Leniency Designs: An Operator's Manual*

Paul Goldsmith-Pinkham, Peter Hull, and Michal Kolesár†

November 5, 2025

High-stakes treatments are often determined by expert decision-makers: doctors decide whether to treat patients, bail judges decide whether to release defendants before trial, patent examiners decide whether to grant patents to firms, child welfare investigators decide whether to place children in foster homes, and loan officers decide whether to approve consumer loan applications. Usually, of course, decision-makers disagree on the appropriate course of action; some are more *lenient* than others, in that they tend to grant the treatment more often. Furthermore, which decision-maker is assigned to a given case is often either deliberately random (to ensure fairness), or, due to idiosyncrasies in the assignment process (such as rotations in shifts), as-good-as-random. A *leniency design* (also known as a judge or examiner instrument design) harnesses such exogenous variation to estimate the causal effects of the high-stakes treatment by using a measure of the decision-makers' leniencies as an instrument for the treatment in an instrumental variables regression. Such designs have exploded in popularity in recent years.[1]

Leniency designs rest on a straightforward idea: many cases are close calls where a lenient decision-maker would say yes, but a stricter one would say no. If decision-makers are randomly assigned, assigned leniency is also as-good-as-random. And if assignment only affects outcomes through the treatment decisions, leniency is a valid instrument. But when implementing this logic, several practical questions arise. How exactly should we measure decision-maker leniency? What controls do we need to include in the analysis? Do we need to cluster standard errors, and if so, at what level? Does the fact that leniency is estimated matter for inference? And if we think treatment effects are heterogeneous, what are our estimates even capturing?

This paper develops a step-by-step guide to leniency designs, drawing on recent econometric literatures. The starting point of this operator's manual is a link between leniency designs and more conventional instrumental variable regressions, in which indicators for decision-maker assignment are used directly as instruments. This link follows from observing that, with no controls in this more conventional specification, the population first-stage regression of treatment on the assignment indicators gives the ideal leniency measure. Hence, the more conventional specification can be understood as also using a leniency measure as the

*We thank Joan Farre-Mensa and Alexander Ljungqvist for comments. Aurel Rochell provided excellent research assistance.

†Goldsmith-Pinkham: Yale University, paul.goldsmith-pinkham@yale.edu. Hull: Brown University, peter_hull@brown.edu. Kolesár: Princeton University, mkolesar@princeton.edu.

[1]Prominent applications include Dobbie and Song (2015), Dobbie et al. (2018, 2021), Farre-Mensa et al. (2020), Autor and Houseman (2010), Doyle (2007), and Mullainathan and Obermeyer (2022). See Table 1 of Frandsen et al. (2023) for more.

instrument. Complementing a recent review by Chyn et al. (2025), we use this link to to develop answers to the above questions, via a robust way of estimating the more conventional instrumental variable regression.

In particular, we show how the unbiased jackknife instrumental variables estimator (UJIVE) of Kolesár (2013) is purpose-built for leveraging exogenous leniency variation while avoiding subtle biases from other leniency constructions with many decision-makers or controls. We further explain how the local average treatment effect framework of Imbens and Angrist (1994) can be used to interpret UJIVE estimates under arbitrary treatment effect heterogeneity. As it turns out, UJIVE is not only useful for estimating treatment effects in this framework: building on Abadie (2002) and Kitagawa (2015), we show how it can also be used to test a key identifying assumption (*average monotonicity*). Finally, we show how knowledge of the decision-maker assignment process—or "design"—can guide standard error calculations, and that the heteroskedastic-robust (i.e., non-clustered) standard errors calculated by UJIVE are often appropriate.

We conclude with a practical checklist for applying examiner designs, assessing key assumptions, and probing external validity, all with a unified UJIVE estimation approach. We illustrate this checklist with a re-analysis of Farre-Mensa et al. (2020), who use quasi-random patent examiner assignment to estimate the value of patents to startups. Patent-granting significantly increases future patent applications, approvals, and citations on both the extensive and intensive margin, though the magnitude and precision of these estimates are sensitive to using the more robust UJIVE estimator. The key average monotonicity condition appears to hold, and the complier subpopulation which contributes treatment effects to the UJIVE estimates appears representative of the broader study population.

Motivating Example: Estimating the Value of Patents

We start by considering the simplest possible version of a leniency design, with two decision-makers who are completely randomly assigned. Concretely, consider a stylized version of the setting in Farre-Mensa et al. (2020), who again estimate the effect of patent-granting to startups on the firms' later inventiveness. We observe a set of n applications, indexed by i, which are submitted by firms to the US Patent Office. The applications are then assigned by a random coin flip to one of two examiners, s and t. The examiners decide whether to grant the patent, as indicated by the dummy variable $x_i \in \{0, 1\}$. We refer to x_i as *treatment*, following standard causal inference lingo. An *outcome* y_i is then realized; for concreteness, let's suppose y_i measures the number of future patent applications filed by the firm.

We want to leverage the randomness in examiner assignment to estimate the causal effect of x_i on y_i. For now, we assume this effect is the same across all firms: being granted a patent causes each firm to submit β additional patent applications, compared to being denied a patent. In practice, it is likely that the effects of patent-granting are heterogeneous (i.e., different for different firms); we allow for this possibility later.

Ignoring treatment effect heterogeneity for now, we write an *outcome equation* for y_i as

$$y_i = \gamma + \beta x_i + \varepsilon_i, \tag{1}$$

where the intercept γ represents the average number of additional patent applications submitted when a firm is denied a patent. The actual number of additional patent applications varies across firms due to various unobservable factors captured by the error term ε_i.

The basic identification challenge is that estimating Equation (1) by simple ordinary least squares need not reveal β because of *selection bias*: startups with high ε_i who would produce many patents in the future even if their current application were denied (because, say, they have higher research and development budgets or employ better scientists) may submit stronger applications that are more likely to be granted. This results in a positive correlation between the error term and the treatment x_i, which, in turn, induces a positive selection bias: a least-squares regression of y_i on x_i will tend to overstate the true causal effect β.

A leniency design addresses selection bias by leveraging differences in the randomly-assigned examiners' tendency to grant patents, isolating variation in the treatment that is unrelated to ε_i. To formalize this, suppose that examiner t is "tough", while examiner s is "soft" in that the overall patent-granting rate of examiner s in the population of patent applications (which we refer to as their *leniency*) exceeds that of examiner t: $p_s > p_t$. Let $\hat{p}_s$ and $\hat{p}_t$ be in-sample estimates of these rates, corresponding to the fraction of patents we observe each examiner granting in the data. Further, let z_i denote the dummy for being assigned to the soft examiner, so it equals 1 if application i is handled by examiner s and 0 if it is handled by examiner t. Then the estimated leniency of the examiner assigned to i is given by $\hat{\ell}_i = \hat{p}_t + (\hat{p}_s - \hat{p}_t)z_i$. A leniency design uses this estimated leniency as an *instrument* for patent-granting x_i. Specifically, we estimate β with an instrumental variable (IV) regression, instrumenting x_i with $\hat{\ell}_i$ in the outcome equation.

To understand why this IV regression works, note that here—with only two randomly assigned examiners—it is equivalent to a simpler IV regression which uses the examiner assignment dummy z_i directly as an instrument. In other words, it does not matter whether we use the estimated leniency $\hat{\ell}_i$ as an instrument or the dummy z_i: the resulting estimates of β are numerically identical. This is because the estimated leniency is just a version of z_i that is scaled by $\hat{p}_s - \hat{p}_t$ and shifted by $\hat{p}_t$, and such instrument scaling and shifting leaves IV estimates unchanged.

Note further that z_i is randomly assigned, like treatment assignments in a simple randomized controlled trial. Thus, so long as examiner assignment only affects the outcome y_i through the patent-granting treatment x_i—the usual IV *exclusion restriction*—z_i will be independent of ε_i and therefore be a valid instrument for estimating β.[2] Exclusion seems reasonable to assume here, unless we thought examiners did other things besides ruling on the patent (such as giving advice to the firms). Furthermore, because z_i is randomly

[2]Consistency of these IV estimates also requires a relevance condition: that z_i and x_i have non-zero correlation. Here relevance holds because we assumed that the examiners vary in their leniency, $p_s > p_t$. This means that the population first-stage coefficient from regressing x_i on z_i, given by $p_s - p_t$, is non-zero.

assigned by a coin flip for each application, it is uncorrelated across applications. It follows that conventional heteroskedasticity-robust standard errors are appropriate for quantifying uncertainty in these IV estimates. There is no need to cluster the standard errors, analogous to how we don't need to cluster them in simple randomized controlled trials. By the numerical equivalence between the leniency version of the IV and the dummy version of it, the same is true when we use the estimated leniency as an instrument.

Real-world leniency designs are naturally more complicated than this stylized example, with more than two (and often many) decision-makers who are not assigned completely at random. Handling these features requires modifications to the above simple IV approach, as we next discuss. Still, the core equivalence between using estimated leniency versus examiner dummies directly as instruments will also prove a useful guide to thinking through estimation and inference in these more complex leniency designs.

Estimation with Many Decision-Makers and Controls

Decision-maker assignment is rarely completely random, but institutional knowledge can sometimes imply it is as-good-as-random once we condition on an appropriate set of control variables. We refer to these as *necessary controls*, as they must be included in every specification leveraging exogenous assignment. As we discuss more below, patent applications in the US are first classified depending on the technology being patented. This determines which group of specialist examiners—the so-called "art unit"—will review the application. Assignment within art units is then effectively random, conditional on the set of examiners working in the art unit at the time of the assignment. Assuming this set changes slowly over time, a researcher may take art unit-by-year fixed effects as the set of necessary controls.[3]

With many examiners, assignment is now captured by a vector z_i of K instruments: one for each examiner, with one examiner omitted in each art unit to prevent multicollinearity. This leads to a *first-stage equation*, a population regression specification linking the treatment, instrument, and controls:

$$x_i = z_i'\pi + w_i'\delta + \nu_i, \tag{2}$$

where the vector w_i comprises art unit-by-year dummies and ν_i is a regression residual. In the special case where we just have one year of data, so that w_i reduces to art unit fixed effects, the first-stage regression coefficients have a simple interpretation: δ_j gives the overall leniency of the omitted reference examiner in art unit j, while π_k measures the difference between the leniency of examiner k and the omitted reference examiner in the art unit that k works in.

If we knew the first-stage coefficients π and δ, we could use the population first-stage fitted values $\ell_i = z_i'\pi + w_i'\delta$ as our leniency measure. By the Frisch-Waugh-Lovell theorem, an IV regression using

[3]One may optionally include *precision controls* in some specifications: pre-assignment characteristics of the firm that help explain variation in the outcome. Including these variables will soak up some variability in the error ε_i and thus help reduce the standard errors. This is the IV analog of including baseline characteristics as controls in regression specifications estimating effects of treatments in randomized controlled trials, while including the necessary controls is analogous to controlling for strata indicators in randomized trials where the treatment assignment probability varies across strata.

this instrument while controlling for w_i is equivalent to running an IV regression without any controls, but instrumenting with the residual from projecting the fitted values onto the covariates, $\tilde{\ell}_i = \tilde{z}_i'\pi$. Here $\tilde{z}_i$ denotes the sample residual from regressing the instrument vector z_i onto the controls w_i. This IV regression leads to the estimator:

$$\hat{\beta}^* = \frac{\sum_i y_i \tilde{\ell}_i}{\sum_i x_i \tilde{\ell}_i}, \tag{3}$$

the ratio of sample covariances between *relative leniency* $\tilde{\ell}_i$ and the outcome y_i versus the treatment x_i. We use the term relative leniency to stress that $\tilde{\ell}_i$ measures the leniency of examiner handling application i *relative* to other examiners who could have handled the application. In contrast, ℓ_i is a measure of *absolute leniency*. With just one year of data, these leniency measures take a simple form: ℓ_i is the overall patent-granting propensity of the examiner assigned to i, while $\tilde{\ell}_i$ is the examiner's patent-granting rate minus the overall patent-granting rate of the art unit handling application i.

Since examiner assignment is as-good-as-random given the controls w_i, and since we are netting out the differences across the controls by using the residual variation in examiner assignment $\tilde{z}_i$, relative leniency is itself as-good-as-randomly assigned. Provided the exclusion restriction holds, $\tilde{\ell}_i$ will therefore be uncorrelated with the outcome error ε_i as in the simple motivating example. By definition of a regression residual, $\tilde{\ell}_i$ is also uncorrelated with the residual ν_i in the first-stage equation. These two facts imply that the expectation of the denominator in eq. (3) equals $\sum_i \tilde{\ell}_i^2$, while the expectation of the numerator equals $\beta \times \sum_i \tilde{\ell}_i^2$.[4] Thus, approximating the expectation of a ratio by a ratio of expectations, we conclude that using the relative leniency instrument leads to an approximately unbiased estimate of β.

In practice, it is infeasible to use the true relative leniency as an instrument because we do not know the first-stage coefficients. A tempting alternative is to simply use least squares estimates of the first-stage coefficients, $\hat{\pi}$ and $\hat{\delta}$, in place of the unknown population values. This is exactly equivalent to using a two-stage least squares (2SLS) estimator which instruments with the full set of assignment dummies z_i. To see this, recall that the 2SLS estimator is equivalent to an instrumental variables estimator that uses a single instrument given by the sample first-stage fitted values, $z_i'\hat{\pi} + w_i'\hat{\delta}$. By the Frisch–Waugh-Lovell theorem, this is in turn equivalent to an IV regression without covariates that replaces the population relative leniency $\tilde{\ell}_i$ in eq. (3) with the sample analog $\hat{\ell}_i = \tilde{z}_i'\hat{\pi}$ (the residual from projecting the fitted values onto the controls).

