A Simple Method for Calculating Economic Magnitudes in Instrumental Variable Research Designs*

Abhiroop Mukherjee† George Panayotov‡ Rik Sen§ Jia Zhang¶

This draft: August 28, 2026

Abstract

Empirical finance papers routinely convey the economic importance of a treatment effect by comparing the causal estimate to a base rate, such as the mean outcome among untreated observations. Surveying all 946 instrumental-variables (IV) papers published in the top-3 finance and top-5 economics journals during 2015–2024, of which 564 are in finance, we find that the base rate is almost always computed over the wrong population: IV estimates pertain to compliers, yet they are compared to base rates that mix compliers with always-takers and never-takers. We show that the correct compliers-only base rate is easily recovered when the treatment is binary, as it often is. Doing so requires only one additional regression on the same data—the researcher's own 2SLS specification, re-run with the outcome Y replaced by $Y(1-D)$, where D is the treatment indicator. We call the ratio of the causal estimate to this correct base rate the complier-adjusted proportional effect (CAPE). For a single binary instrument, CAPE coincides with the proportional LATE of Chen and Roth (2024), who propose it as a unit-invariant percentage effect to side-step the logs-with-zeros issue. We show not only that CAPE's usefulness extends beyond their logs-with-zeros motivation, but also that its complier interpretation applies to multiple and continuous instruments under the partial monotonicity condition of Mogstad, Torgovitsky, and Walters (2021). Replicating nine published papers (including four finance papers) selected for full data availability, we find that conventionally scaled proportional effects differ from CAPE by 0.6 to 249 percentage points in like-for-like comparisons, with errors in both directions.

Keywords: economic magnitudes; instrumental variables; regression discontinuity; local average treatment effects; compliers.
JEL codes: C21, C26, G00.

*We thank Pulak Ghosh, Raj Iyer, S. Lakshmi Naaraayanan, Indira Puri, Amit Seru and seminar participants for helpful comments. All errors are our own.

†The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.
‡The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.
§University of Georgia, United States.
¶The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

1 Introduction

A causal revolution has transformed empirical finance and economics over the past three decades. Instrumental variables (IV) methods—including fuzzy regression discontinuity and related quasi-experimental designs—are now standard tools for estimating the effects of financing, regulation, and institutions. In the top three finance and top five economics journals, our survey (Section 4) identifies 946 IV papers published during 2015–2024 alone, of which 564 study a finance topic. But estimation is only half of what these papers do; the other half is interpretation. To do so, researchers often translate the causal estimate into an effect size by comparing it to a base rate, e.g., by scaling the estimate by the pre-treatment mean of the outcome and reporting the resulting ratio as the percentage effect. For example, Solís (2017) reports that credit access raises college enrollment by 50 percent; Farre-Mensa, Hegde, and Ljungqvist (2020) report that winning the "patent lottery" raises employment growth by 55 percent; and the estimate in Acharya and Xu (2017) implies that NASDAQ-delisted firms produce roughly a third fewer patents. In each case, it is the percentage effect—and not the raw estimate—that is cited in subsequent work, invoked in policy discussions, and used to judge whether an effect is large or small. Yet while identification strategies have received enormous scrutiny, the calculation of correct economic magnitudes implied by the causal estimates has received almost none.

This paper makes a simple point with broad consequences: the reported headline magnitudes are often wrong, because the base rate used is almost always computed over the wrong population. Under the assumptions that give IV estimates a causal interpretation, the estimated effect is a local average treatment effect (LATE): it pertains to *compliers*, the units whose treatment status is moved by the instrument (Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996). The correct proportional effect therefore divides the LATE by the mean outcome compliers would have had absent treatment, $\mathbb{E}[Y_i(0) \mid \text{complier}]$. Common practice instead divides by the full-sample mean, the untreated-sample mean, or the constant from the second-stage regression. Each of these mixes compliers with non-compliers (always-takers and never-takers) and is therefore correct only under the additional—and rarely stated—assumption that the two groups would have had the same mean outcome absent treatment. When that assumption fails, reported percentage effects are biased, in a direction and by a magnitude that cannot be signed a priori. Many researchers are aware of the issue but few explicitly address it; perhaps because they regard it as unfixable, on the grounds that compliers are unobservable.

