% and the grant rate by leniency 3.0, and no other set's share exceeds three percent in absolute value. Outside the two profiles, the largest discrepancy anywhere in the objective is 0.18 in the grant rate, at the narrowest filed-scope bin, and the root-mean-square discrepancy of every other profile is under 0.03 in its own units. The statistic is not diagnosing a model that fits nothing precisely. It is diagnosing two profiles that fit badly, against fourteen that fit closely.

Table M.4 answers the second, and is the reason the first answer matters. It reports the sensitivity of the estimates to the auxiliary statistics in the sense of Andrews et al. (2017), as the movement in each parameter induced by shifting an entire profile by one bootstrap standard deviation, expressed in that parameter's own standard errors. The two sets that supply 92% of the statistic are the two least influential by this measure. Were either wrong by a full bootstrap standard deviation in every bin, no parameter would move by more than 0.007 standard errors, against movements of up to 2.2 standard errors for the profiles that do the identifying work. What the two results establish jointly is leverage, not immunity. The observed hazard residual is not one bootstrap standard deviation but several hundred of them, the residual table's own scale column putting the largest bin at $597\,\hat{\sigma}$, so multiplying the local sensitivity through the observed misfit does not bound the displacement of $\hat{\theta}$, and I do not claim that it does. What the sensitivity shows is that, per unit of misfit, the estimates respond to these two sets far less than to the sets that identify them: the rejection is driven almost entirely by moments with little influence on the estimates. The exposure that remains is concentrated in the profile the model fails. The round caps and the response to a stricter office, whose grant loss is applicants stopping sooner, operate directly on this exit decision, and their magnitudes are the most specification-conditional numbers Section 7.7 reports. The standard errors are unaffected for a separate reason, requiring $\widehat{\Omega}$ only through a 26×26 projection that sits well inside the bootstrap's rank.

Two further points follow from the same table. First, with ν held fixed as the numeraire, the sensitivity covers the other 26 parameters, and the abandonment hazard by round leads none of them. Second, the sensitivity ranking does not follow the weights. The grant rate by leniency carries the largest weight in the objective and leads no parameter, while filed scope by leniency, with unit weight, leads 13 of the 26. Influence on the estimates is a property of the Jacobian and the covariance together, not of the weight alone, which is why the assignment in Table M.7 is stated as a heuristic and this table as the computed version of it.

A profile demeaned across quality, such as rounds by quality, has no level for a uniform shift

to change, so Table M.4 cannot show what it contributes. Table M.5 therefore reports two further measures for c_q, c_{q^2} and β_q: the rise in the standard error when one moment set is dropped, and the movement of the estimate when a profile is tilted across quality.

Sampling uncertainty in the counterfactuals of Section 7.7 is not propagated from $\hat{\theta}$. What Appendix M.4 reports is their sensitivity to the two calibrated constants of the offer.

Table M.2: *Model estimates and sandwich standard errors*

Parameter	Estimate	(SE)	$\lvert t\rvert$
c_0 ‡	0.0090	(0.0006)	16.0
q_{unc}	0.3907	(0.0215)	18.2
α_q	0.0564	(0.0005)	107.3
α_1	0.2911	(0.0113)	25.8
α_2	0.0618	(0.0104)	6.0
b_r	0.0332	(0.0014)	23.8
κ	0.4584	(0.0084)	54.4
c_s	0.2954	(0.0094)	31.3
c_r ‡	0.2485	(0.0020)	126.8
σ_c ‡	0.6485	(0.0031)	209.2
σ_{grant}	0.4086	(0.0013)	324.0
μ	0.4677	(0.0044)	106.2
ν	3.7587	–	–
β_s	−0.0230	(0.0042)	5.4
β_q	−0.0186	(0.0018)	10.5
b_0	0.5866	(0.0020)	293.0
σ_s	0.2878	(0.0010)	296.7
γ_{A}	0.7152	(0.0338)	21.1
γ_s	0.7671	(0.0120)	64.1
γ_{E} ‡	0.9990	(0.0140)	71.2
c_q	0.1337	(0.0061)	22.0
c_{q^2}	0.2123	(0.0191)	11.1
α_3	0.0042	(0.0003)	12.7
$\gamma_{s\text{-sl}}$ ‡	0.0003	(0.0424)	0.0
γ_q	−0.0084	(0.0017)	5.0
b_q	0.0392	(0.0018)	21.6
b_{0q}	0.0064	(0.0029)	2.2

Notes: Indirect-inference sandwich standard errors, equation (35), with $\widehat{\Omega}$ from a block bootstrap over examiners on 499 resamples and the finite-simulation factor $(1+1/S)$ at $S = N_{\text{sim}}/n = 1.30$. $|t|$ is against zero. The Jacobian is a central difference at 2% of each parameter's searched interval, one-sided inward where the central step would leave it. ν is pinned as a normalization and excluded from the inversion. The retained block has full rank (26). ‡ Estimate within 1% of a bound of the searched interval: a symmetric interval misstates a boundary estimate in a known direction.

Table M.3: *Moment fit in natural units, and the decomposition of J*

			Residual, own units			
Moment set	Unit	Bins	RMS	Max	Max/$\hat{\sigma}$	% of J
Abandonment hazard by round	prob.	5	0.1090	0.1937	597	76.6
Filed scope by round (resid.)	scope pctile	5	0.0593	0.0932	108	15.5
Granted scope by filed scope (resid.)	scope pctile	98	0.0232	0.0556	41	2.4
Grant rate by filed scope (partialled)	prob.	98	0.0267	0.1776	40	4.2
Grants by revision round	share of apps	12	0.0034	0.0096	37	−2.8
Scope-leniency slope by round	pctile/quintile	4	0.0103	0.0187	28	1.0
Grant rate by leniency	prob.	99	0.0191	0.0725	18	3.0
Var(grant) by leniency	prob.^{2}	99/100	0.0145	0.0598	18	0.2
Rounds by quality (resid., demeaned)	rounds	100	0.0136	0.0756	14	−0.0
Allowance rate by quality, rounds 0 and 1	prob.	200	0.0045	0.0368	10	−0.5
Var(filed scope) by quality	pctile^2	100	0.0011	0.0043	10	0.2
Filed scope by quality (resid.)	scope pctile	100	0.0025	0.0144	8	−0.0
Granted scope by quality (resid.)	scope pctile	100	0.0021	0.0112	6	0.1
Granted scope by leniency (resid.)	scope pctile	99	0.0140	0.0581	5	0.1
Grant rate by quality (resid.)	prob.	100	0.0031	0.0129	5	0.1
Filed scope by leniency (resid.)	scope pctile	99/100	0.0050	0.0212	4	0.1

Notes: Residuals are $m^{\text{sim}}(\hat{\theta}) - m^{\text{data}}$, in each profile's own units. Profiles marked *resid.* are residualised on the fixed-effect set before binning, so their level is a deviation. A bin count of the form a/b means that $b-a$ bins carry no sampling information (the bootstrap never moves them, or the bin is outside the support on one side) and are excluded. $\hat{\sigma}$ is the bin's block-bootstrap standard deviation over 499 examiner resamples, which at $n = 5.4$M is very small: the Max/$\hat{\sigma}$ column is the quantity J aggregates, and the two residual columns are the same misfit priced in economic units. The final column apportions the coarsened $J = 721,734$ of Appendix M.2 across the sixteen moment sets as $\sum_{i \in b} g_i(\widehat{\Omega}^{-1} g)_i / J$, which sums to 100 by construction and admits negative entries where one set's residual is partly accounted for by another's. Rows are ordered by Max/$\hat{\sigma}$.