In contrast to using the true relative leniency $\tilde{\ell}_i$, however, using the estimated relative leniency $\hat{\ell}_i$ will tend to produce biased treatment effect estimates when there are many examiners. This is the classic many-weak instrument bias problem for 2SLS (e.g., Bekker, 1994; Bound et al., 1995); it arises here because the relative leniency measure $\hat{\ell}_i$ is constructed in part from the treatment status of observation i through the least squares estimate $\hat{\pi}$. With just one year of data, for instance, $\hat{\ell}_i$ is given by the fraction of patents granted by the examiner assigned to i in the sample, minus the fraction of patents granted by the art unit handling

[4] Here, and below, following the econometric literature, we implicitly condition on the observed instruments and covariates z_i and w_i such that randomness only comes from unobserved errors ε_i and ν_i. Appendix A details this setup and gives formal derivations of all claims in this section.

the application—and both fractions are computed including the data from application i. But applicant i's treatment likely correlates with the outcome error ε_i; this is, after all, the reason to use IV instead of simple least squares regression. Thus, the estimated leniency instrument is also likely correlated with the outcome error, and this generates bias in IV estimates in the direction of the naïve ordinary least squares estimates.

The bias from using the estimated leniency instrument can be severe, especially when examiner assignment only modestly affects the probability of treatment. A helpful rule of thumb for gauging the magnitude of the bias obtains when we assume the errors are homoskedastic; then, the 2SLS bias approximately equals the bias of ordinary least squares times $1/\big((1-R^2)\times E[F]\big)$, where $E[F]$ is expectation of the first-stage F statistic for the hypothesis that the first-stage instruments are irrelevant (i.e., that $\pi = 0$ in eq. (2)), and R^2 is the population partial R-squared from adding instruments to the first-stage regression. In practice, since R^2 tends to be near zero, the magnitude of $E[F]$ tells us how much smaller 2SLS bias is relative to ordinary least squares (it's never larger, since $E[F]$ always exceeds one). In fact, this relationship can be used to justify the popular $F > 10$ rule of thumb proposed by Staiger and Stock (1997) for identifying weak instruments. If the expectation of the first-stage F statistic lies below 10, then 2SLS bias exceeds 10% of least squares bias, so the rule of thumb can be thought of as a diagnostic for whether we are in this large bias region.[5] With many examiners, the first-stage F statistic can be modest even when examiner leniency explains economically meaningful variation in the treatment, because the formula for F divides by K.

Knowing that the 2SLS bias comes from using own treatment status to estimate the relative leniency suggests a natural bias-free alternative: instrument with a leniency estimate $\hat{\ell}_{-i}$ that *leaves out* observation i's own value of the treatment, thereby avoiding a mechanical correlation with ε_i. The unbiased jackknife instrumental variable estimator (UJIVE), studied in Kolesár (2013), implements this logic by setting the relative leniency estimate to $\hat{\ell}_{-i} = \tilde{z}_i'\hat{\pi}_{-i}$ where $\tilde{z}_i$ are the same instrument residuals as before and $\hat{\pi}_{-i}$ is a least-squares estimate of π which uses all observations except for i. Because this leave-one-out—or "jackknifed"—estimate is unbiased for π and, in *iid* data, independent of the data for observation i, the same arguments used to show approximate unbiasedness of the infeasible estimator $\hat{\beta}^*$ can also be used to show approximate unbiasedness of the UJIVE estimator.

The unbiasedness property of UJIVE hinges on the fact that it uses leave-out estimation to directly estimate relative leniency $\tilde{\ell}_i$ accounting for controls through the residualized $\tilde{z}_i$. Leave-out estimation of absolute leniency ℓ_i, in contrast, tends to work poorly when the number of covariates is large. In particular, Phillips and Hale (1977) and Angrist et al. (1999) propose an estimator known as JIVE (the jackknife IV estimator) that constructs a leave-out estimate of the absolute leniency and uses it as an instrument in a regression with covariates. By the Frisch-Waugh-Lovell theorem, this is equivalent to 2SLS without controls and instrumenting with least squares residuals from the regression of the leave-out absolute leniency estimate on the controls. Concretely, in a case with one year of data, the residuals of the leave-out absolute leniency

[5] Stock and Yogo (2005) construct a formal statistical test based on F for the null hypothesis that the 2SLS bias is large in the sense of potentially exceeding 10% of least squares bias, under homoskedastic errors. While the precise critical value depends on K, it is close to 10 for most values of K: see their Table 5.1.

are created by subtracting off the average leave-out-leniency estimates for observations in the art unit that handles application i—but this is just the (non-leave-out) average patent-granting rate of the art unit, which depends on the treatment status of application i. The subtraction thus reintroduces the own-observation bias of 2SLS, except that it typically runs in the opposite direction to 2SLS (since the own-observation component is now only the term being subtracted off). Under homoskedasticity, the JIVE bias approximately equals that of 2SLS multiplied by $-E[F] \times L/((E[F]-1)K - L)$ where L is the number of controls. The JIVE bias is negligible when the number of controls is much smaller than the number of instruments (in fact, UJIVE and JIVE coincide in the absence of controls and a constant). But when many controls (e.g., art unit-by-year fixed effects) are needed to extract the randomness in many examiner assignment indicators, this formula shows that the magnitude of JIVE bias can be comparable to the bias of 2SLS.

While UJIVE is a natural solution to the 2SLS bias problem, there are other reasonable approaches. Ackerberg and Devereux (2009) propose a clever modification of JIVE—the improved jackknife IV estimator (IJIVE)—which can greatly reduce its bias in the presence of controls by reversing the order of operations: *first* residualize the instruments and *then* compute a leniency measure based on leave-out fitted values. While reversing the order of operations doesn't entirely eliminate the own-observation bias, in practice IJIVE and UJIVE tend to produce similar estimates. More recently, Chao et al. (2023) propose an alternative to UJIVE, termed FEJIV. We show in the appendix that FEJIV can be interpreted as solving for a relative leniency measure with the smallest mean squared error under homoskedasticity, subject the constraints that the resulting estimator is free of own-observation bias and that the leniency measure is orthogonal to the covariates. The UJIVE leniency measure, in contrast, achieves the minimal mean squared error without the latter orthogonality constraint. The orthogonality property of FEJIV is attactive in that adding a linear function of covariates to the outcome does not affect the estimate; the price for this is that FEJIV leniency is slightly noisier. The resulting estimator also tends to be computationally demanding in large datasets, and imposes stronger data requirements (e.g., the estimator may not exist if there are high-leverage observations).

Another approach is to bias-correct the 2SLS formula, which replaces the infeasible instrument $\tilde{\ell}_i$ in (3) with the estimated relative leniency $\hat{\ell}_i$. When 2SLS does this, it increases the expectation of the denominator from $\sum_i \tilde{\ell}_i^2$ to $\sum_i \tilde{\ell}_i^2 + K \operatorname{var}(\eta_i)$ under homoskedasticity. The quantity $\sum_i \tilde{\ell}_i^2$ corresponds to the numerator of the population partial R-squared statistic; it measures the increase in the explained sum of squares from adding instruments to the first stage. Since 2SLS uses in-sample fitted values to estimate the predictive power of the instruments, it overstates this predictive power, exactly analogous to how the unadjusted R-squared overstates the predictive power of regression (the same issue happens in the numerator, and taking the ratio gives the rule-of-thumb bias formula for 2SLS). Paralleling how the adjusted R-squared fixes the issue by using a degrees of freedom correction, we can use degrees of freedom adjustments in both the denominator and the numerator of the 2SLS formula. Unfortunately, similar to the adjusted R-squared formula, the resulting bias-corrected 2SLS estimator (dating back to Nagar (1959)) works only under homoskedasticity.

In practice, rather than using UJIVE or its cousins that compute the leniency measures internally, it is

common for researchers to first construct an "external" leniency measure and then use it as an instrument in a just-identified IV regression. This is analogous to first computing first-stage fitted values and then using them as a single instrument, rather than directly using the standard 2SLS estimator which estimates everything in one step. Such "manual 2SLS" procedures are widely advised against (see, e.g., Angrist & Pischke, 2009, Chapter 4.2) and we would similarly advise against manual leniency IV estimation. Indeed, as shown above, subtle variations in how the leniency measure is exactly constructed and accounts for covariates can have potentially large consequences for the bias of the ultimate estimator. It is thus simpler and safer to use a one-step implementation like UJIVE, with established unbiasedness properties.

Two more comments on estimation are warranted before proceeding. First, it is common practice to report first-stage F statistic with 2SLS estimates; as we have discussed, its expectation, $E[F]$, directly informs about 2SLS bias. But small first-stage F statistics need not worry a researcher using UJIVE: our arguments for its approximate unbiasedness work even if $E[F]$ is small. In fact, UJIVE remains approximately unbiased and consistent even when the instruments are weak enough that $E[F]$ converges to zero in large samples, so long as $\sqrt{K} \times E[F]$ is large (see, e.g., Mikusheva & Sun, 2022). Second, the leave-out estimation approach leveraged by UJIVE assumes independence across observations. When observations are instead clustered together—a scenario we discuss more when discussing the calculation of standard errors—a leave-own-cluster-out modification of UJIVE may be needed to ensure unbiasedness (Frandsen et al., 2025). Intuitively, correlations between treatment x_i and errors ε_j for observations i and j in the same cluster will generally reintroduce the mechanical bias between the leave-own-observation-out UJIVE instrument and the errors, again tending to bias estimates towards ordinary least squares. Hence, for examiner designs, clustering in the data is not just a consideration for inference but may also guide the choice of estimator.

Heterogeneous Treatment Effects

So far, we have assumed the parameter of interest is a constant treatment effect β. This is of course a strong assumption in practice. It means, for example, that any successful patent application raises the future innovativeness of all startups i by the same constant amount. In reality, the value of patents is likely higher for some startups than for others. Luckily, the UJIVE estimator can retain a causal interpretation even in the presence of such heterogeneous effects—i.e., when patent effects β_i vary arbitrarily across firms i.[6]

The Nobel Prize–winning local average treatment effect theorem of Imbens and Angrist (1994) offers a guide to the general heterogeneous-effect interpretation of leniency designs. The theorem maintains our earlier assumption of as-good-as-random decision-maker assignment and the IV exclusion restriction, while replacing the restriction of constant treatment effects with an assumption of first-stage *monotonicity*. In the patent setting, monotonicity means that if we can find a case that some examiner a would grant a patent

[6]In this section only, we assume the treatment is binary as in much of the literature on heterogeneous treatment effects. This is mostly for notational convenience, since otherwise effect heterogeneity could come both from differences across observations and from different margins of treatment response for a given observation. See Kolesár and Plagborg-Møller (2025, Appendix A.2) for an extension of the local average treatment effect theorem to the case with multivalued or continuous treatment.

to but some other examiner b would deny, it must be that a is more lenient in *all* cases; i.e., there cannot be another case that a would deny but b would grant. In the simple example with just two examiners, this implies we can divide the universe of patents into two groups: a *complier* group of firms with patent applications that are granted when assigned to the soft examiner s but not when assigned to the tough examiner t, and a non-responder group whose treatment status is unaffected by examiner assignment (they are always either granted or denied, regardless of which examiner handles the case). Monotonicity rules out the presence of a third group of *defier* firms that are granted their applications only if assigned to t. The local average treatment effect theorem states that, in the absence of defiers, the IV regression using an indicator variable for assignment to the soft examiner as an instrument estimates the average treatment effect for compliers. Imbens and Angrist (1994) call this a local average treatment effect, to stress that we learn nothing about treatment effects for non-responders. In contrast, if we ran a randomized controlled trial that granted applications by a coin flip, we would learn the overall average treatment effect $E[\beta_i]$.

With many as-good-as-randomly assigned decision-makers, monotonicity implies there are many complier groups: one for each pair of decision-makers. In this case, an extension of the theorem shows that using the relative leniency $\tilde{\ell}_i$ as an instrument—or an approximately unbiased feasible version such as UJIVE—identifies a weighted average of complier-group specific treatment effects, weighted using the squared leniency differences of each decision-maker pair.[7] This is so because, in the population, leniency IV is equivalent to running a series of simple IV regressions with a single binary instrument that restricts the sample to a given decision-maker pair and then averaging these IV regressions together using squared pairwise leniency differences as weights. Since a leniency difference measures the share of compliers, leniency IV thus weights by the square of the group size. As a result, larger complier groups receive more weight. But importantly, the weighting is convex: i.e., there are no complier groups given negative weight.

While it is likely more palatable than assuming constant treatment effects, monotonicity is nevertheless a strong assumption. In particular, results in Vytlacil (2002) imply it is equivalent to assuming each patent examiner has the same ranking of the appropriateness for granting patents; the examiners only differ in the cutoff they use for gauging whether the application is above the bar. The strength of this assumption for leniency designs was, in fact, first noted in the original Imbens and Angrist (1994) analysis (see their Example 2, p. 472). There are two core issues. First, in many settings like our patent example, there are multiple dimensions to decision-makers' evaluation criteria which may make it unlikely that any two decision-makers would have the same ranking. For instance, patent applications are evaluated by their perceived usefulness, novelty, and non-obviousness, and different examiners likely place different weights on these criteria. Second, even if the weights were the same, variation in skill among decision-makers can induce ranking differences due to mistakes by less-skilled decision-makers. For instance, Chan et al. (2022) show that in the case of radiologists variation in skill accounts for about 40% of variation in decision-maker leniency.