To fix ideas, consider a cutoff-based loan pricing rule under which borrowers with a credit score of 700 or above are eligible for an APR discount. Compliers are eligible borrowers who

take the discount; eligible borrowers who do not complete the steps needed to obtain it are non-compliers. Under standard regression discontinuity assumptions, the discontinuity at 700 identifies the LATE of the discount on default, an effect driven entirely by compliers. Converting this LATE into a percentage reduction in default risk requires dividing by the counterfactual default rate the same compliers would have had absent the discount. Because compliers cannot be individually identified, a researcher might instead use the observed default rate among borrowers just below the cutoff. But the below-cutoff group also contains non-compliers, and if their default risk differs from compliers', the denominator—and hence the headline magnitude—will be wrong.

We show that the correct complier-only base rate can in fact be recovered easily whenever the treatment is binary—irrespective of whether the IV is binary or continuous, and even in a multiple-IV setting. Binary treatments are common: of the 946 IV papers published in the 10 year period we survey, 273 (28.9%) use a such a treatment, and out of the 564 finance IV papers 152 (27%) have a binary treatment. The binary treatment case is also relevant for our paper: of the 273 binary treatment IV papers there are 133 (48.7%) that explicitly compare a causal estimate to a base rate, but only 17 of them adjust that base rate to reflect the complier population the estimate actually describes.[1] The corresponding numbers are even more stark in finance: 54 of the 152 binary treatment IV papers compare a causal estimate with a base-rate, but only *one* (Kleiner, Stoffman and Yonker, 2021) makes an adjustment to align the estimate with the base rate.

Recovering the correct complier-only base rate in this case is simple because it only requires one additional regression on exactly the same data underlying the causal estimate: the researcher's own second-stage regression, re-run with the outcome Y replaced by $Y(1-D)$, where D is the treatment indicator. The logic is immediate. For non-compliers—always-takers and never-takers—a change in the instrument changes neither D_i nor Y_i, so $Y_i(1-D_i)$ does not respond. For compliers, moving the instrument from $D_i = 0$ to $D_i = 1$ changes $Y_i(1-D_i)$ by exactly $-Y_i(0)$. Running the second-stage regression—with the same instrument, controls, and fixed effects—on $Y(1-D)$ in place of Y therefore recovers the complier-only untreated mean. The ratio of the original 2SLS coefficient to this quantity, which we call the Complier-Adjusted Proportional Effect (CAPE), is precisely the percentage-effect parameter that researchers seek when they report instead a mean-scaled estimate: it is the local average treatment effect in levels, normalized by the untreated base rate of the population to which that effect actually pertains.

While the building blocks that CAPE uses are not new, our contribution is to tie them

[1] There are other papers that among the 273 that make other interpretation errors, e.g., in the log-with-zeros setting which we do not count among these 133.

together into a simple, easy-to-implement measure with general applicability. That complier potential-outcome means are identified was established by Imbens and Rubin (1997) and Abadie (2002, 2003), but their results only applied to a single binary instrument. A handful of empirical papers have used their methodology, e.g., Katz, Kling, and Liebman (2001); Chetty, Hendren, and Katz (2016); Mukherjee et al. (2022). Furthermore, Chen and Roth (2024) have proposed the same proportional LATE and the transformed-outcome estimator that we employ—instrumenting both Y and $Y(1-D)$—but they only do that as a solution to a specific problem, the unit-dependence of $\log(1+Y)$ when outcomes can be zero. We show that our proposed complier-only base rate calculation (and hence the CAPE estimator) is neither limited in its applicability to the single binary instrument case, nor is it restricted to the log with zeros setting.