Table M.4: *Sensitivity of the estimates to the auxiliary statistics*

	Leading		Second	
Parameter	Moment set	$\Delta\hat{\theta}/\mathrm{SE}$	Moment set	$\Delta\hat{\theta}/\mathrm{SE}$
c_0	Allowance rate by quality, rounds 0 and 1	+0.66	Grant rate by quality (resid.)	−0.38
q_{unc}	Filed scope by leniency (resid.)	+0.15	Allowance rate by quality, rounds 0 and 1	−0.08
α_q	Filed scope by leniency (resid.)	−0.26	Allowance rate by quality, rounds 0 and 1	+0.25
α_1	Allowance rate by quality, rounds 0 and 1	−0.42	Grant rate by quality (resid.)	+0.23
α_2	Filed scope by leniency (resid.)	+0.33	Filed scope by quality (resid.)	−0.31
b_r	Allowance rate by quality, rounds 0 and 1	+0.64	Filed scope by leniency (resid.)	+0.53
κ	Filed scope by leniency (resid.)	−0.95	Filed scope by quality (resid.)	+0.32
c_s	Allowance rate by quality, rounds 0 and 1	+0.92	Filed scope by leniency (resid.)	−0.55
c_r	Granted scope by leniency (resid.)	+0.17	Grant rate by leniency	+0.13
σ_c	Allowance rate by quality, rounds 0 and 1	−0.30	Filed scope by quality (resid.)	+0.17
σ_{grant}	Filed scope by leniency (resid.)	−0.59	Grant rate by leniency	+0.52
μ	Filed scope by leniency (resid.)	−0.82	Allowance rate by quality, rounds 0 and 1	−0.61
β_s	Filed scope by leniency (resid.)	−1.39	Filed scope by quality (resid.)	+0.18
β_q	Filed scope by leniency (resid.)	+0.36	Filed scope by quality (resid.)	−0.22
b_0	Filed scope by leniency (resid.)	+0.21	Grant rate by leniency	+0.11
σ_s	Var(filed scope) by quality	+1.19	Filed scope by leniency (resid.)	+0.63
γ_{A}	Filed scope by leniency (resid.)	+1.48	Filed scope by quality (resid.)	−0.71
γ_s	Granted scope by leniency (resid.)	−1.62	Granted scope by quality (resid.)	+0.66
γ_{E}	Granted scope by leniency (resid.)	+2.21	Granted scope by quality (resid.)	−0.41
c_q	Filed scope by leniency (resid.)	−0.06	Granted scope by leniency (resid.)	−0.05
c_{q^2}	Grant rate by quality (resid.)	+0.17	Granted scope by leniency (resid.)	−0.12
α_3	Grant rate by quality (resid.)	+0.17	Granted scope by leniency (resid.)	−0.14
$\gamma_{s\text{-sl}}$	Granted scope by leniency (resid.)	−1.62	Filed scope by leniency (resid.)	−0.27
γ_q	Granted scope by leniency (resid.)	−1.38	Filed scope by quality (resid.)	+0.21
b_q	Filed scope by leniency (resid.)	−0.20	Granted scope by leniency (resid.)	−0.10
b_{0q}	Filed scope by leniency (resid.)	+0.18	Granted scope by leniency (resid.)	+0.07

Notes: Sensitivity of the estimates to the auxiliary statistics, following Andrews et al. (2017). Entries are $\Sigma = (G'G)^{-1}G'$ evaluated at $\hat{\theta}$, with G the binding-function Jacobian expressed in the criterion's own coordinates, each bin scaled by the square root of its weight in the objective, so the map is that of the weighted estimator actually used. The perturbation shifts every bin of one profile by one block-bootstrap standard deviation, and entries are expressed in units of that parameter's own standard error. An entry of 1.0 therefore means: were the model wrong about that entire profile by one bootstrap standard deviation, the estimate would move by one standard error. This is the inverse of the map in Figure M.1, which reports how strongly each moment *responds* to a parameter. A moment can respond sharply and still contribute little to identification if another parameter moves it more. Every parameter with a nonzero Jacobian column has a row except ν, which is held fixed as the numeraire of the payoff scale (Appendix M.2), including those at a bound, whose rows carry the descriptive reading of Appendix M.1, not a test.

Table M.5: *Local information on the quality factor of the stakes*

	SE without the set / SE			Tilt across quality, $\Delta\hat{\theta}$/SE		
Moment set	c_q	c_{q^2}	β_q	c_q	c_{q^2}	β_q
Filed scope by quality (resid.)	1.30	1.14	7.96	−0.15	−0.13	+2.00
Rounds by quality (resid., demeaned)	1.19	1.06	1.00	+0.48	+0.12	−0.01
Grant rate by quality (resid.)	1.05	1.06	1.00	+0.25	−0.13	−0.28
Var(filed scope) by quality	1.01	1.02	1.02	+0.07	−0.03	−0.06
Allowance rate by quality, rounds 0 and 1	1.01	0.99	1.00	−0.64	−0.37	−0.22
Grant rate by filed scope (partialled)	1.00	1.00	1.00	–	–	–
Scope-leniency slope by round	1.00	1.00	1.00	–	–	–
Granted scope by leniency (resid.)	1.00	0.99	1.00	–	–	–
Grant rate by leniency	1.00	1.00	1.00	–	–	–
Filed scope by round (resid.)	1.00	1.00	1.00	–	–	–
Granted scope by filed scope (resid.)	1.00	1.00	1.00	–	–	–
Var(grant) by leniency	1.00	1.00	1.00	–	–	–
Abandonment hazard by round	1.00	1.00	1.00	–	–	–
Grants by revision round	1.00	1.00	1.00	–	–	–
Filed scope by leniency (resid.)	1.00	1.00	1.01	–	–	–
Granted scope by quality (resid.)	0.93	0.94	1.00	+0.02	−0.02	+0.01

Notes: Local information on the quality factor of the stakes, c_q and c_{q^2}, and on the quality term of the filing displacement, β_q, at $\hat{\theta}$, with the Jacobian, the block-bootstrap $\widehat{\Omega}$ and the numeraire of Table M.4. The first three columns give the standard error from equation (35) when every bin of one moment set is removed, relative to the standard error with all sets. Because the objective does not weight the moments efficiently, removing a set can lower a standard error, as removing granted scope by quality does for c_q. The last three columns give the movement of the estimate, in that parameter's own standard errors, when one profile is shifted by one block-bootstrap standard deviation times a weight that runs linearly from −1 in the lowest quality bin to +1 in the highest, within each round for the allowance rate. A dash marks a profile that is not indexed by quality. Table M.4 shifts every bin by the same amount instead, and a profile demeaned across quality, such as rounds by quality, has no level for that shift to change. Of the allowance-rate tilt, round 1 moves c_q by −0.68 standard errors and round 0 by +0.04. Sets are ordered by the first column.