Luckily, first-stage monotonicity can be both weakened and tested. Without monotonicity, a simple IV

[7] The appendix states the formal result. A general statement appears, for example, in Kolesár (2013), who builds on Imbens and Angrist (1994) by allowing for linear controls w_i.

regression restricted to a pair of decision-makers identifies a weighted *difference* between complier and defier treatment effects with weights given by the fraction of compliers and defiers (because defier treatments are shifted in the opposite direction by decision-maker assignment, relative to the compliers). Since leniency IV averages these pairwise IV regressions, it can be seen as identifying a weighted-average treatment effect for compliers and defiers, but with negative weight put on all defier groups. This is a problem for its causal interpretation, in general. For example, we risk "sign-reversals": even if patent-granting, say, uniformly increases future firm productivity so that β_i is positive for all i, the leniency IV could estimate a negative effect—namely, if the treatment effect for defiers is much larger than that for compliers.

This motivation for the monotonicity assumption suggests that it can be weakened in three ways. First, while the presence of defiers necessarily causes a negative weighting problem in the case with a single pair of examiners, there is a subtlety with multiple examiners. Since each firm appears in multiple pairwise examiner comparisons, the total weight we place on each firm corresponds to its average complier, or defier, status across all pairwise comparisons. Thus, as long as no firm is a defier "on average" the weights placed on individual firms remain positive. This is equivalent to assuming that the average leniency of the examiners who would grant the firm a patent exceeds the average leniency of those who would deny it, a condition Frandsen et al. (2023) term "average monotonicity."[8]

Second, even if some weights are negative, it only presents a problem when the patent applications with negative weights systematically differ in their treatment effects; one may be willing to assume this is not the case.[9] More generally, de Chaisemartin (2017) points out that if one finds a subset of positively weighted applications that matches the treatment effects of the firms that are defiers on average, these compliers cancel out the negative treatment weights on the defiers, and we can interpret leniency IV as estimating a convex-weighted average of treatment effects for the remaining compliers.

Third, even if the negatively weighted observations differ systematically, this need not bias leniency IV estimates substantively unless there are many firms with non-negligibly negative weights and the systematic differences in treatment effects are very large. We never see two patent examiners evaluate the same case, so we cannot directly assess the common ranking assumption. But we do actually have some evidence on this in the context of judge assignment, because some court cases are assigned to a panel of judges: Sigstad (2025) shows that while monotonicity is technically often violated in judicial panels, the ranking disagreements are not severe enough to create substantial bias in leniency IV estimates.

Typically, leniency IV designs do not feature multiple decision-makers being assigned to the same case, precluding such direct monotonicity tests in other contexts. But we can still indirectly test the assumption. One implication of monotonicity is that the average outcomes for cases assigned to two decision-makers with the same leniency must be the same in the population, because under monotonicity, the two decision-makers' decisions must exactly match. More generally, if the leniencies are similar, the average outcomes must also be

[8]While the statement of the condition in Frandsen et al. (2023) is more complicated, we show in the appendix that our formulation is equivalent. With just two decision-makers, average monotonicity reduces to the usual monotonicity condition.

[9]Heckman et al. (2006) call this assumption no "essential" treatment effect heterogeneity. In terms of the Roy (1951) selection model, it imposes no "selection-on-gains."

similar since the number of observations on which the decision-makers disagree must be small. Frandsen et al. (2023) formalize this idea in a statistical test.[10] While useful in some settings, the test has two limitations. First, to show its validity, Frandsen et al. (2023) rule out a large number of decision-makers (i.e., the original motivation for using UJIVE), and its implementation requires bounded outcomes (ruling out, e.g., looking at effects of patent-granting on future startup profits). Further, it tests the original Imbens and Angrist (1994) monotonicity condition, not the weaker average monotonicity condition that is necessary and sufficient for positive individual weights.

To address both limitations, we propose here a test for average monotonicity—based on an earlier proposal by Kitagawa (2015)—which leverages a modification of the UJIVE treatment effect estimator.[11] We develop our proposed monotonicity test by first noting, following Abadie (2002), that the UJIVE estimator can not only estimate average treatment effects for compliers: it can also *characterize* compliers by their baseline characteristics and potential outcomes. Formally, consider a UJIVE estimator with the same treatment, instruments, and controls as before, but with a modified outcome. Rather than y_i, we put on the left-hand side $\tilde{y}_i = v_i \times x_i$, where v_i equals some variable determined before the as-good-as-random assignment of z_i. The local average treatment effect theorem applies to this modified outcome as well, showing that under average monotonicity, the UJIVE estimator identifies a convex weighted average of treatment effects of x_i on $\tilde{y}_i$. But by construction these "effects" are just v_i, since $\tilde{y}_i$ is moved by exactly this amount when x_i is counterfactually increased by one unit. Moreover, the weights that UJIVE puts on these effects are exactly the same as the ones in the original treatment effect specification. Hence, by simply replacing y_i with $\tilde{y}_i$, we can easily compute weighted averages of v_i with the UJIVE weights.

To see why this result is useful for testing monotonicity, note that it is possible to obtain a logically invalid estimate when computing such weighted averages. If v_i is a binary variable, which can only take on values zero or one, then its weighted average must not be negative or greater than one. Such a finding would suggest something has gone wrong in the local average treatment effect theorem, namely monotonicity (assuming we are confident in as-good-as-random assignment and exclusion). More generally, we can set v_i to be an indicator that some baseline variable—or set of variables—takes on a particular set of values and check that the resulting UJIVE estimate with that $\tilde{y}_i$ as the outcome lies between zero and one. For binary x_i, we can even use the original outcome y_i (by setting $\tilde{y}_i = y_i \times x_i$) since then UJIVE will estimate a weighted average of *treated outcomes* (Abadie, 2002); weighted averages of untreated outcomes can be estimated by setting $\tilde{y}_i = y_i \times (x_1 - 1)$. Of course, monotonicity tests like these are as straightforward to implement via UJIVE, and they inherit its approximate unbiasedness property even with many decision-makers or controls.[12]

[10] Strictly speaking, both this test and the average monotonicity test described below are of the joint null hypothesis of as-good-as-random assignment, exclusion, and monotonicity, as a violation of any of these assumptions could drive a test rejection. Often they are interpreted as tests of monotonicity maintaining the first two assumptions, however.

[11] Alternative informal tests have been used by researchers in the literature, such as in Dobbie and Song (2015) and Dobbie et al. (2017), where the researcher plots the average leniency of each decision-maker for one subgroup against their average leniency for the other subgroup. A positive relationship is taken as evidence that decision-makers are not switching their behavior differentially by subgroup. However, the formal properties of these procedures have not been established.

[12] An important caveat to this suggestion is that the power properties of this test have not yet been explored. With a binary instrument, setting $\tilde{y}_i$ to indicators for all possible outcome values interacted with x_i and $x_i - 1$ yields a version of the Kitagawa

We again close this section with two additional comments. First, one might wish to use the above method to estimate average baseline characteristics of compliers even when monotonicity is not a concern. This helps evaluate another more subtle concern with IV estimation given heterogeneous treatment effects: the *external validity*, or representativeness, of the estimated complier-average treatment effect. If compliers are representative of the broader study population through their observable characteristics, again simply by replacing the y_i outcome with some $\tilde{y}_i = v_i \times x_i$ for baseline v_i, this concern can be lessened.[13]

Second, a further subtle issue can arise when the specification of controls in w_i is insufficiently flexible. To ensure we estimate a convex weighted average of treatment effects free of omitted variable bias with a single instrument that is only conditionally as-good-as-randomly assigned, the covariates w_i need to enter flexibly enough so that the conditional mean of the instrument vector given w_i, $E[z_i \mid w_i]$, is linear in the covariates (Kolesár (2013) shows this condition is sufficient, while Blandhol et al. (2025) show it is necessary). When the instrument is a vector of dummies, a sufficient condition for nonnegative weights is that the mean of each element of z_i is linear in w_i and the remaining elements of z_i (Goldsmith-Pinkham et al., 2024). In our application below, this would require interacting the examiner assignment with art unit-by-year fixed effects, similar to the specification in Angrist and Krueger (1991). However, as with the potential monotonicity violations discussed above, more parsimonious control specifications (such as only using examiner dummies and art unit-by-year fixed effects without interacting them) need not generate meaningful bias, and they may in fact yield more precise estimates. In practice, researchers can explore this potential bias-variance tradeoff by checking robustness to alternative control specifications.

Inference

Leniency designs with multiple decision-makers raise subtle inference problems. To see why, we start with a benchmark: suppose we knew the relative leniency $\tilde{\ell}_i$, and hence could compute the infeasible estimator $\hat{\beta}^*$ in eq. (3). With *iid* data, the correct standard errors are given by the square root of $\sum_i \hat{\varepsilon}_i^2 \tilde{\ell}_i^2 / (\sum_i \tilde{\ell}_i^2)^2$, where $\hat{\varepsilon}_i$ is the residual from projecting $y_i - x_i \hat{\beta}^*$ on the covariates. Since the covariances in the numerator and denominator of eq. (3) are approximately normally distributed by the central limit theorem, the delta method tells us their ratio is also approximately normal with this standard error.

Standard software packages compute heteroskedasticity-robust standard errors for 2SLS using the feasible version of this formula, plugging in $\hat{\ell}_i$ for $\tilde{\ell}_i$ and the 2SLS estimate for $\hat{\beta}^*$. These standard errors are correct when treatment effects are homogeneous and many-weak instrument bias is negligible. Under these conditions, 2SLS is as efficient as the infeasible estimator $\hat{\beta}^*$.

(2015) test. With multiple instruments, our test is different from the multivalued-instrument extension of the Kitagawa (2015) test (see, e.g., Coulibaly et al., 2024), since we test average monotonicity not the stronger conventional monotonicity.

[13] Another complementary way of gauging external validity is to report how complier-average treatment effects, computed by simple IV regressions restricted to a given examiner pair, vary with the pairs' leniencies. Ordering the examiners by their leniency and reporting the estimates for pairs of neighboring examiners then gives an estimate of the marginal treatment effect curve (Heckman & Vytlacil, 2005). Intuitively, a flat curve suggests that effects for non-responders are likely similar to the UJIVE estimate. However, the properties of such procedures—or their more sophisticated versions (e.g., Mogstad et al., 2018)—have not to our knowledge been formalized in the case of many decision-makers or controls.

Unfortunately, with many instruments, the same issue that causes bias in the 2SLS estimator also causes its standard errors to be small relative to the infeasible estimator. Recall that overfitting inflates the denominator in eq. (3) from $\sum_i \tilde{\ell}_i^2$ to $\sum_i \tilde{\ell}_i^2 + K \operatorname{var}(\eta_i)$ under homoskedasticity. Since the same denominator appears in the standard error formula, overfitting shrinks standard errors too. Thus, 2SLS mimics not only the ordinary least squares estimate, but also its precision. As Bound et al. (1995) emphasize, this can mask the weak instrument problem: researchers see tight standard errors around an estimate close to ordinary least squares and may wrongly conclude the causal effect is precisely estimated with minimal selection bias.

UJIVE fixes the denominator problem by construction, but the numerator of the default heteroskedasticity-robust standard error formula still has two issues. First, with heterogeneous treatment effects, we cannot match the infeasible estimator's precision even when instruments are strong: the correct numerator picks up a second term reflecting variation in complier treatment effects across complier groups. Conveniently, Imbens and Angrist (1994) already derived the correct numerator for the standard error in their appendix (see also Lee, 2018). Second, with many examiners, a third term appears due to estimation noise in $\hat{\ell}_i$ (this is the Bekker (1994) many instrument term; we give the details in the appendix). Luckily, if we use UJIVE along with the plug-in heterogeneity-robust formula, it turns out that this also takes care of the many instrument term—we use this formula in our empirical illustration below.[14]

There are still two potential issues with the UJIVE standard error formula remaining. The first concerns the reliability of the delta method, which lets us argue that if the covariances in the numerator and denominator of eq. (3) are each approximately normal, then so is their ratio. Here the issue is a classic one, dating back to Anderson and Rubin (1949), and arises when the first stage is very weak as measured by $\sqrt{K} \times E[F]$. Suppose, for example, that there is essentially no variation in leniency across patent examiners. Then, even for the infeasible estimator $\hat{\beta}^*$, the delta method does not apply as the numerator and denominator of eq. (3) are essentially zero in expectation; ratios of mean-zero normals follow a Cauchy distribution, which has much thicker tails than a normal distribution.

A vast literature has tackled this weak instrument problem over the years, beginning with Anderson and Rubin (1949) in the case with a fixed number of instruments and most recently, several papers proposing jackknife extensions for many instruments (e.g. Mikusheva & Sun, 2022; Matsushita & Otsu, 2024; Yap, 2025). The fix turns out to be quite simple. To test that the causal effect equals a particular value β_0, we compute the residual $\hat{\varepsilon}_{i0}$ by projecting $y_i - x_i\beta_0$ onto the covariates. If the null is true, $\hat{\varepsilon}_{i0}$ should be uncorrelated with relative leniency, so we just need to check whether this correlation is significantly different from zero. This turns out to be equivalent to computing the UJIVE standard error just like before and checking whether UJIVE is significantly different from β_0, with one tweak: we use $\hat{\varepsilon}_{i0}$ instead of the UJIVE residual—this is the test proposed by Yap (2025).[15]

[14]The plug-in formula for the standard error's numerator has an own-observation bias that makes it overshoot its target, similar to the upward bias in the denominator of the 2SLS estimator. By a lucky coincidence, this overshooting more than accounts for the many instrument term making the formula valid (albeit slightly conservative). Correcting the slight upward bias involves a much more complicated leave-three-out procedure: see Anatolyev and Sølvsten (2023) and Yap (2025).