Our method can easily be applied to solve the correct base rate identification problem in almost all common settings that applied researchers with binary treatments face—even in settings with multiple binary instruments or continuous instruments. This extension is not trivial in practice: out of the 273 published IV papers with binary treatment, 93 (34.1%) have multiple binary instruments or a single continuous instrument. In the continuous instrument case, in particular, our method allows researchers to avoid discretization, which was previously used (Dahl, Kostøl, and Mogstad, 2014) to deliver base rate alignment. The advantage of avoiding discretization is multifold (see Section 3.3): CAPE uses all of the instrument's variation, requires no endpoint or percentile choice, accommodates the original design's covariates and clustering directly, and aggregates margin-specific untreated means with exactly the Angrist–Imbens (1995) weights of the published coefficient, rather than with weights defined by a chosen pair of endpoints.

Our first contribution, which allows the generalization described above, is methodological. We formalize CAPE's properties in the settings applied researchers actually face, particularly where these properties do not follow as corollaries of existing results. With a single binary instrument and homogeneous proportional effects, CAPE is consistent for the common proportional effect. Under heterogeneity, CAPE retains a clean interpretation thanks to a *weight-matching* property: the 2SLS regressions for Y and for $Y(1-D)$ share the same first stage and therefore aggregate margin-specific effects and base rates with identical weights, so numerator and denominator compare like with like. With multiple binary instruments, we build on Mogstad, Torgovitsky, and Walters (2021), who show that the Imbens–Angrist (1994) monotonicity condition can hold with more than one instrument only if choice behavior is effectively homogeneous, and who propose the weaker condition of partial monotonicity—monotonicity in each instrument holding the others fixed—which we adopt. Under partial monotonicity, weight matching implies that both coefficients converge

to weighted averages over instrument-specific complier subgroups with identical weights. CAPE is thus a ratio of identically weighted averages and lies within the range of subgroup-level proportional effects—a bounded-range property that no observed-sample scaling can guarantee. With mutually independent instruments, the shared weights are automatically non-negative; with correlated instruments, CAPE inherits whatever weighting the design induces—no better and no worse than the underlying 2SLS coefficient. With continuous instruments, the same logic delivers CAPE as a ratio of margin-specific effects aggregated with the Angrist–Imbens (1995) weights, again matched across numerator and denominator because the weights depend only on the first stage. The practical procedure is identical in every case. A byproduct of the same identity is a general diagnostic for *who the compliers are*: replacing $Y(1-D)$ with $X(1-D)$ for any predetermined characteristic X recovers complier means of that characteristic, including in multiple- and continuous-instrument settings where the scalar κ-weighting of Abadie (2003) does not directly apply. Finally, we develop inference for CAPE—delta-method, bootstrap, and Fieller intervals—the last of which is important because CAPE is a ratio whose denominator can be small.

Our second contribution is to document practice and quantify the stakes. We replicate nine quasi-experimental studies published in top finance and economics journals—four of them finance papers—selected for data availability from a systematic search, and compare published or implied proportional effects with CAPE across fourteen representative settings spanning binary, share, count, and growth outcomes and single binary, multiple binary, and continuous instruments. The differences are economically consequential and run in both directions. In Farre-Mensa, Hegde, and Ljungqvist (2020), the five-year employment-growth effect of a first patent grant is 176 percent of the complier base rate rather than the 116 percent implied by the broad untreated mean, because the "middling" applications whose fate examiner leniency decides have relatively low untreated growth. In Solís (2017), the enrollment effect of credit access rises from the reported 48.8 percent to 57.3 percent once the base rate excludes always-takers—test retakers with better-educated, higher-income parents, whose untreated enrollment prospects exceed compliers'. Just as important, CAPE sometimes *validates* published interpretations: in Stein and Yannelis (2020), the employment effect of bank access moves only from 14.1 to 14.7 percent, indicating that compliers and non-compliers had comparable untreated labor-force attachment. In that case, CAPE tells researchers that the convenient denominator happened to be the right one.