Table M.6: *Local identification diagnostics at the estimate*

Parameter	$\hat{\theta}$	Position in bounds	Inference
c_0	0.009001	0.0%	at bound: SE from a forward difference
q_{unc}	0.3907	39.1%	interior
α_q	0.05643	94.1%	interior
α_1	0.2911	31.6%	interior
α_2	0.06183	3.1%	interior
b_r	0.03318	9.7%	interior
κ	0.4584	28.2%	interior
c_s	0.2954	14.8%	interior
c_r	0.2485	99.4%	at bound: SE from a backward difference
σ_c	0.6485	99.8%	at bound: SE from a backward difference
σ_{grant}	0.4086	79.3%	interior
μ	0.4677	26.3%	interior
ν	3.759	92.0%	held fixed as the numeraire: no SE
β_s	−0.02297	4.9%	interior
β_q	−0.01863	48.1%	interior
b_0	0.5866	86.6%	interior
σ_s	0.2878	29.3%	interior
γ_{A}	0.7152	43.0%	interior
γ_s	0.7671	73.9%	interior
γ_{E}	0.999	99.7%	at bound: SE from a backward difference
c_q	0.1337	44.6%	interior
c_{q^2}	0.2123	26.5%	interior
α_3	0.004217	21.1%	interior
$\gamma_{s\text{-sl}}$	3.062×10^{-4}	0.0%	at bound: SE from a forward difference
γ_q	−0.008443	59.4%	interior
b_q	0.03918	82.6%	interior
b_{0q}	0.006358	55.3%	interior
Condition number of $\Lambda' W \Lambda$: 3.05×10^6			

Notes: Local identification diagnostics at the estimate. $\Lambda = \partial m_{\text{sim}}/\partial \theta'$ is the binding-function Jacobian, computed by finite differences at 2% of each parameter's searched interval with common random numbers, so that differenced simulation noise cancels: central differences, and one-sided inward differences for the parameters whose central step would leave the interval. W is the estimation weight matrix. Columns of Λ are expressed in bound-range units, so a unit move is the same fraction of the searched interval for every parameter and the condition number is not an artefact of scaling. It is computed over the 26 parameters other than ν, which is held fixed as the numeraire, and describes the local curvature of this objective at this estimate under these weights. It is not a rank test and does not establish global identification. "Position in bounds" is $(\hat{\theta} - \underline{\theta})/(\overline{\theta} - \underline{\theta})$. For parameters at a bound a symmetric standard error misstates a boundary estimate in a known direction. Their standard errors are reported and flagged in Table M.2.

Table M.7: *Identification: leading moment for each parameter*

Parameter	Primarily disciplined by	Moment weight
$\sigma_{\text{grant}}, \mu$	Grant rate by leniency, with the variance of the grant outcome by leniency	200, 1
$\alpha_1, \alpha_2, \alpha_3$	Grant rate by filed scope, net of quality and leniency	1
b_r, b_0	Grants by revision round	10
α_q, b_q, b_{0q}	Allowance rate by quality in rounds 0 and 1, with the grant rate by quality	0.828, 1
c_s, c_r, σ_c	Filed scope by quality among grants, which granted scope by quality follows. Table M.4 adds the allowance rate by quality in rounds 0 and 1 for c_s and σ_c	1
c_0	At its lower bound. The scope-leniency slope by round and the grant rate by filed scope, net of quality and leniency	5, 1
c_q, c_{q^2}	Filed scope by quality among grants and revision rounds by quality, with the grant rate by quality for c_{q^2}	1, 2, 1
κ, ν	The level of filed scope, from filed scope by quality and by leniency, relative to the costs, with ν the numeraire	1, 1
q_{unc}	Filed scope by quality, with filed scope by leniency in Table M.4	1, 1
σ_s	Variance of filed scope by quality	1
β_q, β_s	Filed scope by quality and by leniency	1, 1
$\gamma_s, \gamma_{s\text{-sl}}$	Granted scope by filed scope	4
$\gamma_{\text{A}}, \gamma_{\text{E}}$	The narrowing of scope across rounds, filed scope by round and the scope-leniency slope by round	5, 5
γ_q	Granted scope by quality	1
$\gamma_l, \underline{l}$	Calibrated to the round-zero granted-to-filed ratio by leniency (not estimated)	–
$\chi, \sigma_\eta, \bar{q}, \sigma_q, q_{\text{err}}$	Calibrated	–

Notes: Leading moment per parameter. Identification is joint, and Table M.4 reports the computed sensitivity. For the applicant's parameters, the leading sets are those with the largest share of the parameter's local information in the Jacobian of the objective, net of the other parameters. "Moment weight" is the weight the moment set receives in the weighted indirect-inference objective. The offer's leniency gradient γ_l and its floor $\underline{l}$ are calibrated to the round-zero granted-to-filed ratio by leniency rather than estimated, so they have no weight of their own.

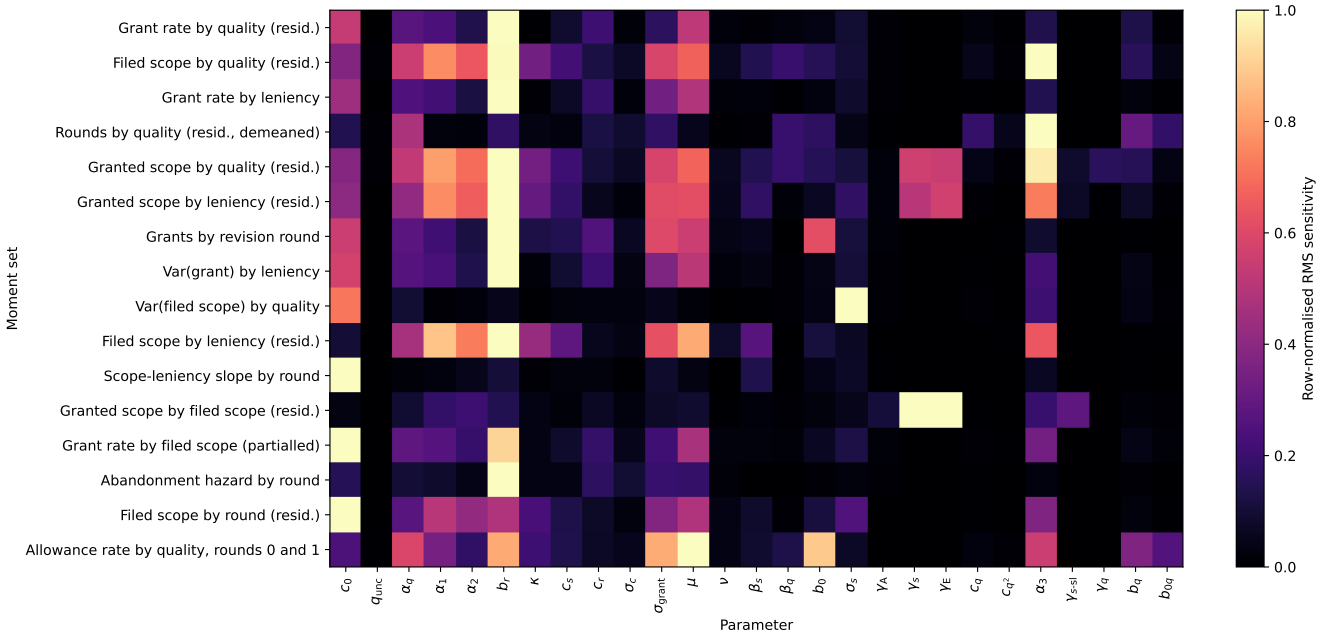
Table M.8: *Identification: parameter blocks and their informative moments*