[15]The test proposed by Matsushita and Otsu (2024) uses the same idea, but doesn't use the heterogeneity-robust formula for

Recently, for the single-instrument case, Angrist and Kolesár (2024) give a more optimistic take on the weak instrument problem: they observe that, for the case with known leniency, the usual standard error only leads to overly optimistic inference if the correlation between $\sum_i \varepsilon_i \tilde{\ell}_i$ (the numerator of $\hat{\beta}^* - \beta$) and $\sum_i x_i \tilde{\ell}_i$ (the denominator) is very high; if not, the weak instruments blow up the standard error with $\hat{\varepsilon}_i$ even more than if we used the robust Yap (2025) approach that uses $\hat{\varepsilon}_{i0}$ (which in this case coincides with the Anderson-Rubin test). This correlation parameter can be thought of as a measure of endogeneity, since it corresponds to the correlation between ε_i and x_i under homoskedasticity. But severe selection biases leading to very high endogeneity are unlikely in many applications; hence, Angrist and Kolesár (2024) argue, the delta method approximation may be good enough for the usual standard errors not to be misleading. We show in the appendix that an analogous argument applies to the UJIVE estimator with many instruments and controls: the only difference is the correlation parameter uses the UJIVE leave-out leniency measure rather than $\tilde{\ell}_i$; while the correlation parameter no longer directly links to endogeneity due to the more complicated variance structure, it can be bounded by placing bounds on the treatment effects. In most applications these weak instrument issues are moot, as the variation in leniency (as measured by $\sqrt{K} \times E[F]$) will be substantial. But if variation is small and high correlation cannot be ruled out, using $\hat{\varepsilon}_{i0}$ to compute standard errors with the above weak-instrument test may be prudent.[16]

A final complication arises when the data is not *iid* but clustered. This can happen for two reasons. Either the sampling is clustered (say the patent office, instead of sharing the full data, only shares application data submitted on a randomly selected subset of dates, and the researcher would like to generalize to the full set of data), or the assignment is clustered (there is a lottery for each date, and the examiner who wins the lottery is assigned all applications filed on that date). Then, even if the relative leniency $\tilde{\ell}_i$ was known, a clustered version of the standard error formula is needed. Often, a researcher observes the full population of interest (in our application, we observe all patents in a specific time frame), so the main consideration is the assignment process.

Abadie et al. (2020, 2023) show that if the examiner assignment process is *iid*, one does not need to worry about the correlation patterns of the residual ε_i across i. What matters for the standard error is the correlation pattern of the *product*, $\tilde{\ell}_i \varepsilon_i$, which will be uncorrelated due to relative leniency being consequentially *iid*. This contrasts with traditional considerations for whether and how to cluster, which treat the patent examiner assignment (and hence relative leniency) as fixed, necessitating worries about correlation patterns in the residuals. These worries go away under the random instrument view, and one can use the same rule of thumb as in randomized experiments: cluster at the level of variation in the assignment. Furthermore, whether clustering matters for the magnitude of the standard error does not help determine

the standard error. As a result, it is not robust to treatment effect heterogeneity, and neither is the many-instrument robust version of the Anderson and Rubin (1949) test considered in Mikusheva and Sun (2022).

[16] Mikusheva and Sun (2022) develop a weak-instrument pretest, designed to check validity of delta-method based standard errors regardless of the value of the correlation parameter, from a variant of the JIVE estimator. As they discuss, using the pre-test comes with a power sacrifice: if one switches between conventional and weak-instrument robust inference procedures using the pre-test, one needs to use larger critical values than the conventional 1.96 to account for possible type I and type II errors in the pretest.

whether it is *needed*, so clustering the standard errors "just in case" may yield overly conservative inference. This reasoning also applies when leniency is unknown, but the number of examiners is fixed. While the Abadie et al. (2020, 2023) arguments have not been formally extended to the many examiner case, it seems natural to let the sampling and assignment process guide the choice of whether and how to cluster.

A Checklist for Leniency Designs

We now present a step-by-step guide to leniency design implementation, illustrated with data from Farre-Mensa et al. (2020). Their setting examines how initial patent approval by a US Patent Office examiner affects subsequent innovation by US startups. Our reanalysis sample contains 32,514 first-time patent applications filed after 2001 with final decisions by 2013.[17] Table 1 summarizes key variables. The treatment—approval of the first-time application—occurs for 65% of applicants. Following Farre-Mensa et al. (2020), we track several key outcomes: whether startups file additional patents (40% do), whether they have any subsequent patents that were approved (30% do), and counts of new applications, approvals, and citations. Besides standard covariates like patent class and claim counts, we observe venture capital funding rounds prior to application. For the subset of applications also submitted abroad, we observe a proxy of application quality: European or Japanese patent office decisions. We also construct a measure of local entrepreneurship using the number of startups in each state. As we show below, these additional observables can help gauge the plausibility of the leniency design's identifying assumptions and assess external validity.

Our checklist consists of five steps; to illustrate it, we use an original R package for UJIVE, available at https://github.com/kolesarm/ManyIV.

1. Identify necessary controls for as-good-as-random assignment, and what estimator and standard errors are appropriate given the quasi-experimental design

A credible leniency design begins with institutional knowledge that justifies quasi-random decision-maker assignment. This justification naturally identifies any required controls. In the US patent office, for example, once patent applications are allocated to art units the assignment process to individual examiners is not standardized. However, Lemley and Sampat (2012) argue, based on interviews with examiners, that the assignment process is effectively random conditional on the set of examiners working in the art unit at the patent office at the time of the assignment.[18] While the potential examiner set is not directly observable, we can assume the set of examiners working in a given art unit changes only slowly across time and specify the necessary controls as art unit-by-year fixed effects. Without these necessary controls, the analysis would

[17] Our analysis uses the full sample from the original paper as found in its replication package. The initial dataset has 34,435 observations. We drop 1 observation that is missing citation data. We then drop 1,851 observations with singleton covariates or instruments, and 69 observations with leverage of 1. This causes us to drop 378 collinear controls and 1,676 collinear IVs. See the code documentation at https://github.com/kolesarm/ManyIV for discussion on how this recursive algorithm is implemented to drop these observations and variables.

[18] Many art units use the last digit of the application serial number for assignment. Others similarly use rules that imply an effectively random assignment, such as a "first-in-first-out" rule that assigns each incoming application to the first available examiner. See also the institutional discussions in Sampat and Williams (2019) and Farre-Mensa et al. (2020).

conflate systematic differences across art units or changes over time with examiner-specific variation. From our discussion on heterogeneous effects, recall that it can be important to have sufficiently flexible controls to ensure a local average treatment effect interpretation of leniency design estimates.

Quasi-random assignment is only one piece of instrument validity in an examiner design; the second is the exclusion restriction, requiring decision-maker assignment to only affect relevant outcomes through the specified treatment. Here too, a clear argument should be made from institutional knowledge. In the patent setting example, exclusion is plausible since patent examiners make relatively narrow approval decisions and likely have no direct impact on the future innovativeness of the applying firm.

The assignment mechanism can also guide the appropriate level for clustering standard errors. Individual-level randomization in the patent setting, where patents are routed idiosyncratically to examiners within art unit-year, makes heteroskedasticity-robust standard errors appropriate. Based on our rule of thumb, clustering is not needed because each case is assigned independently. Contrast this with settings where groups of observations are assigned together: if all applications from a given month were assigned to a randomly chosen examiner, we would cluster by month. As discussed above, a justification for clustering standard errors may also justify modifying the UJIVE estimator (i.e. to a leave-cluster-out version).

2. Verify balance on other observables

The as-good-as-random assumption implies that the average predetermined characteristics should be equal across decision-makers, conditional on the necessary controls. We test this implication directly: for each observable characteristic, we run our UJIVE specification with that characteristic as the outcome. Significant coefficients indicate imbalance and potential threats to identification. This unified approach—using the same UJIVE specification for both balance tests and treatment effect estimation—offers important advantages over alternatives. Consider the common practice of regressing observables on constructed leniency measures from a first stage. With many examiners and fixed effects, this can generate mechanical correlations that masquerade as true imbalance.[19] UJIVE avoids these finite-sample biases while maintaining the interpretability of our test: the magnitude of any imbalance directly translates to potential bias in our treatment effects. If venture capital funding has a positive coefficient in this UJIVE balance check, for example, this would imply that the randomly assigned examiners' tendency to approve patents is also positively correlated with patents' ex ante venture capital funding. Such a finding would raise questions about whether the examiners are truly randomly assigned, and the magnitude of the estimated coefficient can be used in quantifying sensitivity to omitted variables.

Table 2 presents balance tests for the Farre-Mensa et al. (2020) reanalysis. Each row reports a UJIVE coefficient from regressing a predetermined covariate on patent approval, instrumenting with examiner indicators and controlling for art unit-by-year fixed effects. The results support quasi-random assignment: seven

[19] Even if one uses a leave-out leniency measure, the coefficient will still be biased due to estimation error in the leniency measure, which generates an errors-in-variables bias. Another approach to testing balance is to regress the predetermined characteristics on the full set of examiner indicators and controls, and report the joint F-test for the examiner indicators. With many examiners, however, the default F-statistic is not valid (Anatolyev & Sølvsten, 2023).

of eight coefficients are statistically insignificant at the 5% level, and all are insignificant at the 1% level. As importantly, the magnitudes are economically negligible: coefficients are typically 10 times smaller than the estimated treatment effects discussed below. Consider the venture capital variable, which might raise particular concern if certain examiners systematically reviewed applications from better-funded startups. We estimate a close-to-zero effect, suggesting no systematic differences (coefficient = -0.024, standard error = 0.035). Similar null results emerge for technical complexity (independent claims), external quality validation (European patent approval), and local entrepreneurial environment (state startup density). The absence of systematic imbalance strengthens our confidence that examiner assignment generates plausibly exogenous variation in patent approval.

Balance tests can also examine post-assignment variables to detect exclusion restriction violations. These tests ask whether decision-makers affect outcomes through channels beyond the specified treatment. In the patent setting, for example, one concern is that examiner assignment affects not only whether an application is approved but also the speed of the application review—which plausibly has independent effects on follow-up innovation outcomes by reducing applicant uncertainty. A UJIVE regression which uses a measure of review speed as the outcome, keeping again the treatment, instrument, and controls, could be used to assess the significance of such potential exclusion restriction violations.[20]

3. Estimate treatment effects by UJIVE and alternative estimators

The UJIVE estimator is the preferred choice for leniency designs as it avoids the subtle potential biases of alternative approaches when there are many decision-makers and controls. Comparing UJIVE estimates and standard errors to those from these alternative approaches—such as 2SLS—is illustrative of these biases. As usual with instrumental variable designs, it can also be instructive to compare the UJIVE estimates to ordinary least squares estimates of the treatment effect as a way to probe the importance of selection bias.

Table 3 presents treatment effect estimates across four specifications. Column 1 reports our preferred UJIVE estimates using examiner indicators as instruments with art unit-by-year controls. Columns 2 and 3 implement 2SLS with alternative instruments: the Farre-Mensa et al. (2020) constructed leniency measure, based on examiner approval rates from earlier periods in Column 2, and the full set of examiner indicators in Column 3. Column 4 shows ordinary least squares results.

The UJIVE estimates reveal positive and significant effects of patent approval on subsequent innovation. Approved startups are more likely to file future patents, receive future approvals, and generate citations, with significant effects on both extensive and intensive margins. The pattern of bias across estimators is instructive. Ordinary least squares produces higher coefficients for subsequent applications but lower coefficients for approvals and citations, suggesting moderate selection that operates differently across outcomes. The 2SLS estimates using examiner indicators (Column 3) fall between ordinary least squares and UJIVE,

[20] Farre-Mensa et al. (2020) conduct a different test, including review speed as a second treatment while also including a second "leniency" instrument capturing the assigned examiner's average review speed for past cases (see also Hegde et al. (2022)). Such multiple-treatment specifications can be difficult to interpret when the effects of patent-granting and review speed are heterogeneous (Bhuller & Sigstad, 2024).

pulled toward the biased ordinary least squares results. This intermediate position confirms our theoretical prediction: with many instruments, 2SLS suffers from finite-sample bias that partially reproduces the selection problem it aims to solve. The UJIVE estimates avoid this bias, providing our most credible evidence that initial patent approval causally stimulates follow-on innovation. These effects operate through multiple channels—approved firms not only attempt more patents but also succeed in getting them approved and generating scientific impact through citations.[21] The 2SLS estimates using the Farre-Mensa et al. (2020) measure of leniency tend to be *larger* than UJIVE. Since their leniency construction is similar to that of JIVE, this may reflect a many-covariate bias.

Standard errors also tell an important story in Table 3. The many-instrument 2SLS (Column 3) produces standard errors roughly 3–4 times smaller than UJIVE (Column 1)—a difference that reflects statistical pathology rather than efficiency gains. As we have discussed, 2SLS standard errors and estimates are both pulled towards ordinary least squares with many examiners, creating a false sense of precision around the biased estimates. The shrinkage in standard errors occurs even when using the leave-out Farre-Mensa et al. (2020) leniency measure in Column 2, since these do not reflect estimation error in the leniency measure. UJIVE avoids both failures: it delivers unbiased point estimates and standard errors that accurately capture sampling variation.

