The Inspector Lottery That Isn't: Small-Sample Bias in Examiner Leniency Designs Applied to England's Planning Appeals

APEP Autonomous Research* @olafdrw

April 9, 2026

Abstract

Examiner leniency instruments have become a workhorse of applied economics, yet their reliability hinges on sufficient cases per examiner. I test this design in England's Planning Inspectorate, scraping 2,227 appeals and extracting inspector identities from decision letter PDFs—yielding 720 inspectors, but only 860 cases (198 inspectors) survive minimum-cases restrictions for leave-one-out estimation. The leniency instrument produces a *negative* first stage ($\hat{\gamma} = -0.054$, $F = 5.6$), driven by mechanical mean reversion: the median inspector decides only 2.2 cases per cell. Lagged leniency—the same inspector's prior-year allow rate—is strongly positive (0.225, $p < 0.001$), confirming persistent inspector styles that cannot be recovered from thin within-cell variation. The results demonstrate minimum data requirements for credible examiner designs and introduce a scalable pipeline for linking planning appeals to individual decision-makers.

JEL Codes: R31, R52, C26

Keywords: examiner leniency, instrumental variables, planning appeals, housing supply, small-sample bias, England

*Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch (cumulative: 4h 39m).

1. Introduction

Examiner leniency designs have become a workhorse of applied economics. From judges and criminal sentencing (Kling, 2006; Dobbie et al., 2018) to patent examiners (Sampat and Williams, 2019) and disability claims (Maestas et al., 2013), the logic is elegant: quasi-random assignment of cases to decision-makers with different propensities to approve creates exogenous variation in outcomes. But elegance depends on precision. When each examiner decides hundreds of cases, leave-one-out leniency scores are stable estimators of latent strictness. When each examiner decides three or four, they are noise.

This paper documents what happens when an examiner design meets a small sample. England's Planning Inspectorate (PINS) assigns roughly 300 professional inspectors to adjudicate planning appeals from a national pool, using workload-based allocation that developers cannot influence. The institutional design is textbook for a leniency IV: quasi-random assignment, large examiner pool, binary outcomes. But when I scrape 2,227 decided appeals from the PINS case portal and extract inspector identities from 1,451 decision letter PDFs, the median inspector appears in only 2.2 cases in my sample—far below the hundreds typical of Dobbie et al. (2018) or Sampat and Williams (2019).

The leave-one-out leniency score produces a *negative* first stage: inspectors with higher allow rates among their other cases are *less* likely to allow the focal case ($\hat{\gamma} = -0.054$, $p = 0.018$). This sign reversal is inconsistent with a standard examiner design but entirely consistent with mechanical mean reversion. When an inspector has three cases in a cell and two are allowed, the leave-one-out score for the third case is 1.0—not because the inspector is lenient, but because the other two cases happened to be allowed. The negative first stage reflects the fact that extreme leave-one-out scores are driven by sampling noise, not persistent styles.

Two pieces of evidence distinguish mean reversion from genuine compensating assignment. First, balance tests pass cleanly: inspector leniency is uncorrelated with filing lag and case type within LPA–case-type–year cells (Table 2, Panel B). If PINS were strategically assigning lenient inspectors to harder cases, we would expect leniency to predict observable case difficulty. Second, *lagged* leniency—the same inspector's allow rate in the prior year—is a strong positive predictor of current-year outcomes (coefficient 0.225, $p < 0.001$). Inspectors *do* have persistent decision-making styles; the problem is that contemporaneous leave-one-out scores cannot recover these styles from three or four cases per cell.

This paper contributes to a growing literature on the practical requirements of examiner designs (Frandsen et al., 2023; Borusyak and Hull, 2023). The planning inspector setting is ideal in principle: quasi-random assignment from a national pool, no geographic specialization,

standardized case procedures. The failure of the first stage is therefore informative—not about the institution, but about the sample. The full PINS case archive contains over 100,000 decided appeals; the design will work when the full population is assembled. This paper demonstrates both the promise of the setting and the minimum data requirements for credible implementation.