Parameter block	Identifying moments	Source of identification
Grant rule $\alpha_q, \alpha_1, \alpha_2, \alpha_3, \mu, b_r, b_0, b_q, b_{0q}, \sigma_{\text{grant}}$	Grant rate by leniency, variance of the grant outcome by leniency, grant rate by filed scope net of quality and leniency, grants by revision round, allowance rate by quality in rounds 0 and 1, grant rate by quality	σ_{grant} and μ from the slope and level of the grant rate in leniency and the variance of the grant outcome, the scope penalty from the grant rate in filed scope net of quality and leniency, b_r and b_0 from the round profile, and the three quality terms from the allowance rate in the first two rounds together with the eventual grant rate, b_{0q} shifting only the first round, b_q every round, and the weight of α_q growing with f_r.
Applicant's trade-off $\nu, \kappa, c_0, c_s, c_r, c_q, c_{q^2}, \sigma_c, q_{\text{unc}}$	Filed scope by quality and by leniency, granted scope by quality, the scope-leniency slope by round, grant rate by filed scope net of quality and leniency, allowance rate by quality in rounds 0 and 1, revision rounds by quality, grant rate by quality	Locally, the costs of a round and the value of scope are informed mostly by the scope applicants file. A broader claim takes more rounds, so the cost parameters move the level of filed scope, and the quality factor of the stakes moves how filed scope varies with quality. The level of filed scope relative to these costs gives the value of scope, with ν the numeraire, and κ is only weakly separated from c_s. q_{unc} governs how closely requests track true quality. c_0 is at its lower bound.
Filing and granted scope $\sigma_s, \beta_q, \beta_s, \gamma_s, \gamma_{s\text{-sl}}, \gamma_{\text{A}}, \gamma_{\text{E}}, \gamma_q$	Variance of filed scope by quality, filed scope by quality and by leniency, granted scope by filed scope, granted scope by quality and by leniency, filed scope by round, scope-leniency slope by round	The filing taste from the dispersion of filed scope within a quality bin, the displacement terms from the slopes of filed scope in quality and leniency, the base offer and its slope in filed scope from the granted-to-filed ratio, the decay and the concession rate from the narrowing of scope across rounds, and γ_q from granted scope by quality. γ_l and $\underline{l}$ are calibrated.

The complier shares by quality and by filed scope estimated in Section 5.4 carry no weight in the objective. The model must reproduce them through the mechanism above without having been fitted to them.

Figure M.1 displays the full empirical moment-sensitivity matrix underlying this assignment, evaluated at the estimate by finite differences under common random numbers.

Figure M.1: *Moment sensitivity to parameters (empirical Jacobian)*



Notes: Row-normalized root-mean-square sensitivity of each targeted moment set (rows) to each estimated parameter (columns), $\partial m^{\text{sim}}/\partial\theta'$ evaluated at $\hat{\theta}$ by finite differences under common random numbers, central except for the five estimates on a bound, where the difference is one-sided inward. Brighter cells mark the parameters to which a moment set's simulated counterpart responds most strongly. This is the empirical analogue of the leading-moment assignment in Table M.7.

M.3 Fit

The fit is reported in two forms. Table M.9 gives the level and the slope of every binned profile that is conditioned on quality or on filed scope, data against model. Most of those profiles are flat or near-flat, so two numbers exhaust them and a panel would show sampling noise around a level. Figures M.2, M.3 and M.4 then plot the profiles whose *shape* carries information: the three that describe the grant rate, the four that describe scope and prosecution length, and the allowance rate by quality in the first two rounds. The remaining components of the objective are slopes computed across revision rounds, or hazards, which do not read as profiles at all.

Start with what the model gets right. It reproduces the level moments the offer mechanism targets to within a tenth of a percentile of scope: mean granted scope (0.426 in both series), mean filed scope (0.545 in both), and the scope concession (0.119 in both, Table M.9). Every profile that is flat in quality in the data is flat in quality in the model, and the grant rate matches to within a tenth of a percentage point (0.711 against 0.710). The grant rate falls in filed scope net of quality and leniency (Figure M.2b), and granted scope rises steeply in filed scope, with slopes of 0.74 in the data and 0.76 in the model (Figure M.3a). The grant rate by leniency, which carries the dominant weight, is matched almost exactly in level and slope, and filed scope falls gently across

the leniency range in both series (slopes −0.016 in the data and −0.020 in the model, Figure M.3c). The share of all applications granted in each round matches to within a percentage point in every round (Figure M.2c).

Seven residuals remain, and they are not of equal importance. The first two carry 92% of the coarsened over-identification statistic of Appendix M.2. The largest by far is the abandonment hazard by round. The model abandons too few applications early in prosecution and too many late: 0.14 against 0.22 of the applications still pending at $r = 1$, and 0.48 against 0.29 at $r = 4$, the last round scored, a root-mean-square gap of 0.109 across the five rounds. It accounts for 76.6% of the statistic and is the leading profile for no estimate in the sensitivity of Appendix M.2. Second, mean filed scope by the number of rounds an application takes rises in both series, but the model overstates the rise, from 0.43 at $r = 0$ to 0.72 at $r = 4$, against 0.47 to 0.63 in the data. It accounts for 15.5%. Third, the grant rate by filed scope, net of quality and leniency, is too low among the narrowest filings (0.68 against 0.73 at the 5th percentile), and its hump peaks later, near the 36th percentile against a data plateau from the 12th to the 31st (Figure M.2b). Fourth, granted scope rises in filed scope at the data's average rate, but among the broadest fifth of filings it rises less steeply (0.63 against 0.89) and ends below the data (Figure M.3a). Fifth, revision rounds rise in quality at about the data's average rate (0.116 against 0.099), but the data's rise is concentrated in the top quintile of quality (a slope of 0.50 there against 0.07 below), the model's much less so (0.20 against 0.10), and the model overstates mean rounds (1.34 against 1.27). Sixth, granted scope by examiner leniency is flat in the data below the median examiner and rises in the model, so the model sits too low over strict examiners. Above the median the two rise alike (0.106 against 0.100, Figure M.3b). Seventh, the allowance rate in the second round falls in quality by 1.1 percentage points in the model against 1.9 in the data, while the levels of both rounds and the first round's fall are matched (Figure M.4).