4. Test monotonicity to assess the plausibility of a LATE interpretation

To ensure that the UJIVE estimates have a clear causal interpretation under heterogeneous effects, we also need a version of first-stage monotonicity. Since the Imbens and Angrist (1994) first-stage monotonicity condition is likely too strong in the examiner setting, we implement a UJIVE test of the weaker average monotonicity assumption: that the average leniency of the examiners who would grant the firm a patent exceeds the average leniency of those who would deny it.

Figure 1 tests average monotonicity using outcome-specific UJIVE regressions. For each possible outcome value, we create an indicator for that value (e.g., a dummy for zero subsequent applications) and interact it with treatment status as our dependent variable, maintaining examiner instruments and art unit-by-year controls. The specification is estimated as in Tables 2 and 3. The fact that all of these UJIVE estimates are positive, with tight 95% confidence intervals, builds support for average monotonicity holding.

5. Characterize compliers to assess external validity

Having established internal validity, we examine whether our complier population represents the broader sample. We estimate complier characteristics using UJIVE with modified outcomes: we regress the interaction of each covariate with the treatment indicator on the treatment itself, maintaining examiner instruments and controls. This recovers weighted averages of complier characteristics for treated compliers using the same weights as our main estimates. Using one minus the treatment as the treatment variable recovers the charac-

[21] In unreported results, we examine how the estimates change when we additionally control for pre-application covariates. Adding such precision controls can reduce standard errors, as footnote 3 discusses. This is not generally the case here, however, since adding the controls also reduces the effective estimation sample given the algorithm the UJIVE estimator uses to eliminate collinearities (see footnote 17).

teristics for untreated compliers. We efficiently pool the two specifications, following Angrist et al. (2023).[22]

Table 4 reveals that compliers closely resemble the full sample. Column 2 reports complier means for eight characteristics, while Column 1 shows population means. None of the differences are statistically significant, while all the complier estimates fall within logical bounds—again supporting average monotonicity. Compliers have similar rates of venture capital funding, comparable technical complexity in their applications, and equivalent representation across patent classes. This representativeness matters for interpretation: our treatment effects likely approximate average effects for the full population of first-time patent applicants, not just an idiosyncratic subset whose outcomes depend on examiner assignment. The validity of our estimates thus appears to extend beyond the specific complier group to inform broader questions about how patent rights affect startup innovation.

Conclusion

Leniency designs allow for estimation of causal effects when expert decision-makers are as-good-as-randomly assigned and exert discretion on a high-stakes treatment. But, as with many things in both life and econometrics, the devil is in the details. Our review emphasizes how subtle choices in estimation and inference can have first-order consequences in practice. The UJIVE estimator emerges as a natural solution to the many-weak instrument and many control problems that plague two-stage least squares estimation, while correctly estimated heteroskedastic-robust standard errors are a natural choice when assignment operates at the individual level. Our practical checklist and reanalysis of Farre-Mensa et al. (2020) shows how UJIVE can also be used for a number of important auxiliary analyses, such as checking balance, testing average monotonicity, and characterizing compliers. We hope this toolkit gives applied researchers more confidence when approaching leniency designs, and we encourage them to consult this manual whenever questions of operation arise.

[22]For a characteristic v_i, one can show this can be done by using $\tilde{x}_i = 2x_i - 1$ as the transformed treatment variable that takes on values $\{-1, 1\}$ instead of $\{0, 1\}$. We then run a UJIVE regression of $v_i \times \tilde{x}_i$ onto $\tilde{x}_i$, maintaining the controls and examiner instruments.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, *97*(457), 284–292. https://doi.org/10.1198/016214502753479419

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, *88*(1), 265–296. https://doi.org/10.3982/ECTA12675

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, *138*(1), 1–35. https://doi.org/10.1093/qje/qjac038

Ackerberg, D. A., & Devereux, P. J. (2009). Improved JIVE estimators for overidentified linear models with and without heteroskedasticity. *Review of Economics and Statistics*, *91*(2), 351–362. https://doi.org/10.1162/rest.91.2.351

Anatolyev, S., & Sølvsten, M. (2023). Testing many restrictions under heteroskedasticity. *Journal of Econometrics*, *236*(1), 105473. https://doi.org/10.1016/j.jeconom.2023.03.011

Anderson, T. W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, *20*(1), 46–63. https://doi.org/10.1214/aoms/1177730090

Angrist, J., Hull, P., & Walters, C. (2023). Methods for measuring school effectiveness. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (pp. 1–60, Vol. 7). Elsevier. https://doi.org/10.1016/bs.hesedu.2023.03.001

Angrist, J., & Kolesár, M. (2024). One instrument to rule them all: The bias and coverage of just-ID IV. *Journal of Econometrics*, *240*(2), Article 105398. https://doi.org/10.1016/j.jeconom.2022.12.012

Angrist, J. D., Imbens, G. W., & Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, *14*(1), 57–67. https://doi.org/10.1002/(SICI)1099-1255(199901/02)14:1<57::AID-JAE501>3.0.CO;2-G

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. https://doi.org/10.2307/2291629

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72

Autor, D. H., & Houseman, S. N. (2010). Do temporary-help jobs improve labor market outcomes for low-skilled workers? Evidence from "Work First". *American Economic Journal: Applied Economics*, *2*(3), 96–128. https://doi.org/10.1257/app.2.3.96

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, *62*(3), 657–681. https://doi.org/10.2307/2951662

Bhuller, M., & Sigstad, H. (2024). 2SLS with multiple treatments. *Journal of Econometrics*, *242*(1), 105785. https://doi.org/10.1016/j.jeconom.2024.105785

Blandhol, C., Bonney, J., Mogstad, M., & Torgovitsky, A. (2025, January). *When is TSLS actually LATE?* (Working Paper No. 29709). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w29709

Bound, J., Jaeger, D. A., & Baker, R. M. (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*(430), 443–450. https://doi.org/10.1080/01621459.1995.10476536

Chan, D. C., Gentzkow, M., & Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, *137*(2), 729–783. https://doi.org/10.1093/qje/qjab048

Chao, J. C., Swanson, N. R., & Woutersen, T. (2023). Jackknife estimation of a cluster-sample IV regression model with many weak instruments. *Journal of Econometrics*, *235*(2), 1747–1769. https://doi.org/10.1016/j.jeconom.2022.12.011

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K., & Woutersen, T. (2012). Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory*, *12*(1), 42–86. https://doi.org/10.1017/S0266466611000120

Chyn, E., Frandsen, B., & Leslie, E. (2025). Examiner and judge designs in economics: A practitioner's guide. *Journal of Economic Literature*, *63*(2), 401–439. https://doi.org/10.1257/jel.20241719

Coulibaly, M., Hsu, Y.-C., Mourifié, I., & Wan, Y. (2024, May). *A sharp test for the judge leniency design*. arXiv: 2405.06156.

de Chaisemartin, C. (2017). Tolerating defiance? Local average treatment effects without monotonicity. *Quantitative Economics*, *8*(2), 367–396. https://doi.org/10.3982/QE601

Dobbie, W., Goldin, J., & Yang, C. S. (2018). The effects of pre-trial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, *108*(2), 201–240. https://doi.org/10.1257/aer.20161503

Dobbie, W., Goldsmith-Pinkham, P., & Yang, C. S. (2017). Consumer bankruptcy and financial health. *The Review of Economics and Statistics*, *99*(5), 853–869. https://doi.org/10.1162/REST_a_00669

Dobbie, W., & Song, J. (2015). Debt relief and debtor outcomes: Measuring the effects of consumer bankruptcy protection. *American Economic Review*, *105*(3), 1272–1311. https://doi.org/10.1257/aer.20130612

Dobbie, W., Liberman, A., Paravisini, D., & Pathania, V. (2021). Measuring bias in consumer lending (N. Gennaioli, Ed.). *The Review of Economic Studies*, *88*(6), 2799–2832. https://doi.org/10.1093/restud/rdaa078

Doyle, J. J., Jr. (2007). Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, *97*(5), 1583–1610. https://doi.org/10.1257/aer.97.5.1583

Evdokimov, K., & Kolesár, M. (2018, January). *Inference in instrumental variable regression analysis with heterogeneous treatment effects*. https://www.princeton.edu/~mkolesar/papers/het_iv.pdf

Farre-Mensa, J., Hegde, D., & Ljungqvist, A. (2020). What is a patent worth? Evidence from the U.S. patent "lottery". *The Journal of Finance*, *75*(2), 639–682. https://doi.org/10.1111/jofi.12867

Frandsen, B., Lefgren, L., & Leslie, E. (2023). Judging judge fixed effects. *American Economic Review*, *113*(1), 253–277. https://doi.org/10.1257/aer.20201860

Frandsen, B., Leslie, E., & McIntyre, S. (2025). Cluster jackknife instrumental variables estimation. *Review of Economics and Statistics*. https://doi.org/https://doi.org/10.1162/rest.a.263

Goldsmith-Pinkham, P., Hull, P., & Kolesár, M. (2024). Contamination bias in linear regressions. *American Economic Review*, *114*(12), 4015–51. https://doi.org/10.1257/aer.20221116

Heckman, J. J., Urzúa, S., & Vytlacil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, *88*(3), 389–432. https://doi.org/10.1162/rest.88.3.389

Heckman, J. J., & Vytlacil, E. J. (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica*, *73*(3), 669–738. https://doi.org/10.1111/j.1468-0262.2005.00594.x

Hegde, D., Ljungqvist, A., & Raj, M. (2022). Quick or broad patents? Evidence from u.s. startups. *The Review of Financial Studies*, *35*(6), 2705–2742. https://doi.org/10.1093/rfs/hhab097

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475. https://doi.org/10.2307/2951620

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, *83*(5), 2043–2063. https://doi.org/10.3982/ECTA11974

Kolesár, M. (2013, November). *Estimation in an instrumental variables model with treatment effect heterogeneity* [Working paper, Princeton University]. https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf

Kolesár, M., & Plagborg-Møller, M. (2025). Dynamic causal effects in a nonlinear world: The good, the bad, and the ugly. *Journal of Business and Economic Statistics*, *43*(4).

Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E., & Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business and Economic Statistics*, *33*(4), 474–484. https://doi.org/10.1080/07350015.2014.978175

Lee, S. (2018). A consistent variance estimator for 2SLS when instruments identify different LATEs. *Journal of Business & Economic Statistics*, *36*(3), 400–410. https://doi.org/10.1080/07350015.2016.1186555

Lemley, M. A., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. *The Review of Economics and Statistics*, *94*(3), 817–827. https://doi.org/10.1162/REST_a_00194

Matsushita, Y., & Otsu, T. (2024). Jackknife Lagrange multiplier test with many weak instruments. *Econometric Theory*, *40*(2), 447–470. https://doi.org/10.1017/S0266466622000433

Mikusheva, A., & Sun, L. (2022). Inference with many weak instruments. *The Review of Economic Studies*, *89*(5), 2663–2686. https://doi.org/10.1093/restud/rdab097

Mogstad, M., Santos, A., & Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, *86*(5), 1589–1619. https://doi.org/10.3982/ECTA15463

Mullainathan, S., & Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, *137*(2), 679–727. https://doi.org/10.1093/qje/qjab046

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, *27*(4), 575–595. https://doi.org/10.2307/1909352

Phillips, G. D. A., & Hale, C. (1977). The bias of instrumental variable estimators of simultaneous equation systems. *International Economic Review*, *18*(1), 219–228. https://doi.org/10.2307/2525779

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, *65*(329), 161–172. https://doi.org/10.1080/01621459.1970.10481070

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*(2), 135–146. https://doi.org/10.1093/oxfordjournals.oep.a041827

Sampat, B., & Williams, H. L. (2019). How do patents affect follow-on innovation? Evidence from the human genome. *American Economic Review*, *109*(1), 203–236. https://doi.org/10.1257/aer.20151398

Sigstad, H. (2025, March). *Monotonicity among judges: Evidence from judicial panels and consequences for judge IV designs* (SSRN Working paper), BI Norwegian Business School. https://doi.org/10.2139/ssrn.4534809

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557–586. https://doi.org/10.2307/2171753

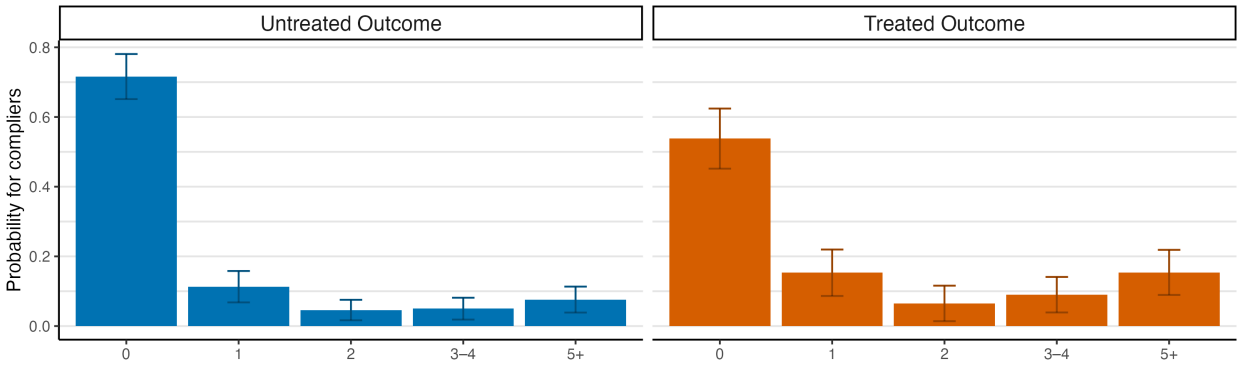
Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear iv regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of thomas rothenberg* (pp. 80–108). Cambridge University Press. https://doi.org/10.1017/CBO9780511614491.006