The paper also introduces a novel data source. The PINS Appeal Case Portal provides structured case-level data, but inspector identities are locked inside decision letter PDFs in a standardized header format. I develop an automated extraction pipeline that recovers inspector names from 79% of PDFs, yielding 720 unique inspectors—the first researcher-accessible dataset linking planning appeals to individual decision-makers in England. This dataset, and the extraction methodology, are contributions independent of the IV results.

The remainder of the paper proceeds as follows. Section 2 describes the institutional setting. Section 3 presents the data construction. Section 4 develops the identification strategy and explains why the first stage fails. Section 5 reports results. Section 6 discusses implications.

2. Institutional Background

The planning system. England's planning system operates through roughly 300 local planning authorities (LPAs), typically corresponding to district or borough councils. When a developer submits a planning application, the LPA evaluates it against the local plan and national policy, granting or refusing permission. Approximately 87% of applications are approved at the initial stage (Department for Levelling Up, Housing and Communities, 2024).

The appeals process. A developer whose application is refused may appeal to the Planning Inspectorate (PINS), an executive agency of the Department for Levelling Up, Housing and Communities. PINS employs approximately 300–400 full-time equivalent planning inspectors who adjudicate appeals through written representations (75% of appeals), hearings, or public inquiries. The inspector's decision is binding, though it can be challenged via judicial review on narrow legal grounds.

Inspector assignment. PINS assigns inspectors from a *national* pool based on workload availability, required procedure type, and specialist expertise—not geographic proximity to the appeal site. This design was implemented specifically to prevent regulatory capture. The developer cannot request a specific inspector, object to the assignment, or predict who will be assigned. Inspectors bring different professional backgrounds: chartered surveyors (MRICS, FRICS), architects (RIBA), and town planners (MRTPI, FRTPI), contributing to

heterogeneous decision-making styles.

Why this matters for housing. England faces a persistent housing shortage. The gap between housing supply and government targets has driven house prices to over eight times average earnings (Hilber and Vermeulen, 2016). Roughly 20,000 appeals are filed annually, and the overall allow rate—approximately 30% in the sample period—determines whether a substantial flow of refused developments are ultimately built. The Planning and Infrastructure Bill currently before Parliament proposes reforms to the appeals process, including expanded inspector capacity.

3. Data

PINS appeal cases. I scrape case-level data from the PINS Appeal Case Portal (`acp.planninginspectorate.gov.uk`), a publicly accessible database of all planning appeals in England. For each case, I extract the local planning authority, case type, filing date, decision date, and outcome using CSS selectors confirmed against the portal's HTML structure. I sample 3,500 case IDs from the range 3,220,000–3,360,000 (spanning decisions from approximately 2019 to 2023) and retain 2,500 cases with valid decided outcomes, of which 2,227 have binary outcomes (allowed or dismissed).

Inspector extraction. Inspector identities are not recorded in the portal's structured data. However, every decision letter PDF follows a standardized header: "Appeal Decision / Site visit made on [DATE] / by [INSPECTOR NAME and CREDENTIALS] / an Inspector appointed by the Secretary of State." I download decision letters for 2,000 cases and extract inspector names via regular expressions matching this header format, successfully recovering inspector identities for 1,625 cases (65% of the full sample, 79% of PDFs attempted). Extraction failures arise from non-standard formatting in enforcement notices and third-party decision letters.

Analysis sample. After restricting to cases with inspector names and requiring each inspector to appear in at least three cases (for leave-one-out calculation), the analysis sample contains 860 cases decided by 198 inspectors across 264 LPAs. The overall allow rate is 30.2%.

Housing outcomes. I download Land Registry Price Paid Data for 2019–2024 (6.05 million transactions), aggregating to the district-quarter level: number of new-build transactions, total residential transactions, and median transaction price.

Table 1 reports summary statistics.

Table 1: Summary Statistics

	Mean	SD
Panel A: Case-Level		
Appeal allowed (=1)	0.302	0.460
Inspector leniency (standardized)	0.000	0.986
Inspector leniency (raw)	0.307	0.343
Filing lag (days)	122.092	73.259
Case type: Householder	0.295	
Case type: Planning	0.662	
Case type: Enforcement	0.000	
Case type: Other	0.043	
Panel B: Inspector-Level		
Cases per inspector	4.3	1.8
Inspector allow rate	0.3	0.2
LPAs per inspector	4.2	1.7

Cases: 860. Inspectors: 198. LPAs: 264.