Table M.9: *Moment fit: level and slope of the profiles conditioned on quality and on filed scope*

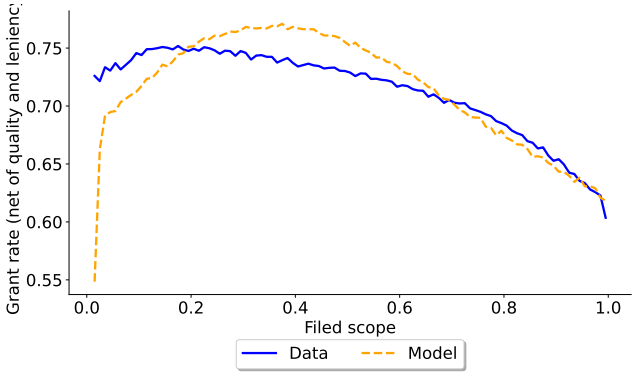
			Level		Slope	
Moment	Binned by	w	Data	Model	Data	Model
Grant rate	Quality	1	0.711	0.710	0.022	0.019
Mean filed scope	Quality	1	0.545	0.545	−0.003	−0.003
Variance of filed scope	Quality	1	0.068	0.068	−0.003	0.000
Revision rounds†	Quality	2	1.271	1.336	0.099	0.116
Granted scope	Quality	1	0.426	0.426	−0.008	−0.008
Allowance rate, round 0†	Quality	0.828	0.114	0.115	−0.015	−0.014
Allowance rate, round 1†	Quality	0.828	0.486	0.487	−0.019	−0.011
Scope concession (filed − granted)	Quality	–	0.119	0.119	0.005	0.006
Grant rate†	Filed scope	1	0.713	0.715	−0.114	−0.096
Granted scope†	Filed scope	4	0.402	0.394	0.740	0.759

Notes: Model column computed from the simulated panel at the estimates (the vector used for the policy counterfactuals of Section 7.7), with the moment definitions of the estimation objective. Data column reports the empirical targets under the v2 sample conventions (core sample 2005–2020, within-year pooled scope percentiles, application and examiner controls partialled out). Scope is the within-year percentile of average claim length. "Level" is the mean across the 100 percentile bins, and "slope" is the coefficient of a linear fit across those bins, with the conditioning variable on the 0–1 percentile scale. Revision rounds are scored net of the controls and relative to their mean, so their level column reports mean rounds instead. Grant rate by filed scope is net of quality and leniency. The allowance rate is the share of applications pending in the round that are allowed in it. w is the weight the moment set carries in the indirect-inference objective (the two allowance rows form one set), and "–" marks a row that is not a moment set. Rows marked † also appear as figures.

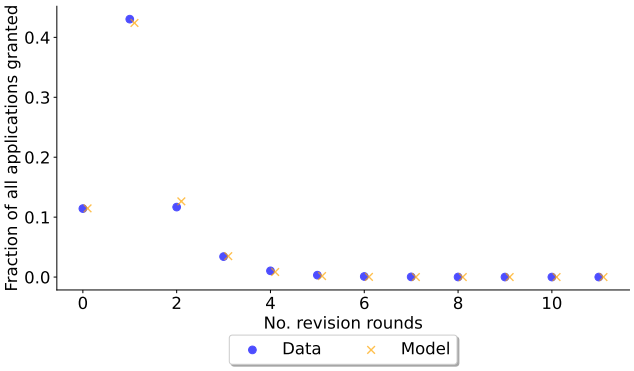
Figure M.2: *Targeted moment fit: grant rate*



(a) By examiner leniency



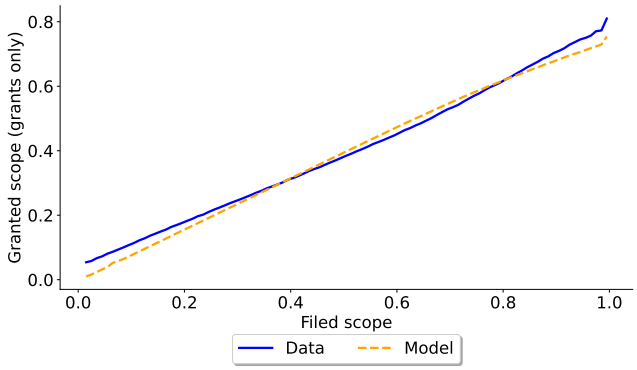
(b) By filed scope, net of quality and leniency



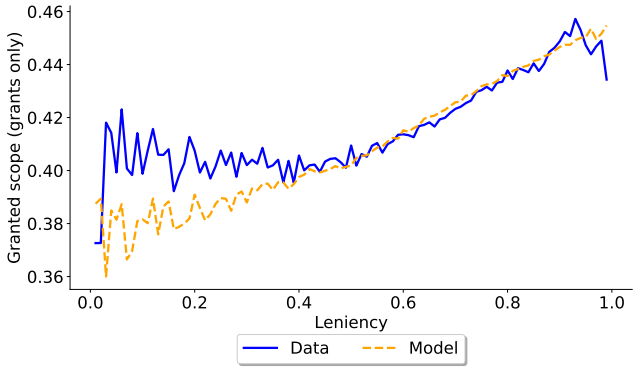
(c) By prosecution round

Notes: Solid line, data. Dashed line, model at the estimates. All three profiles enter the estimation objective. The grant rate by *quality* rises only mildly in both data and model, about two percentage points end to end, and is reported in Table M.9 instead. It can also be read off Figure M.5, which plots it separately within each leniency quintile. Filed scope is a within-year percentile. In Panel b both series are net of quality and leniency, and the data series also of the fixed-effect set of Section 5, as the objective scores the profile, so the panel shows the dependence of the grant rate on filed scope alone. Panel a carries the dominant weight in the objective (Table M.7). Panel c plots the fraction of *all* applications granted at each round, not the round-conditional hazard.

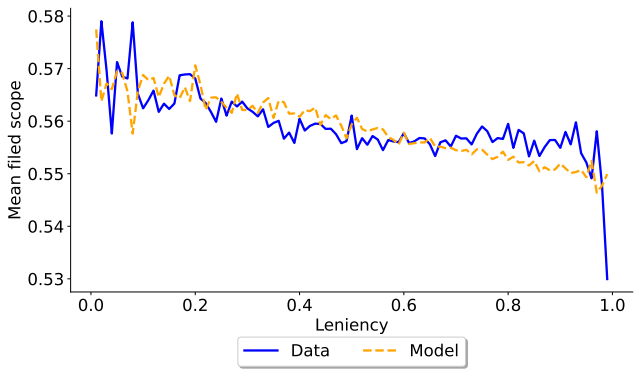
Figure M.3: *Targeted moment fit: scope and prosecution length*



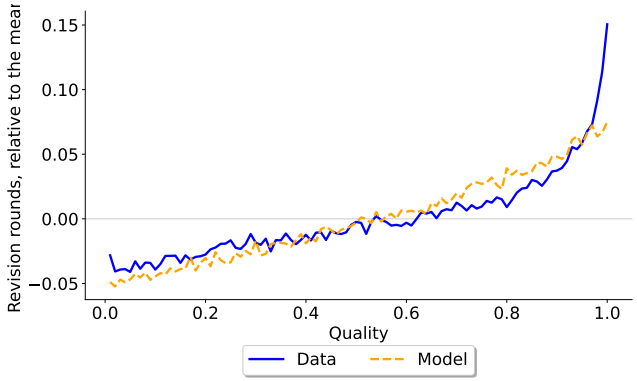
(a) Granted scope by filed scope (residualized)