Vytlacil, E. J. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, *70*(1), 331–341. https://doi.org/10.1111/1468-0262.00277

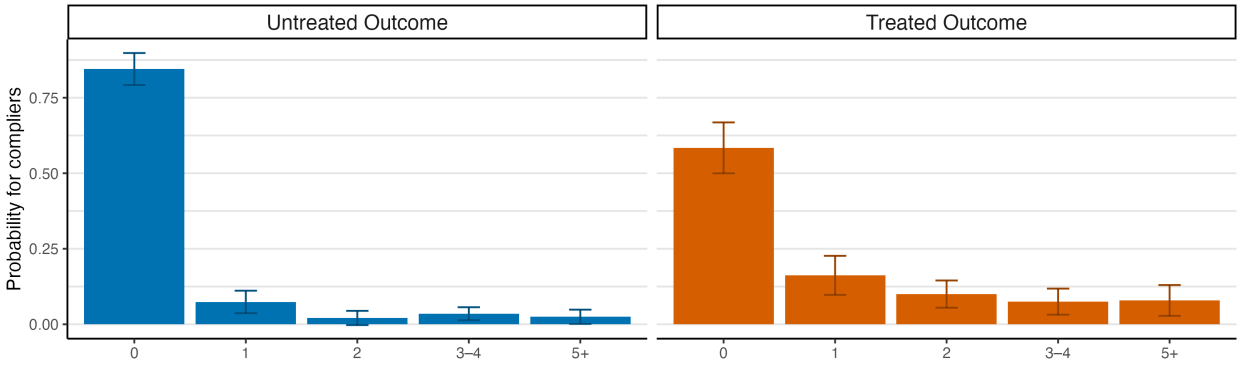
Yap, L. (2025, April). *Inference with many weak instruments and heterogeneity*. arXiv: 2408.11193.

Figure 1: Monotonicity Checks

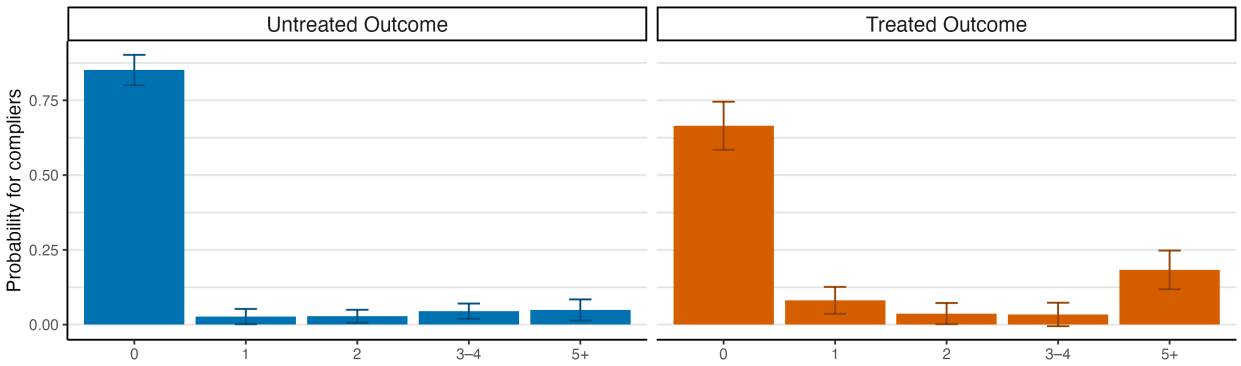
(a) # of subsequent applications



(b) # of subsequent approved applications



(c) # of citations to subsequent patents



Notes: This figure visualizes the distributions of treated and untreated outcomes of each outcome variable in Panel A of Table 1. Outcomes for treated and untreated compliers are estimated separately with UJIVE as described in the text. Vertical bars indicate 95% confidence intervals that are robust to heteroskedasticity and treatment effect heterogeneity.

Table 1: Farre-Mensa et al. (2020) Reanalysis Sample

	Mean (1)	Std. Dev. (2)	# Obs. (3)
Panel A: Outcomes			
Any subsequent application	0.404	0.491	32,514
# subsequent applications	2.339	9.594	32,514
Any subsequent approved application	0.299	0.458	32,514
# subsequent approved applications	1.276	6.021	32,514
Any citation to subsequent patents	0.252	0.434	32,514
# citations to subsequent patents	5.786	52.808	32,514
Panel B: Treatment			
Application approved	0.649	0.477	32,514
Panel C: Covariates			
Patent class group I	0.162	0.368	32,514
Patent class group II	0.410	0.492	32,514
Patent class group III	0.445	0.497	32,514
# independent claims in application	3.730	3.174	27,226
# VC rounds before application	0.124	0.577	32,514
# startups in HQ state	317.3	354.3	32,514
Approval by European Patent Office	0.372	0.484	3,287
Approval by Japanese Patent Office	0.400	0.490	1,206

Notes: This table reports means and standard deviations for the sample used in reanalyzing Farre-Mensa et al. (2020). Variables are defined as follows. *# subsequent applications* is the number of applications with a filing date greater than the first-action date of a firm's first application and *any subsequent application* equals one if any applications were made. *# subsequent approved applications* is the number of approved applications with a filing date greater than the first-action date of a firm's first application and *any subsequent approved application* equals one if any applications were approved. *# citations to subsequent patents* is the number of citations received by all subsequent patent applications over the five years after each patent application's public disclosure date, and *any citation to subsequent patents* equals one if any such citation was received. Patent class groups are defined by related subject matter, as provided by the US Patent Office (see https://www.uspto.gov/sites/default/files/patents/resources/classification/classescombined.pdf). *Patent class group I* includes chemical and related arts, *Patent class group II* includes communications, radiant energy, weapons, electrical, and computer arts, and *Patent class group III* includes material science, mechanical manufacturing and power, and related arts. *# independent claims in application* counts independent claims made in the patent application. *# VC rounds before application* is the number of venture capital funding rounds the startup had secured prior to its first patent application.*# startups in HQ state* counts the total number of startups headquartered in the same state of the applicant in the same year as the applicant. *Approval by European Patent Office* and *Approval by Japanese Patent Office* take value 1 if a patent was approved by a European and Japanese patent office, respectively, take value 0 if an application was made but not approved, and are missing if no application was made. *Years from application to first action* counts the years between the patent application and the first action on the patent application.

Table 2: Balance Checks

	Coefficient (1)	Std. Error (2)	# Obs. (3)
Patent class group I	-0.040	0.017	32,514
Patent class group II	0.011	0.021	32,514
Patent class group III	0.032	0.021	32,514
Log(# independent claims in application)	0.066	0.084	27,226
Log(1 + # VC rounds before application)	-0.024	0.035	32,514
Log(# startups in HQ state)	-0.273	0.143	32,514
Approval by European Patent Office	0.013	0.212	3,287
Approval by Japanese Patent Office	0.525	1.258	1,206

Notes: This table reports results of UJIVE regressions of each covariate in Panel C of Table 1 on application approval, instrumenting with examiner indicators. All estimates control for art unit-by-year fixed effects. The reported standard errors are robust to heteroskedasticity and treatment effect heterogeneity.

Table 3: Treatment Effect Estimates

	UJIVE	2SLS		OLS
	Examiner indicators (1)	FMHL approval rate (2)	Examiner indicators (3)	(4)
Any subsequent application	0.173	0.265	0.232	0.234
	(0.055)	(0.023)	(0.016)	(0.006)
Log(1 + # subsequent applications)	0.323	0.456	0.374	0.357
	(0.100)	(0.037)	(0.027)	(0.009)
Any subsequent approved application	0.259	0.250	0.240	0.223
	(0.050)	(0.020)	(0.014)	(0.005)
Log(1 + # subsequent approved applications)	0.356	0.362	0.323	0.291
	(0.081)	(0.029)	(0.021)	(0.007)
Any citation to subsequent patents	0.183	0.210	0.173	0.164
	(0.049)	(0.020)	(0.014)	(0.005)
Log(1 + # citations to subsequent patents)	0.419	0.480	0.372	0.339
	(0.125)	(0.044)	(0.033)	(0.011)

Notes: This table reports estimates of the effect of application approval on each outcome in Panel A of Table 1. Column 1 estimates effects with UJIVE, instrumenting with examiner indicators. Columns 2 and 3 instead estimate effects with 2SLS, instrumenting with examiners' approval rate as computed by Farre-Mensa et al. (2020) and examiner indicators respectively. Column 4 estimates effects by ordinary least squares. All estimates control for art unit-by-year fixed effects. The sample size is 32,514 in all specifications. Standard errors, reported in parentheses, are robust to heteroskedasticity and treatment effect heterogeneity.

Table 4: Complier Characteristics

	Sample Mean (1)	Complier Mean (2)	# Obs. (3)
Patent class group I	0.162	0.199	32,514
		(0.031)	
Patent class group II	0.410	0.415	32,514
		(0.038)	
Patent class group III	0.445	0.395	32,514
		(0.036)	
# independent claims in application	3.730	3.660	27,226
		(0.224)	
# VC rounds before application	0.124	0.158	32,514
		(0.039)	
# startups in HQ state	317.3	329.3	32,514
		(21.9)	
Approval by European Patent Office	0.372	0.201	3,287
		(0.097)	
Approval by Japanese Patent Office	0.400	0.430	1,206
		(0.518)	

Notes: This table reports means of each covariate in Panel C of Table 1 for the full sample and for compliers. Complier means are estimated with UJIVE as described in the text. Standard errors, reported in parentheses, are robust to heteroskedasticity and treatment effect heterogeneity.

Appendix A Derivations

Estimation with Many Decision-makers

To analyze the estimators in the main text, we follow the literature on many instruments (e.g. Bekker, 1994; Chao et al., 2012), and condition on the realizations on the instruments and covariates, $\{z_i, w_i\}_{i=1}^n$, throughout. To ease notation, we keep the conditioning implicit, writing, e.g., $E[y_i]$ rather than $E[y_i \mid \{z_i, w_i\}_{i=1}^n]$. To simplify the analysis, we also assume that the first-stage is correctly specified, so that the residual ν_i is mean zero, and not just uncorrelated with the instruments and covariates,

$$E[\nu_i] = 0. \tag{4}$$

Since we are assuming that the instrument is as-good-as-randomly assigned conditional on the covariates, it follows that $E[\varepsilon_i]$ depends only on w_i and not on z_i. To simplify derivations, we further assume that the dependence is linear:

$$E[\varepsilon_i] = w_i'\alpha \tag{5}$$

for some vector α. Similar to, for example, Chao et al. (2012) we could allow for asymptotically negligible non-linearities in both Equations (4) and (5) at the cost of complicating the derivations without affecting the results. Finally, we assume that the errors (ε_i, ν_i) are independent across i.

First consider the infeasible estimator $\hat{\beta}^*$. Since $\tilde{\ell}_i$ depends only on covariates, it follows using eq. (4) that $E[\sum_i \tilde{\ell}_i x_i] = \sum_i \tilde{\ell}_i \ell_i = \sum_i \tilde{\ell}_i^2$. Similarly, using Equation (5), we have $E[\sum_i \tilde{\ell}_i \varepsilon_i] = \sum_i \tilde{\ell}_i w_i'\alpha = 0$, where the last equality uses the fact that $\tilde{\ell}_i$ and w_i are orthogonal by definition of a sample residual, $\sum_i \tilde{\ell}_i w_i = 0$. Approximating the expectation of a ratio by a ratio of expectations, we therefore obtain

$$E[\hat{\beta}^*] - \beta \approx \frac{0}{\sum_i \tilde{\ell}_i^2} = 0.$$

While using the ratio of expectations to approximate the expectation of a ratio may appear ad hoc, one can show that the approximation becomes exact asymptotically under the many instrument asymptotics proposed by Bekker (1994), where the dimension K of the instrument vector and the dimension L of the covariate vector grow with sample size—the results we give below mirror the asymptotic bias results given in Evdokimov and Kolesár (2018), for instance. Therefore, an alternative interpretation of the approximation $\approx$ is that it corresponds to the probability limit under the Bekker (1994) sequence.

To derive the bias expressions for the other estimators, it is convenient to use matrix notation. We denote the vectors of n treatment and outcome observations by x and y, respectfully, and denote the matrix of covariates and instruments with rows w_i' and z_i' by W and Z, respectively. To state the results, it will be convenient to link the sum of squares of the relative leniency, $\sum_i \tilde{\ell}_i^2$ to the first-stage F statistic and the partial R^2. Recall that the first-stage F statistic for testing the null that $\pi = 0$ is, under homoskedastic errors, given by $F = \hat{\pi}'\tilde{Z}'\tilde{Z}\hat{\pi}/K \operatorname{var}(\eta_i)$, with expectation $E[F] = \sum_i \tilde{\ell}_i^2/K \operatorname{var}(\eta_i) + 1$. The population partial R^2 from adding the instruments to the first-stage regression is given by $R^2 = (\sum_i \tilde{\ell}_i^2)/(\sum_i \tilde{\ell}_i^2 + n \operatorname{var}(\eta_i))$ so

$$\frac{\sum_i \tilde{\ell}_i^2}{\operatorname{var}(\eta_i)} = K(E[F] - 1) = \frac{nR^2}{1 - R^2}. \tag{6}$$

This identity allows us to state the bias of a large class of estimators using the first-stage F statistic, as the following lemma shows.