Outside finance, the stakes can be even larger. In Esposito et al. (2023), the effect of film screenings on navy enlistment falls from a simple-ratio 385 percent (or 224 percent against the sample mean) to 168 percent, because compliers are urban, populous counties that were mechanically more likely to record an enlistment even without treatment. In An-

gelucci, Meraglia, and Voigtländer (2022), a multiple-instrument setting, the proportional effect of municipal self-governance on parliamentary representation falls from 475 to 227 percent. Sampat and Williams (2019) illustrate that continuous instruments pose no obstacle to CAPE—and that log-versus-level scaling can matter more than denominator choice. Asher and Novosad (2020) provide a within-paper diagnostic: CAPE materially revises transport-access magnitudes while closely validating the two main occupational shares, and an imprecise private-bus base warns against treating every point ratio as stable. Across the fourteen settings in Table 3, like-for-like denominator corrections range from 0.6 to 248.7 percentage points and run in both directions; Panel B separately exposes the outcome-scale problem in log specifications.

Given CAPE is the same estimator as Chen and Roth (2024) proposed, it also speaks to the literature on proportional effects when outcomes can be zero. Many finance outcomes—patents, citations, acquisitions, VC rounds—are counts with substantial mass at zero, and the standard practice of reporting $e^{\hat{\beta}}-1$ from a $\log(1+Y)$ regression does not identify a unit-invariant percentage effect (Chen and Roth, 2024; Cohn, Liu, and Wardlaw, 2022). While Chen and Roth (2024) already show that the CAPE regression provides a unit-invariant proportional-effect estimand in such settings, their complier interpretation rests on Abadie (2002) and is therefore established only for a single binary instrument; our weight-matching results supply the multi- and continuous-instrument interpretation they leave unaddressed.[2] We further provide econometric results on testing and diagnostics in the CAPE framework and quantify the difference CAPE makes in practice. In Acharya and Xu (2017), for example, the log-implied 32 percent decline in patenting after delisting maps into a CAPE of -95 percent: the level effect of delisting is nearly the entire untreated patent output of the complier firms. This difference must be interpreted with care. Unlike in the earlier replications, where CAPE is a *correction* that aligns the causal estimate with its proper base rate, here it does not correct the estimate from the standard $\log(1+Y)$ specification. Rather, it provides a conceptually distinct, well-defined estimand of the proportional treatment effect that avoids the pitfalls of the $\log(1+Y)$ estimate when zeros are present.

Overall, this paper speaks to three literatures. Within the econometrics of local average treatment effects (Imbens and Angrist, 1994; Angrist and Imbens, 1995; Angrist, Imbens, and Rubin, 1996; Imbens and Rubin, 1997; Abadie, 2003; Heckman and Vytlacil, 2005), our weight-matching and component-wise-monotonicity results extend complier-mean identification to the ratio estimand under multiple and continuous instruments; more broadly, we convert a known identification fact into an interpretable, general-purpose magnitude param-

[2]Chen and Roth (2024) motivate this regression as a solution to the logs-with-zeros problem; its use for aligning the base rate with the complier population—our focus—does not arise in their paper.

eter with an implementation researchers can readily use. Within the literature on outcome transformations and proportional effects (Cohn, Liu, and Wardlaw, 2022; Mullahy and Norton, 2024), we generalize and justify the complier normalization that Chen and Roth (2024) implement in a single application. And within the methodological literature in empirical finance (Roberts and Whited, 2013; Atanasov and Black, 2016; Jiang, 2017; Mitton, 2022; Baker, Larcker, and Wang, 2022), we shift attention from the identification step, where the credibility conversation has concentrated, to the interpretation step, where economic significance is actually measured—and where, we show, a pervasive and easily correctable error resides. For empiricists, the takeaway is a simple reporting recommendation: alongside the 2SLS coefficient, report the estimated complier base rate, the CAPE ratio with a delta method, bootstrap or Fieller confidence interval (as appropriate), and—to gauge how different the complier population is from the full sample—complier means of key covariates. Each requires only re-running regressions already in the replication package.