(b) Granted scope by examiner leniency



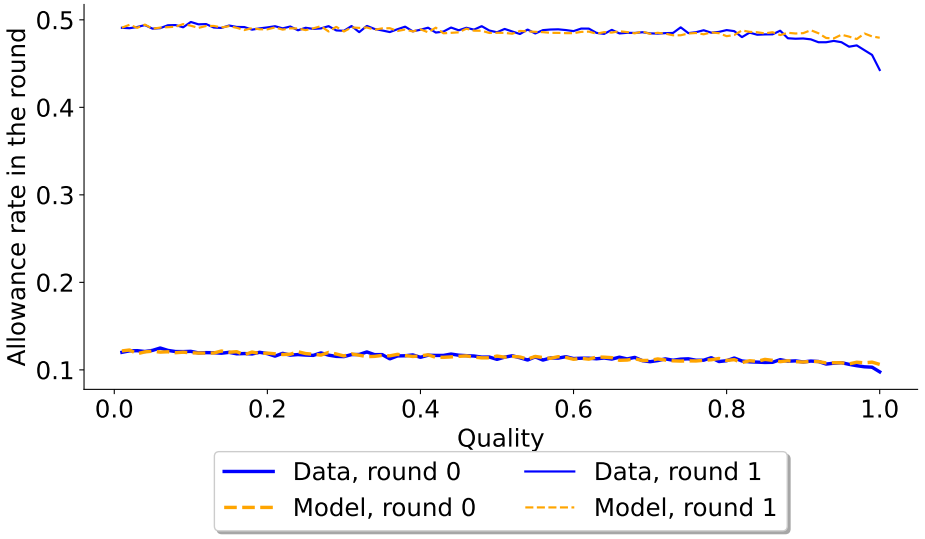
(c) Mean filed scope by examiner leniency



(d) Revision rounds by quality, relative to the mean

Notes: Solid line, data. Dashed line, model at the estimates. All four profiles enter the estimation objective, each plotted as the objective scores it. Scope is the within-year percentile of average claim length, pooled over filed and granted claims, and the data series are residualized on the fixed-effect set. In Panel a filed scope is binned on its residualized axis. Panel d plots revision rounds relative to their mean, and Table M.9 reports mean rounds. Panels a and b condition on a grant, as granted scope is only defined there. Panels c and d are over all applications, matching their estimation targets. The two profiles of filed scope conditioned on *quality*, its mean and its variance, are near-flat in both data and model and are reported in Table M.9 instead. Relative to its sampling error, Panel a carries the largest residual among the profiles plotted here: among the broadest filings the model's granted scope rises less steeply than the data's. In Panel b the data are flat below the median examiner, where the model rises.

Figure M.4: *Targeted moment fit: allowance rate by quality in the first two rounds*



Notes: Solid lines, data. Dashed lines, model at the estimates. The allowance rate in round r is the share of the applications pending in round r that are allowed in it, plotted for $r = 0$ (thick lines) and $r = 1$ (thin lines) against the quality percentile. The two rounds form one moment set, the sixteenth, with weight 0.828 in the objective. In the data the allowance rate falls in quality in both rounds, on a linear fit by 1.5 percentage points in round 0 and by 1.9 in round 1, which is 13% and 4% of the round's allowance rate. The model matches both levels and the first-round fall (1.4 percentage points) and understates the second-round fall (1.1).

Moments outside the objective. Because indirect inference fits what it is shown, the informative diagnostics are the profiles the estimator was not asked to match. One appears in Table M.9: the scope concession by quality, filed minus granted scope among grants. Its level and its rise in quality are matched jointly (0.119 in both series, slopes 0.005 against 0.006) rather than traded against one another. Granted scope by quality is not among the checks, because it is a weighted target of the objective (Section 7.5), where the offer's quality gradient γ_q reproduces its level and its mild decline (0.426 and -0.008 in both series).

Why a loaded schedule rather than a belief-path offer. As Section 7.5 notes, the record contains filed scope at entry and granted scope at issue, and nothing between, so any model of the intra-prosecution concession path is identified only through its endpoint implications. That bounds what the offer's quality gradient can be asked to represent, and it supplies a test. The natural alternative to the reduced-form gradient γ_q is an offer conditioned on the examiner's round-r posterior. On endpoint data that family has exactly one distinguishing implication: the grant-round profile of the within-round quality gradient of the pass-through ratio, b_r. Under Bayesian updating the posterior's weight on true quality is $f_r = (r+1)\kappa/(1+(r+1)\kappa)$ with $\kappa = \sigma_q^2/\sigma_\eta^2$, so $f_0/f_1 = (1+2\kappa)/(2(1+\kappa)) > 1/2$ at every signal noise, and an offer decay $\gamma_E \leq 1$ only shrinks later rounds relative to earlier ones. Every member of the family therefore predicts $|b_0| \geq |b_1|/2$:

the first-round gradient cannot be less than half the second's. Table M.10 reports the estimated profile on the estimator's own construction. The bound is violated: $|b_1|/2 - |b_0| = 0.0044$, two and a half times its standard error of 0.0017. The family is therefore rejected at every $(\sigma_\eta, \gamma_{\mathrm{E}})$, including the calibrated $\sigma_\eta = 0.5$, at which $f_0/f_1 = 5/8$. The rejection is exactly as wide as the family: the bound assumes signals of equal precision across rounds, and it can fail only if the first signal is nearly uninformative while a later one is substantially sharper, a configuration the endpoint record can neither support nor rule out. Two qualifications apply. The constant-gradient form the model carries fails the same profile from the other side, predicting a gradient largest at $r = 0$ and declining where the data are flat at $r = 0$ with a hump at rounds one and two. Neither family fits the profile, so replacing the reduced form with a belief-path offer would add no identifiable content on endpoint data while losing the pooled-slope fit. The pooled slope is the only object both forms identify, it is what the estimate of γ_q fits, and the round profile is reported as a limitation of both. And b_r conditions on a grant at round r, so the profile is a grant-selected object, and the belief-path prediction concerns the same selected object, so the comparison is like for like. The repair is data rather than modeling: the intermediate amendments exist, unparsed, in the application file wrappers.

The device's reach is bounded by the model's own structure. γ_q enters neither the grant rule (16) nor the applicant's problem, whose value function is computed against the full filed claim (Section 7.3), so no counterfactual can move the concession's quality gradient, and none of the exercises in Section 7.7 needs it to. Under the stricter office, requests, persistence and grants do not depend on γ_q, and only granted scope among grants can. As Appendix M.1 shows, the inversion is exact because the offer is proportional to the claim at the estimates. Dividing granted scope by the offer ratio recovers the round-r claim, and grants at the bottom of the scope grid, which concede nothing, are held fixed. Varying γ_q over two standard errors of its estimate moves the narrowing of granted patents under the stricter office, and its split between the stricter decisions and the applicants' response, by less than 1.5% of their values. Over the whole interval searched, $[-0.05, 0.02]$, the narrowing stays at 4.3% and the response's share of it at 87%. The gradient's role is to make the simulated concession schedule match the observed one.