Lemma 1. *Consider an estimator $\hat{\beta}_G = y'Gx/x'Gx$ based on a relative leniency measure Gx such that $Gx = \tilde{\ell} + G\nu$. Then, under homoskedastic errors, $E[\hat{\beta}_G] - \beta \approx \frac{\operatorname{tr}(G)\operatorname{cov}(\varepsilon_i,\nu_i)}{\sum_i \tilde{\ell}_i^2 + \operatorname{tr}(G)\operatorname{var}(\eta_i)} = \frac{\operatorname{tr}(G)}{K(E[F]-1)+\operatorname{tr}(G)} \frac{\operatorname{cov}(\varepsilon_i,\nu_i)}{\operatorname{var}(\eta_i)}$.*

The lemma follows by noting that $E[x'Gx] = \sum_i \tilde{\ell}_i^2 + E[\nu'G\nu] = \sum_i \tilde{\ell}_i^2 + \sum_i G_{ii} E[\nu_i^2]$, and $E[\varepsilon'Gx] = \sum_i G_{ii} \operatorname{cov}(\varepsilon_i, \nu_i)$. Under homoskedastic errors, $E[\nu_i^2]$ and $\operatorname{cov}(\varepsilon_i, \nu_i)$ do not depend on the index i, which yields the first expression. The second expression follows directly from (6).

The ordinary least squares estimator $\hat{\beta}_{OLS} = \sum_i \tilde{x}_i y_i / \sum_i \tilde{x}_i^2$ uses the relative leniency measure Mx, where $M = I - W(W'W)^{-1}W'$ denotes the $n \times n$ annihilator matrix associated with the covariates. Since

the trace of M is $n-L$ where L is the covariate dimension, it follows from Lemma 1 that its bias is approximately

$$E[\hat{\beta}_{OLS}] - \beta \approx \frac{1 - L/n}{K(E[F]-1)/n + 1 - L/n} \frac{\text{cov}(\varepsilon_i, \nu_i)}{\text{var}(\eta_i)} \approx (1 - R^2) \frac{\text{cov}(\varepsilon_i, \nu_i)}{\text{var}(\eta_i)},$$

where the second approximation uses (6) and the assumption that L/n is close to zero.

The 2SLS estimator uses the relative leniency measure $\tilde{z}_i \hat{\pi}$, which in matrix notation may be written Hx, with $H = \tilde{Z}(\tilde{Z}'\tilde{Z})^{-}\tilde{Z}'$ denoting the projection matrix associated with the instrument $\tilde{z}_i$. This matrix has trace equal to K, which yields

$$E[\hat{\beta}_{TSLS}] - \beta \approx \frac{1}{E[F]} \frac{\text{cov}(\varepsilon_i, \nu_i)}{\text{var}(\eta_i)}$$

Hence, the relative bias of 2SLS, $(E[\hat{\beta}_{TSLS}] - \beta)/(E[\hat{\beta}_{OLS}] - \beta)$, is approximately given by $\frac{1}{(1-R^2)E[F]}$, as claimed in the main text.

To show approximate unbiasedness of UJIVE, write its leniency measure using matrix notation as $G_{UJIVE}x$ with $G_{UJIVE} = H - \text{diag}(H_{ii}/(M_{ii} - H_{ii}))(M - H)$. The unbiasedness property then follows directly from Lemma 1 by noting that $\text{tr}(G_{UJIVE}) = 0$.

For JIVE, the relative leniency measure can be written as $G_{JIVE}x$ with $G_{JIVE} = M(I - D_Q)^{-1}(H_Q - D_Q)$, where $H_Q = H + W(W'W)^{-1}W'$ is the projection matrix associated with the full vector of right-hand side variables (z_i, w_i), and D_Q is its diagonal. Since $\text{tr}(G) = \text{tr}((I - D_Q)^{-1}(H - D_Q M)) = -L$, we obtain

$$E[\hat{\beta}_{JIVE} - \beta] \approx -\frac{L}{K(E[F]-1) - L} \frac{\text{cov}(\varepsilon_i, \nu_i)}{\text{var}(\eta_i)},$$

which equals the 2SLS bias times $-\frac{E[F]}{K(E[F]-1)/L-1}$, as claimed in the main text.

For IJIVE, the G matrix takes the form $G_{IJIVE} = M(I - \text{diag}(H))^{-1}(H - \text{diag}(H))M$, with trace equal to $\text{tr}((I - \text{diag}(H))^{-1}(H - \text{diag}(H)M)) = \sum_i \frac{H_{ii}(1-M_{ii})}{1-H_{ii}}$. While this quantity is always positive, under balanced design, i.e. when the diagonal elements of H_{ii} are all approximately equal, it simplifies to $LK/(n-K)$, which is much smaller than $\text{tr}(G_{JIVE})$.

To derive the bias-corrected 2SLS estimator, observe that the proof of Lemma 1 shows that under homoskedasticity, the denominator and numerator of the 2SLS formula have expectations $\sum_i E[\hat{\ell}_i x_i] = \sum_i \tilde{\ell}_i^2 + K \,\text{var}(\nu_i)$, and $\sum_i E[\hat{\ell}_i y_i] = \sum_i \tilde{\ell}_i^2 \beta + K \,\text{cov}(\nu_i \beta + \varepsilon_i, \nu_i)$, respectively. An unbiased estimator of $\text{var}(\nu_i)$ obtains as the degrees-of-freedom adjusted sample variance of the first-stage residuals, $\hat{\nu} = (M - H)x$, given by $\widehat{\text{var}}(\nu_i) = \sum_i \hat{\nu}_i^2/(n - K - L) = x'(M - H)x/(n - K - L)$. Since $\nu_i \beta + \varepsilon_i$ corresponds to the reduced form residual from regressing y onto Z and W, the degrees-of-freedom adjusted sample covariance between the first-stage residuals and reduced-form residuals, $(M - H)y$, given by $y'(M - H)x/(n - K - L)$, gives an unbiased estimator of $\text{cov}(\nu_i \beta + \varepsilon_i, \nu_i)$. Subtracting estimates of 2SLS numerator and denominator bias gives the bias-corrected 2SLS estimator,

$$\hat{\beta}_{B2SLS} = \frac{\sum_i \hat{\ell}_i y_i - K y'(M - H)x/(n - K - L)}{\sum_i \hat{\ell}_i x_i - K x'(M - H)x/(n - K - L)} = \frac{y'Hx - K y'(M - H)x/(n - K - L)}{x'Hx - K x'(M - H)x/(n - K - L)},$$

which uses the leniency measure $G_{B2SLS}x$ with $G_{B2SLS} = H - K(M - H)/(n - K - L)$. This has trace equal to 0, and $\hat{\beta}_{B2SLS}$ is therefore unbiased under homoskedasticity. This version of the bias-corrected 2SLS estimator corresponds to the version studied in Kolesár et al. (2015).

Finally, we show that the UJIVE estimator and the FEJIV estimator of Chao et al. (2023) can be interpreted as minimizing the mean-squared error of the estimated leniency under homoskedasticity. It follows from the proof of Lemma 1 that a leniency measure Gx is unbiased if the diagonal of G is zero, $G_{ii} = 0$, and if $GW = 0$ and $G\tilde{Z} = \tilde{Z}$ (the latter two conditions are equivalent to thecondition $Gx = \tilde{\ell} + G\nu$ in Lemma 1). Under these conditions, the mean squared error of the leniency measure is $E[(Gx - \tilde{\ell})'(Gx - \tilde{\ell})] = E[\nu' G' G \nu]$. When ν_i is homoskedastic, this expectation simplifies to $\text{var}(\nu_i)\,\text{tr}(G'G)$. Therefore, minimizing the mean squared error subject to an unbiasedness constraint is equivalent to minimizing $\text{tr}(G'G)$ subject to $\text{diag}(G) = 0$, $GW = 0$, and $G\tilde{Z} = \tilde{Z}$. The first-order condition for the Lagrangian associated with this

problem is given by

$$G' = W\Pi' + \tilde{Z}A' + \text{diag}(\lambda),$$

where Π is the matrix of the Lagrange multipliers associated with the constraint $GW = 0$, A is matrix of the Lagrange multipliers associated with the constraint $G\tilde{Z} = \tilde{Z}$, and λ is the vector of Lagrange multipliers associated with the constraint $\text{diag}(G) = 0$. Multiplying the first-order condition by W' and $\tilde{Z}'$, respectively, allows us to solve for Π and A: $\Pi' = -(W'W)^{-1}W'\,\text{diag}(\lambda)$ and $A' = (\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'(I - \text{diag}(\lambda))$. Plugging this back into the first-order condition then yields $G' = H + (M - H)\,\text{diag}(\lambda)$. Since the diagonal of G is zero, this implies $\lambda_i = H_{ii}/(M_{ii} - H_{ii})$. Thus, $G_{UJIVE} = H + \text{diag}(H_{ii}/(M_{ii} - H_{ii}))(M - H)$ solves the minimization problem: the UJIVE leniency measure minimizes the mean squared error subject to an unbiasedness constraint.

Now consider the same minimization problem, but subject to the additional constraint that $W'G = 0$ and $\tilde{Z}'G = \tilde{Z}'$. The first additional constraint implies that $W'Gx = 0$, so that the leniency measure is orthogonal to the covariates. The second additional constraint ensures that the in-sample covariance of the leniency measure with $\tilde{Z}$ is the same as the in-sample covariance of x with $\tilde{Z}$: $\tilde{Z}'Gx = \tilde{Z}'x$. These additional constraints imply that the resulting estimator corresponds to the MINQUE estimator in Rao (1970); they also ensure that the optimal G matrix is symmetric. Under these additional constraints, the first-order condition becomes

$$G' = W\Pi' + \tilde{\Pi}'W + \tilde{Z}A' + \tilde{A}'\tilde{Z}' + \text{diag}(\lambda).$$

where $\tilde{\Pi}$ is the matrix of Lagrange multipliers associated with the constraint $W'G = 0$ and $\tilde{A}$ is the matrix of Lagrange multipliers associated with the constraint $\tilde{Z}'G = \tilde{Z}$. Solving for the Lagrange multipliers as before, we obtain $G' = H - (M - H)\,\text{diag}(\lambda)(M - H)$. Since the diagonal of G is zero, this implies $\text{diag}(H) = \text{diag}((M - H)\,\text{diag}(\lambda)(M - H))$, which is equivalent to $\text{diag}(H) = ((M - H) \odot (M - H))\lambda$, where $A \odot B$ denotes the Hadamard (elementwise) product of two matrices, $(A \odot B)_{ij} = A_{ij}B_{ij}$. Provided that $(M - H) \odot (M - H)$ is invertible, we therefore obtain the solution

$$G_{FEJIV} = H - (M - H)\,\text{diag}(\lambda)(M - H), \qquad \lambda = ((M - H) \odot (M - H))^{-1}\,\text{diag}(H),$$

which corresponds precisely to the FEJIV estimator of Chao et al. (2023). Because the minimization imposes additional constraints, it follows that when ν_i is homoskedastic, $E[(G_{UJIVE}x - \tilde{\ell})'(G_{UJIVE}x - \tilde{\ell})] \leq E[(G_{FEJIV}x - \tilde{\ell})'(G_{FEJIV}x - \tilde{\ell})]$. Since the matrix $(M - H) \odot (M - H)$ has dimension $n \times n$, its inversion may be challenging in large datasets (if n is in the tens or hundreds of thousands). However, the additional constraint $W'G$ ensures that the FEJIV estimator is invariant adding linear functions of the covariates to the outcome equation: this desirable property is not shared by UJIVE.

Heterogeneous Effects

We now relax the assumption that treatment effects are constant and give a simple statement of the local average treatment effect theorem of Imbens and Angrist (1994). As before, assume that there are K examiners, with $k(i)$ denoting the examiner assigned to observation i. In a departure from previous derivations, we no longer condition on the instruments and covariates, but treat the sample $\{y_i, x_i, z_i, w_i\}_{i=1}^n$ as *iid*. Let $\ell(k, w) = E[x_i \mid k(i) = k, w_i = w]$ denote the absolute leniency of examiner k for observations covariates w and let $p_k(w) = P(k(i) = k \mid w_i = w)$ denote the conditional probability of being assigned examiner k.