4. Identification Strategy

The leniency instrument. For each case i decided by inspector j in cell c (defined by case type $\times$ decision year), I construct the leave-one-out leniency score:

$$Z_{ij} = \frac{1}{n_{jc} - 1} \sum_{k \in \mathcal{C}_{jc}, k \neq i} \text{Allowed}_k \tag{1}$$

where $\mathcal{C}_{jc}$ is the set of cases assigned to inspector j in cell c, and $n_{jc} = |\mathcal{C}_{jc}|$. I standardize Z_{ij} to mean zero and unit variance within each cell. The first-stage regression is:

$$\text{Allowed}_i = \alpha + \gamma Z_{ij} + \boldsymbol{\delta}_{\ell tc} + \varepsilon_i \tag{2}$$

where $\boldsymbol{\delta}_{\ell tc}$ denotes fixed effects for LPA $\times$ case type and year $\times$ case type. Standard errors are clustered at the LPA level.

The small-sample problem. The credibility of examiner leniency instruments depends on each examiner deciding enough cases within each cell for the leave-one-out score to approximate latent strictness. In the canonical applications, this condition is easily satisfied: Dobbie et al. (2018) have 420,000 cases across 207 judges (roughly 2,000 per judge); Sampat and Williams (2019) have 1.45 million patent applications across approximately 6,000 examiners. In contrast, my sample has 860 cases across 198 inspectors, with a median of 2.2 cases per

inspector in the sample. The typical leave-one-out score is based on one or two other cases.

When the leave-one-out denominator is small, the score is mechanically dominated by sampling variation in the other cases rather than by persistent inspector styles. Consider an inspector with three cases in a cell. If two are allowed and one is dismissed, the leave-one-out scores are 0.5 for the allowed cases and 1.0 for the dismissed case—producing a *negative* correlation between the leniency score and the focal outcome that is entirely mechanical.

Balance. Despite the small-sample problem, the quasi-randomness of inspector assignment can still be tested. If PINS assignment is workload-based and unrelated to case characteristics, the leniency score should be uncorrelated with observable case features within cells. Table 2, Panel B confirms this: leniency is uncorrelated with filing lag ($\hat{\beta} = -1.75$, $p = 0.57$) and case type composition ($\hat{\beta} = -0.010$, $p = 0.62$).

5. Results

5.1 First Stage

Table 2 reports the first-stage results. A one-standard-deviation increase in the leniency score *decreases* the probability of the appeal being allowed by 5.4 percentage points ($p = 0.018$) with interacted fixed effects and by 4.1 percentage points ($p = 0.034$) with additive fixed effects. The first-stage F-statistic is 5.6, well below the conventional threshold of 10.

The negative sign is robust across specifications (Table 4): it persists when using overall (non-cell-specific) leniency, when excluding same-LPA cases from the leniency calculation, and in both householder and full planning appeal subsamples. The leave-one-inspector-out exercise confirms no single inspector drives the result: dropping any of the 20 most prolific inspectors yields coefficients in the narrow range $[-0.061, -0.049]$.

5.2 Monotonicity and Lagged Leniency

The monotonicity condition requires that the allow rate increases across leniency quintiles. It does not: the allow rate is 32.5% in the first (strictest) quintile and 26.7% in the fifth (most lenient)—flat or slightly declining. This pattern is consistent with the mean-reversion interpretation: extreme quintiles are populated by inspectors with the noisiest leniency scores, not the most extreme true styles.

The key diagnostic is lagged leniency. I compute each inspector's allow rate in the prior calendar year and use it to predict current-year outcomes. The coefficient is 0.225 ($p < 0.001$), confirming that inspector styles *are* persistent: an inspector who allowed more cases last year is significantly more likely to allow cases this year. The fact that lagged leniency works while

Table 2: First Stage: Inspector Leniency Predicts Appeal Outcomes

	Coefficient	N	Fixed Effects
Panel A: First Stage (Dep. Var.: Appeal Allowed)			
Leniency (cell-specific)	-0.0538	860	LPA×Type + Year×Type
	(0.0226)		
Leniency (simple FEs)	-0.0414	857	LPA + Type + Year
	(0.0194)		
First-stage F	5.6 / 4.6		
Panel B: Balance Tests (Dep. Var. in Left Column)			
Filing lag (days)	-1.7523	857	
	(3.0783)		
Householder type (=1)	-0.0103	857	
	(0.0205)		