The last untargeted check requires a figure. Figure M.5 plots the grant rate by quality *separately within each leniency quintile*: the joint surface, of which only the two marginals enter the objective. What is at stake here is a specification assumption rather than a hard prediction. The grant rule (16) is additively separable: leniency shifts the threshold and carries no interaction with quality, and Proposition 1 does not go through without that separability. Separability is *imposed*, not tested: the model has no quality-by-leniency parameter, so it cannot fit such an interaction and cannot report that it found none. What this panel can do is weaker but not empty. If a substantial interaction were present in the data, the fitted model would be unable to reproduce the joint

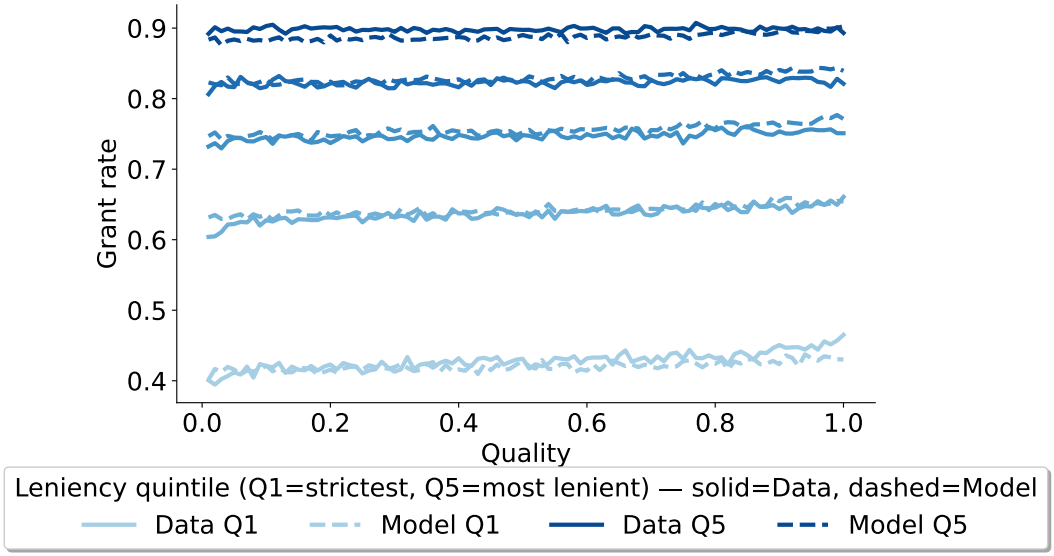
Table M.10: *The concession's quality gradient by grant round: testing the belief-path alternative*

Grant round r	b_r	s.e.	N	clusters
0	−0.0002	0.0016	574,297	8,785
1	−0.0092	0.0011	2,238,709	9,524
2	−0.0085	0.0018	615,460	9,321
3	−0.0062	0.0032	180,796	8,978
4	−0.0018	0.0056	55,114	7,874
5	−0.0067	0.0098	16,963	5,561
Belief-path bound $\lvert b_0\rvert \geq \lvert b_1\rvert/2$: violation $0.0044 > 2 \times$ s.e. $= 0.0035$				

Notes: The Table reports the within-round quality gradient of the concession: for each grant round r (transaction counts at issue), the slope b_r from regressing the pass-through ratio, granted over filed scope clipped to $[0, 2]$ on the granted sample with filed scope ≥ 0.05, on the quality rank centered at one half. Scope percentiles use the estimator's own construction: filed and granted claim lengths ranked against a pooled within-publication-year base and flipped so that higher values denote broader scope. The sample is the 2005–2020 estimation frame. Standard errors are clustered on art unit × publication year. The bound in the bottom row is parameter-free: under Bayesian updating the posterior's weight on true quality satisfies $f_0/f_1 = (1 + 2\kappa)/(2(1 + \kappa)) > 1/2$ for every signal noise, and an offer decay $\gamma_{\mathrm{E}} \leq 1$ only shrinks later rounds, so every belief-path offer predicts $\lvert b_0\rvert \geq \lvert b_1\rvert/2$. The observed violation exceeds twice its standard error, rejecting the family at every $(\sigma_\eta, \gamma_{\mathrm{E}})$.

surface, and the failure would show up as within-quintile tilt against the model's within-quintile profile. The panel shows a mild version of exactly that signature. The data's within-quintile rise in quality flattens as leniency grows, from a slope of 0.038 in the strictest quintile to essentially zero in the most lenient, faster than the probit's own curvature delivers, while the model's slope stays between 0.014 and 0.022 in every quintile. The interaction is modest in level terms, with mean within-quintile gaps below 1.1 percentage points everywhere. I stop short of claiming the discrepancy would appear "here and nowhere else in the fit". An omitted interaction distorts the estimated parameters, and those parameters generate the targeted marginals too, so a misspecification of this kind could be partly absorbed elsewhere rather than surfacing cleanly in this one panel. The evidence is that separability holds in levels and bends in slopes: the five levels line up from 0.43 in the strictest quintile (0.42 in the model) to 0.90 in the most lenient (0.89), the largest pointwise gap is 3.5 percentage points, in the strictest quintile, and the one signed discrepancy is the slope flattening just described. An untargeted check of an imposed assumption can deliver no more. The quintile levels themselves are close to implied by the targeted grant rate by leniency, so the panel's informative content is the within-quintile comparison, not the spacing.

Figure M.5: *An untargeted joint moment: grant rate by quality within leniency quintile*



(a)

Notes: Solid lines, data. Dashed lines, model at the estimates. The joint surface does not enter the estimation objective, only its two marginals do. Applications are split into quintiles of examiner leniency (Q1 strictest, Q5 most lenient) and the grant rate is plotted against the quality percentile within each. Only the extreme quintiles are labeled, and the shading runs light to dark from Q1 to Q5. The model contains no quality-by-leniency interaction, so the within-quintile agreement is not imposed. The spacing of the five levels, by contrast, is close to implied by the targeted grant rate by leniency of Figure M.2a. The complier share, the one untargeted moment that speaks directly to the leniency channel the counterfactuals exploit, is reported in the main text (Section 7.6, Figure 5).

Bandwidth robustness of the complier curves. Table M.11 recomputes both curves of Figure 5 at Gaussian kernel bandwidths $h \in \{0.05, 0.07, 0.10\}$, holding the evaluation grid fixed. Between the 10th and the 90th percentile of filed scope the share rises by 43% to 48% in the data and by 42% to 45% in the model, and across quality it falls by 7% to 8% in the data and by 4% to 5% in the model. The model's dip over narrow filings is present at all three bandwidths, with its minimum between the 25th and the 27th percentile. Its depth is not similarly stable: the fall from the narrowest filings to the minimum shrinks from 1.9 to 1.1 percentage points as the bandwidth widens, the boundary sensitivity any kernel estimate carries at the edge of its support.

Table M.11: *Bandwidth sensitivity of the complier curves, data and model*

	Data			Model		
Bandwidth h	10th pct.	90th pct.	Rise (%)	10th pct.	90th pct.	Rise (%)
Filed scope						
0.05	5.87	8.69	+47.9	6.41	9.28	+44.7
0.07	5.91	8.64	+46.2	6.43	9.20	+43.1
0.10	5.97	8.53	+42.8	6.37	9.04	+42.1
Quality						
0.05	7.18	6.67	−7.1	7.21	6.86	−4.9
0.07	7.20	6.65	−7.6	7.20	6.87	−4.6
0.10	7.22	6.67	−7.6	7.20	6.89	−4.3

Notes: Share of compliers among granted patents at the 10th and 90th percentile of filed scope and of quality, and the rise between them, for the data curve and the model curve of Figure 5 at three Gaussian kernel bandwidths h. $h = 0.07$ is the bandwidth of the figure. Data: the profile estimated in Section 5.4, scaled so that its grant-weighted average equals the aggregate share. Model: the share of the model's own granted applications that the reference examiner would refuse, from the simulated panel at the estimates paired with the same applications under the reference examiner, the examiner whose grant rate equals the average (n = 7,000,000), expressed in the data's units by one factor for both axes. Levels and rises are in percent.