Consider using the population version of relative leniency as an instrument: $\ddot{\ell}_i = E[x_i \mid k(i) = k, w_i] - E[x_i \mid w_i]$ (while the relative leniency measure $\tilde{\ell}_i$ uses the sample residual from the regression of absolute leniency onto the covariates w_i as an instrument, $\ddot{\ell}_i$ uses the population residual). Then $\ddot{\ell}_i = \ell(k(i), w_i) - \sum_{j=1}^J p_j(w_i)\ell(j, w_i)$. Using $\ddot{\ell}_i$ as an instrument in a population IV regression of y_i onto x_i without any controls identifies

$$\beta^* = \frac{E[\ddot{\ell}_i y_i]}{E[\ddot{\ell}_i x_i]}.$$

The next result shows that β^* can be written as a weighted average of simple IV regressions restricted to

the subpopulation with $w_i = w$ and $k(i) = k$ or $k(i) = j$. These simple IV regressions identify

$$\beta(j,k,w) = \frac{E[y_i \mid k(i) = k, w_i = w] - E[y_i \mid k(i) = j, w_i = w]}{\ell(k,w) - \ell(j,w)}.$$

Lemma 2. *The leniency IV estimand β^* may be written as a weighted average of pairwise IV regressions,*

$$\beta^* = \frac{\sum_{k>j} E[\omega(j,k,w_i)\beta(k,j,w_i)]}{\sum_{k>j} E[\omega(j,k,w_i)]}, \quad \omega(j,k,w) = p_j(w)p_k(w)(\ell(k,w) - \ell(j,w))^2. \tag{7}$$

Proof. By iterated expectations,

$$\begin{aligned} E[\ddot{\ell}_i y_i] = E[\ddot{\ell}_i E[y_i \mid k(i), w_i]] &= E\left[\ddot{\ell}_i \sum_j p_j(w_i) \left(E[y_i \mid k(i), w_i] - E[y_i \mid k(i) = j, w_i]\right)\right] \\ &= E\left[\ddot{\ell}_i \sum_j p_j(w_i)(\ell(k(i), w_i) - \ell(j, w_i))\beta(k(i), j, w_i)\right], \end{aligned}$$

where the second equality uses $E[\ddot{\ell}_i \mid w_i] = 0$. Plugging in $\ddot{\ell}_i = \ell(k(i), w_i) - \sum_{j=1}^J p_j(w)\ell(j, w_i)$ yields

$$E[\ddot{\ell}_i y_i] = E\left[\sum_{k,j} p_j(w_i)p_k(w_i)\left(\ell(k,w_i) - \sum_{j'=1}^J p_{j'}(w)\ell(j',w_i)\right)(\ell(k,w_i) - \ell(j,w_i))\beta(k,j,w_i)\right].$$

Splitting the double sum $\sum_{k,j}$ into two sums, $\sum_{k>j}$ and $\sum_{k<j}$, and rearranging terms then yields

$$E[\ddot{\ell}_i y_i] = \sum_{k>j} E\left[p_j(w_i)p_k(w_i)(\ell(k,w_i) - \ell(j,w_i))^2\beta(k,j,w_i)\right].$$

By analogous arguments $E[\ddot{\ell}_i x_i] = \sum_{k>j} E[p_j(w_i)p_k(w_i)(\ell(k,w_i) - \ell(j,w_i))^2]$, yielding the result. □

The result shows that leniency IV identifies a weighted average of simple IV regressions, where we restricted the sample to observations assigned to one of two examiners in a given pair (j,k), and restrict the covariates to equal to a particular value. The weights are proportional to the product of the conditional assignment probabilities to j and k times the squared differences in examiner leniency.

Lemma 2 is an algebraic result, and holds without any assumptions on the model. To give a causal interpretation to β^*, we need a causal interpretation for $\beta(j,k,w)$. To this end, if the instruments are as-good-as-randomly assigned conditional on w_i, and an exclusion restriction holds (but we do not impose monotonicity), Proposition 3 in Angrist et al. (1996) shows that $\beta(j,k,w)$ is given by a weighted difference between treatment effects for compliers (those who are treated when assigned to k but not when assigned to j), and defiers (those who are treated when assigned to j but not when assigned to k), where the weight on defiers is negative and equal to the proportion of defiers divided by the leniency difference while the weight on compliers is positive and equals to the proportion of compliers divided by the leniency difference,

$$\beta(j,k,w) = \frac{E[\beta_i C_i(j,k,w)]}{\ell(k,w) - \ell(j,w)},$$

where β_i is the individual treatment effect, and $C_i(j,k,w)$ denotes a variable that equals 1 if an observation is a complier and -1 if they are a defier, so that $\ell(k,w) - \ell(j,w) = E[C_i(j,k,w)]$. Under monotonicity, there are no defiers, so that $\beta(j,k,w)$ corresponds to the average treatment effect for compliers. Combining this interpretation with Lemma 2 then yields the version of the local average treatment effect theorem we

referred to in the main text. In the presence of defiers, we instead obtain:

$$\beta^* = \frac{E[\lambda_i \beta_i]}{E[\lambda_i]}, \quad \lambda_i = \sum_{k>j} p_j(w_i) p_k(w_i)(\ell(k, w_i) - \ell(j, w_i)) C_i(j, k, w_i). \tag{8}$$

Thus, the total weight on observation i equals λ_i, the "average" complier status of observation i: a weighted difference between the number of times the observation i is a complier, minus the number of times the observation is a defier.

To derive a necessary and sufficient condition for the individual weight λ_i to be positive, let $x_i(k)$ denote the potential treatment indicator that equals 1 if i examiner k would treat observation i. Then $C_i(j, k, w) = x_i(k) - x_i(j)$, so that

$$\begin{aligned}
\lambda_i &= \sum_k x_i(k) p_k(w_i)(\ell(k, w_i) - \sum_j p_j(w_i)\ell(j, w_i)) \\
&= \sum_k x_i(k) p_k(w_i) \sum_j (1 - x_i(j)) p_j(w_i) \left[\bar{\ell}_i(1) - \bar{\ell}_i(0)\right] \\
&= P(x_i = 1 \mid w_i) P(x_i = 0 \mid w_i) \left[\bar{\ell}_i(1) - \bar{\ell}_i(0)\right],
\end{aligned}$$

where $\bar{\ell}_i(1) = \sum_k x_i(k) p(k, w_i) \ell(k, w_i) / \sum_k x_i(k) p(k, w_i)$ is the average leniency of examiners who would treat observation i, and $\bar{\ell}_i(0) = \sum_k (1 - x_i(k)) p(k, w_i) \ell(k, w_i) / \sum_k (1 - x_i(k)) p(k, w_i)$ is the average leniency of those who would not treat it. Thus the weights are positive if and only if $\bar{\ell}_i(1) \geq \bar{\ell}_i(0)$. The first line in the above display may also be written $\lambda_i = \text{cov}(x_i, \ell(k(i), w_i) \mid w_i)$, so that the condition $\bar{\ell}_i(1) \geq \bar{\ell}_i(0) \geq 0$ is equivalent to the average monotonicity condition in Frandsen et al. (2023) that, if $w_i = 1$ is the only covariate, $\text{cov}(x_i, \ell(k(i), 1)) \geq 0$. This equivalence was noted in Sigstad (2025).

Inference

To derive a standard error for UJIVE that allows for treatment effect heterogeneity, let

$$y_i = z_i' \pi_Y + w_i' \delta_Y + \nu_{Yi}, \qquad E[\nu_{Yi}] = 0 \tag{9}$$

denote the *reduced form* regression of the outcome on instruments and covariates. To avoid technical complications, as in the estimation section of the Appendix, we condition on the instruments and covariates, and in analogy to eq. (4), we assume that the reduced form is correctly specified. The parameter of interest is given by the estimand of the infeasible estimator $\hat{\beta}^*$,

$$\beta^* = \frac{\sum_i E[\tilde{\ell}_i y_i]}{\sum_i E[\tilde{\ell}_i x_i]} = \frac{\sum_i \tilde{\ell}_i z_i \pi_Y}{\sum_i \tilde{\ell}_i^2}.$$

If treatment effects are constant, so that eq. (5) holds, then it follows from eq. (1) and eq. (9) that $\pi_Y = \pi\beta$, and $\beta^* = \beta$. Recall that the UJIVE estimator is given by $\hat{\beta}_{UJIVE} = y' G_{UJIVE} x / x' G_{UJIVE} x$. Since $G_{UJIVE} x = \tilde{\ell} + G_{UJIVE} \nu$, it follows from eqs. (2) and (9) that we may decompose the numerator and denominator of the estimator as

$$\begin{pmatrix} y' G_{UJIVE} x \\ x' G_{UJIVE} x \end{pmatrix} = \sum_i \begin{pmatrix} \tilde{\ell}_i y_i \\ \tilde{\ell}_i x_i \end{pmatrix} + \sum_i \begin{pmatrix} r_{Yi} \nu_i \\ r_i \nu_i \end{pmatrix} + \sum_{i,j} \begin{pmatrix} G_{ij} \nu_{Yi} \nu_j \\ G_{ij} \nu_i \nu_j \end{pmatrix},$$

where G_{ij} is the (i, j) element of G_{UJIVE}, while $r_{Yi} = \sum_j G_{ji}(z_j' \pi_Y + w_j' \delta_Y)$ and $r_i = \sum_j G_{ji}(z_j' \pi + w_j' \delta)$ denote the "signal" in the reduced form and the first stage. The second and third terms represent additional noise components in the estimator relative to the infeasible estimator $\hat{\beta}^*$. Since the third term is uncorrelated with the first two, it follows that the covariance matrix of the left-hand side may be written

$$\Sigma = \sum_i \text{var} \begin{pmatrix} \tilde{\ell}_i y_i + r_{Yi} \nu_i \\ \tilde{\ell}_i x_i + r_i \nu_i \end{pmatrix} + \sum_{i,j} \begin{pmatrix} G_{ij}^2 \sigma_{y,i}^2 \sigma_{x,j}^2 + G_{ij} G_{ji} \sigma_{xy,i} \sigma_{xy,j} & G_{ij}(G_{ij} + G_{ji}) \sigma_{yx,i} \sigma_{x,j}^2 \\ G_{ij}(G_{ij} + G_{ji}) \sigma_{yx,i} \sigma_{x,j}^2 & G_{ij}(G_{ij} + G_{ji}) \sigma_{x,i}^2 \sigma_{x,j}^2 \end{pmatrix}$$

where $\sigma^2_{y,i} = \text{var}(\eta_{Y,i})$, $\sigma^2_{x,i} = \text{var}(\eta_i)$, and $\sigma_{yx,i} = \text{cov}(\eta_{Y,i}, \eta_i)$. Furthermore, the numerator and denominator can be shown to be asymptotically normal by a martingale central limit theorem (Evdokimov & Kolesár, 2018, Lemma D.5) so that

$$\begin{pmatrix} y'G_{UJIVE}x \\ x'G_{UJIVE}x \end{pmatrix} \approx \mathcal{N}\left(\begin{pmatrix} \beta^* \sum_i \tilde{\ell}_i^2 \\ \sum_i \tilde{\ell}_i^2 \end{pmatrix}, \Sigma \right). \tag{10}$$

It then follows by an application of the delta method that in large samples, $\hat{\beta}_{UJIVE}$ has an approximately normal distribution,

$$\hat{\beta}_{UJIVE} \approx \mathcal{N}(\beta^*, \sigma^2_{UJIVE}),$$

where

$$\sigma^2_{UJIVE} = \frac{\sum_i \text{var}(\tilde{\ell}_i(y_i - x_i\beta^*) + (r_{Y,i} - r_i\beta^*)\nu_i) + \sum_{i,j}(G^2_{ij}\sigma^2_{y-x\beta^*,i}\sigma^2_{x,j} + G_{ij}G_{ji}\sigma_{y-x\beta^*,x,i}\sigma_{y-x\beta^*,x,j})}{(\sum_i \tilde{\ell}_i^2)^2}$$

In contrast, by the same arguments, the standard error for the infeasible estimator is given by the square root of $\sigma^2_* = \sum_i \text{var}(\tilde{\ell}_i(y_i - x_i\beta^*))/\sum_i \tilde{\ell}_i^2$.

The $(r_{Y,i} - r_i\beta^*)\nu_i$ term in the numerator arises due to variation in complier treatment effects across complier groups. In particular, under regularity conditions, $\text{var}((r_{Y,i} - r_i\beta^*)\nu_i) \approx \text{var}(\tilde{z}_i'(\pi_Y - \pi\beta^*)\nu_i)$. Under constant treatment effects, the reduced form coefficients are proportional to the first stage, $\pi_Y = \pi\beta$, so that the first term in the above display may be replaced by $\sum_i \text{var}(\tilde{\ell}_i(y_i - x_i\beta))$, or equivalently $\sum_i \varepsilon_i^2\tilde{\ell}_i^2$. The last term in the numerator is the Bekker (1994) many instrument term. It is negligible if the instruments are strong enough in the sense that $E[F]$ is large.

Note that Equation (10) corresponds exactly to Equation (3) in Angrist and Kolesár (2024) who consider an instrumental variables regression with a single instrument, with the variation in the first stage leniency, $\sum_i \tilde{\ell}_i^2$ playing the role of the first-stage regression coefficient on the single instrument, with Σ playing the role of the covariance matrix of between the reduced form and the first stage. It then follows from their analysis that delta-method based inference is not overly optimistic even if the instruments are weak, provided that the correlation between $(y - x\beta^*)'G_{UJIVE}x$ and $x'G_{UIJIVE}x$, given by

$$\rho = \frac{\Sigma_{12} - \Sigma_{22}\beta^*}{\sqrt{\Sigma_{22}(\Sigma_{11} - 2\beta^*\Sigma_{12} + \beta^{*2}\Sigma_{22})}}$$

is not too large: as long as $|\rho| < 0.76$, rejection rates for nominal 5% tests stay below 10% regardless of the instrument strength. In contrast to the single-instrument case, Σ now contains additional terms relative to the infeasible estimator, so that ρ no longer maps to the endogeneity parameter under homoskedasticity. Nonetheless, since Σ is consistently estimable, for any particular null hypothesis of interest, one can plug in the hypothesized value of β^* into the above expression along with estimates of Σ to verify that using the UJIVE standard errors doesn't lead to overrejection. Alternatively, since ρ is monotone in β^*, bounding the treatment effect parameter yields bounds in the plausible values of ρ: if these exclude large value of ρ, confidence intervals based on UJIVE standard errors will have good coverage rates.