Notes: Panel A reports the first stage of the inspector leniency IV. The dependent variable is an indicator for the appeal being allowed. Leniency is the leave-one-out inspector allow rate within case-type×year cells, standardized to mean zero and unit variance. Panel B reports balance tests: leniency should not predict case characteristics within cells. Standard errors clustered at LPA level in parentheses.

contemporaneous leave-one-out fails is precisely what the small-sample bias story predicts. Lagged leniency aggregates over a full year of cases (not two or three), producing more stable estimates of latent strictness.

5.3 Robustness

Table 4 reports four robustness checks. First, using overall (non-cell-specific) leave-one-out leniency produces a qualitatively similar negative coefficient (-0.073) but with a large standard error (0.095), consistent with the noise interpretation. Second, excluding same-LPA cases from the leniency calculation yields -0.047 (SE: 0.092), ruling out the hypothesis that within-LPA case correlation drives the result. Third, the planning-appeal subsample ($N = 569$) shows a stronger negative coefficient (-0.058, $p = 0.018$) than the householder subsample (-0.042, $p = 0.46$). Fourth, the leave-one-inspector-out range is tight: $[-0.061, -0.049]$.

6. Conclusion

This paper attempts an examiner leniency IV in a promising institutional setting—England's Planning Inspectorate—and documents what happens when the design meets insufficient data. The negative first stage is not evidence against quasi-random assignment (balance tests pass) or against persistent inspector styles (lagged leniency is strongly positive). It is

Table 3: Monotonicity and Lagged Leniency

	Allow Rate	N
Panel A: Allow Rate by Leniency Quintile		
Quintile 1 (strictest)	0.325	194
Quintile 2	0.282	181
Quintile 3	0.324	142
Quintile 4	0.315	178
Quintile 5 (most lenient)	0.267	165
Panel B: Lagged Leniency (Prior-Year Allow Rate)		
	Coefficient	SE
Lagged leniency → current outcome	0.225	(0.063)
p-value	< 0.001	

Notes: Panel A reports the mean allow rate within each quintile of the contemporaneous leave-one-out leniency score. Under a valid instrument, allow rates should increase monotonically across quintiles; the flat/declining pattern is consistent with small-sample noise. Panel B reports the coefficient from regressing the current appeal outcome on the same inspector's allow rate in the prior calendar year, with LPA and year fixed effects. Standard errors clustered at LPA level.

evidence that leave-one-out scores based on two or three cases per cell are mechanically noisy, producing a sign reversal through mean reversion.

The result has three implications. First, it underscores the data requirements for credible examiner designs: the leave-one-out score must be based on enough cases to approximate the examiner's true propensity, not just the realized outcomes of a handful of other cases. Rules of thumb from the literature suggest 30–50 cases per examiner as a minimum (Frandsen et al., 2023); the planning inspector setting falls far short with a median of 2.2.

Second, the PINS setting *is* promising for future work. The full case archive contains over 100,000 decided appeals across roughly 300 inspectors—enough for stable leniency estimation if the complete population is assembled through comprehensive scraping rather than the sample approach used here. The automated PDF extraction pipeline developed in this paper achieves 79% success rates and can be scaled.

Third, the lagged leniency result confirms that inspector heterogeneity is real and economically meaningful. Inspectors' prior-year allow rates predict current outcomes with a coefficient of 0.225, implying substantial variation in strictness. Whether this variation translates into housing supply differences remains an open question that the full-sample design can answer.