M.4 Sensitivity to the offer's leniency gradient and floor

The offer's leniency gradient γ_l and its floor $\underline{l}$ are calibrated rather than estimated (Appendix M.2). Table M.12 varies them one at a time around the calibrated values, $\gamma_l = 0.1407$ by an assumed ± 0.05 and $\underline{l} = 0.5$ by ± 0.1, without re-estimating the model. The band is a range of calibrations, not a sampling interval. Both constants enter only the post-grant offer, so grants, rounds and requests are the same in every row, and only granted scope moves. Under the shift to today's 45th percentile, granted patents become 4.2% to 4.4% narrower across the rows, the stricter decisions alone account for 10.3% to 15.4% of the narrowing, and the change in requests for 74.97% to 79.3%. The lower end, at $\gamma_l = 0.1907$, is just below three quarters. Under the shift to the 25th percentile, granted patents become 21.6% to 22.6% narrower, and the applicants' response accounts for at least 83% of the narrowing in every row. The linear offer without a floor, the earlier specification, gives a fall of 4.4%, 13.9% of it from the stricter decisions alone. The parameters are held at their point estimates, so the rows carry no sampling uncertainty.

M.5 Counterfactual losses by filed scope

Table M.13 splits the change in grants under the stricter office of Section 7.7 by group of today's filed scope. For each group it reports the total change, the part the stricter examiners cause at today's requests and persistence, and the part that comes from the applicants' response.

Table M.12: *Sensitivity of the stricter-office results to the offer's leniency gradient and floor*

	Median at todayś 45th percentile				Median at todayś 25th percentile			
Offer	Fall	Rule	Request	Response	Fall	Rule	Request	Response
Calibrated ($\gamma_l = 0.1407$, $\underline{l} = 0.5$)	−4.3	12.9	77.1	87.1	−22.1	14.3	72.3	85.7
$\gamma_l = 0.0907$	−4.2	10.3	79.3	89.7	−21.6	11.5	74.6	88.5
$\gamma_l = 0.1907$	−4.4	15.4	75.0	84.6	−22.6	16.8	70.2	83.2
$\underline{l} = 0.4$	−4.3	13.5	76.6	86.5	−22.3	15.2	71.6	84.8
$\underline{l} = 0.6$	−4.3	11.9	77.9	88.1	−21.8	12.8	73.5	87.2
No floor ($\underline{l} = 0$, linear offer)	−4.4	13.9	76.3	86.1	−22.4	15.9	71.0	84.1

Notes: Sensitivity of the stricter-office results to the two calibrated constants of the examiner's offer, the leniency gradient γ_l and the leniency floor $\underline{l}$, one at a time around the calibrated values, without re-estimating the model. Both enter only the post-grant offer, so grants, rounds and requested scope are the same in every row by construction and only granted scope moves. For a stricter office at today's 45th and 25th leniency percentile: *Fall* is the percentage change in mean granted scope among grants, *Rule* the share of that fall produced by the stricter decisions alone at the status-quo policy, *Request* the share produced by the change in requested scope, and *Response* the share produced by the applicants' response as a whole, all in percent. The last row is the linear offer without a floor, reported for comparison with the earlier specification.

Table M.13: *Change in grants under the stricter office, by group of filed scope*

	Share of filings	Change in grants (%)					
		Median at today's 45th			Median at today's 25th		
	(%)	Total	Rule	Response	Total	Rule	Response
Narrowest group	15.8	−5.0	−3.0	−2.0	−30.6	−18.1	−12.6
Second group	24.0	−3.4	−2.7	−0.7	−22.1	−16.9	−5.2
Third group	25.1	−3.4	−3.0	−0.4	−21.8	−18.9	−2.9
Fourth group	18.2	−4.1	−3.6	−0.5	−25.0	−21.6	−3.4
Broadest group	16.8	−5.2	−4.0	−1.2	−30.8	−23.8	−7.0
All filings	100.0	−4.0	−3.2	−0.9	−25.2	−19.5	−5.7

Notes: Change in the number of grants against today's grants in each group, in percent, by group of today's filed scope. The groups are cut at the data's quintiles of filed scope, so their shares of the model's filings, in the first column, are not equal. The stricter office shifts every examiner's leniency down by the same amount, which moves the median examiner to today's 45th percentile (the main shift) or 25th percentile. "Rule" keeps today's requests and decisions to continue or abandon and applies only the stricter examiners. "Response" is the rest of the total, from applicants choosing their requests and their persistence anew, so the two parts add up to the total up to rounding. The row for all filings equals the grants columns of Table 1. Filed inventions are held fixed, and every difference is paired on one simulated population.

N Limitations of the design

The qualifications attached to individual results in the main text share a few common causes, collected here. What follows is not a list of things left for future work with this data. Each is a limit of the design itself, and none is repaired by more observations of the same kind.

Ordinal scope measure. Scope is a within-year percentile rank. Every result defined through the ordering alone, the scope bins and the complier profiles over them, is exactly invariant to any monotone rescaling of claim length. Every *quantity* that sums scope across applications, including the mean requested and granted scope of Section 7.7, is normalization-dependent. Appendix J finds that exposure material for the shares of the granted-scope budget, and whether the narrowing of granted patents under a stricter office moves across scales is not measured. No cardinal, legally validated measure of claim scope exists, and constructing one is not a data problem.

Monotonicity, scope choice, and entry. *Individual* monotonicity is untestable. The average monotonicity the data can test permits offsetting compliers and defiers, and the additive-leniency threshold rule of Section 5.1 is maintained as the assumption that delivers it. Filed scope is chosen rather than randomized and cannot be separated from unobserved applicant type, so the scope gradient of Section 6.2 is a composition fact, not an effect of scope. And entry is not modeled: no filing-rate moment pins down the outside option, so the set of filed inventions is fixed by assumption.

Which results the model's limits reach. The counterfactual results are computed inside the model, so they are conditional on the estimated grant rule, the filing and offer equations, and the fixed set of filed inventions. Within that, the limits reach them unequally. The result least exposed is that granted patents narrow mostly because applicants ask for less. Under the stricter office the response carries 87% of the fall in granted scope among grants, requests account for 77% of the fall, and the exit decision, the margin the model fits worst, accounts for 3%. Requests are solved under the ratchet, with the applicant's belief updated along one path of claims (Section 7.3). The applicants' part of the grant loss is more exposed. It is 22% of the loss and works through applicants who stop sooner, so it rests on the abandonment hazard, which carries three quarters of the over-identification statistic, and on applicants who never learn which examiner they were assigned. Within art unit and year, the data's exits after the first rejection respond 1.5 times as strongly to the examiner's leniency as the model's. The comparison with an office that tests quality rests on the weight that office gives quality, set so that quality varies the acceptance index as much as the scope penalty does today.