For housing policy, the existence of persistent inspector heterogeneity means that the "inspector lottery" is real: a developer's chances depend partly on which inspector is assigned, not only on the merits of the proposal. The Planning and Infrastructure Bill's proposal to expand inspector capacity should consider not only the number of inspectors but the

Table 4: Robustness of the First Stage

Specification	Coefficient	SE	N	F-stat
Baseline (cell-specific LOO)	-0.0538	0.0226	860	5.6
Overall LOO (no cells)	-0.0730	0.0950	859	
Excluding same-LPA cases	-0.0468	0.0921	859	
Leave-one-inspector-out range	[-0.0607, -0.0488]			

Notes: Each row reports the first-stage coefficient from a regression of appeal outcome (allowed=1) on inspector leniency with LPA×case-type and year×case-type fixed effects, except where noted. "Excluding same-LPA cases" computes each inspector's leniency from all cases outside the current LPA. Leave-one-inspector-out range drops each of the 20 most prolific inspectors in turn. Standard errors clustered at LPA level.

consistency of their decisions.

References

Borusyak, Kirill and Peter Hull, "Non-Random Exposure to Exogenous Shocks," *Econometrica*, 2023, *91* (6), 2155–2195.

Department for Levelling Up, Housing and Communities, "Planning Applications in England: October to December 2023," Technical Report, DLUHC 2024.

Dobbie, Will, Jacob Goldin, and Crystal S. Yang, "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 2018, *108* (2), 201–240.

Frandsen, Brigham R., Lars J. Lefgren, and Emily C. Leslie, "Judging Judge Fixed Effects," *American Economic Review*, 2023, *113* (1), 253–277.

Hilber, Christian A. L. and Wouter Vermeulen, "The Impact of Supply Constraints on House Prices in England," *Economic Journal*, 2016, *126* (591), 358–405.

Kling, Jeffrey R., "Incarceration Length, Employment, and Earnings," *American Economic Review*, 2006, *96* (3), 863–876.

Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand, "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt," *American Economic Review*, 2013, *103* (5), 1797–1829.

Sampat, Bhaven and Heidi L. Williams, "How Do Grant Decisions Affect the Diffusion of Ideas? Evidence from Examiner Leniency," *American Economic Review*, 2019, *109* (5), 1848–1887.

Table 5: Standardized Effect Sizes

Outcome	$\hat{\beta}$	SE	SD(Y)	SDE	SE(SDE)	Classification
Panel A: Pooled						
Appeal allowed (first stage)	-0.0538	0.0226	0.460	-0.1170	0.0492	Moderate negative
Panel B: Heterogeneous (by appeal type)						
Allowed: Householder appeals	-0.0417	0.0563	0.487	-0.0857	0.1156	Moderate negative
Allowed: Planning appeals	-0.0582	0.0242	0.445	-0.1308	0.0543	Moderate negative

- *Notes:* **Country:** United Kingdom (England). **Research question:** Does quasi-random assignment to more lenient planning inspectors increase the probability of appeal success and subsequent housing construction in England? **Policy mechanism:** England's Planning Inspectorate assigns professional inspectors from a national pool to adjudicate planning appeals; inspectors bring systematically different professional judgments, creating stable variation in strictness that is orthogonal to case characteristics within local planning authority and case-type cells. **Outcome definition:** Panel A pooled outcome is a binary indicator for whether the planning appeal was allowed (1) or dismissed (0); Panel B splits by householder vs. full planning appeals. **Treatment:** Binary (allowed vs. dismissed), instrumented by leave-one-out inspector leniency score. **Data:** UK Planning Inspectorate Appeal Case Portal, case-level records with inspector identities extracted from decision letter PDFs, covering England 2019–2023. **Method:** Inspector leniency IV following Dobbie, Goldin, and Yang (2018); leave-one-out leniency constructed within case-type-by-year cells; LPA-clustered standard errors; balance tests confirm quasi-random assignment. **Sample:** Planning appeals decided by PINS inspectors in England; restricted to inspectors with at least 3 observed decisions for leniency estimation. SDE $= \hat{\beta}/\text{SD}(Y)$ where SD(Y) is the unconditional standard deviation of the outcome. Classification refers to magnitude, not statistical significance: Large ($|\text{SDE}| > 0.15$), Moderate (0.05–0.15), Small (0.005–0.05), Null (< 0.005).

Appendix: Standardized Effect Sizes

Acknowledgements

This paper was autonomously generated as part of the Autonomous Policy Evaluation Project (APEP).

Contributors: @olafdrw

First Contributor: https://github.com/olafdrw

Project Repository: https://github.com/SocialCatalystLab/ape-papers