Finally, we also want to be explicit about what CAPE cannot do. First, as noted above, CAPE applies only to binary-treatment settings; extending the logic to more general frameworks is left to future research. Second, a word of caution: CAPE delivers a design-consistent scaling parameter for judging the economic magnitude of a LATE estimate, and it does so only if the original design delivers a valid LATE. If the underlying LATE is compromised by questionable identifying assumptions, CAPE inherits those flaws. Conversely, a wide gap between a CAPE estimate and a paper's reported effect size does not imply that the LATE is wrong; on the contrary, CAPE provides the correct economic magnitude under the maintained assumption that the original LATE is valid.

The remainder of the paper proceeds as follows. Section 2 sets out the framework, defines the estimand, and derives the estimator with a single binary instrument. Section 3 develops the theory under heterogeneity, multiple binary instruments, and continuous instruments, and discusses covariates and inference. Section 4 documents how often CAPE-relevant settings arise in top finance and economics journals. Section 5 presents the replications. Section 6 reports simulation evidence on when and why the simple ratio fails. Section 7 distills practical guidance, and Section 8 concludes.

2 Interpreting Magnitudes: The Problem and the Estimator

2.1 Setup and estimand

We adopt the potential-outcomes framework of Angrist, Imbens, and Rubin (1996). There are N units. $\mathbf{Z}$ is an N-vector of assignments (the *instrument*), with ith component Z_i. $D_i(\mathbf{Z})$ indicates unit i's treatment status, and $Y_i(\mathbf{D}, \mathbf{Z})$ its outcome. D_i and Z_i are binary. We maintain:

Assumption 1 (SUTVA)**.** *(a) If $Z_i = Z_i'$ then $D_i(\mathbf{Z}) = D_i(\mathbf{Z}')$. (b) If $Z_i = Z_i'$ and $D_i = D_i'$ then $Y_i(\mathbf{D}, \mathbf{Z}) = Y_i(\mathbf{D}', \mathbf{Z}')$.*

Assumption 2 (Exclusion restriction)**.** $\mathbf{Y}(\mathbf{D}, \mathbf{Z}) = \mathbf{Y}(\mathbf{D}, \mathbf{Z}')$ *for all* $\mathbf{Z}, \mathbf{Z}'$ *and all* $\mathbf{D}$.

Under Assumptions 1–2 we write $D_i(Z_i)$ and $Y_i(D_i)$.

Assumption 3 (Random instrument)**.** *Z_i is i.i.d. Bernoulli(p), independent of potential outcomes and potential treatment states. (Exchangeability or conditional independence would suffice; we assume full randomness for clarity.)*

Assumption 4 (Relevance)**.** $\mathbb{E}[D_i(1) - D_i(0)] \neq 0$.

Assumption 5 (Monotonicity)**.** $D_i(1) \geq D_i(0)$ *for all i.*

Under Assumptions 1–5, the IV (Wald) estimand is the local average treatment effect (LATE):

$$\frac{\mathbb{E}[Y_i(D_i(1), 1) - Y_i(D_i(0), 0)]}{\mathbb{E}[D_i(1) - D_i(0)]} = \mathbb{E}[\, Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1 \,]. \tag{1}$$

The proportional-effect parameter that magnitude discussions aim at is

$$\theta \;\equiv\; \frac{\mathbb{E}\big[Y_i(1) - Y_i(0) \mid D_i(1) - D_i(0) = 1\big]}{\mathbb{E}\big[Y_i(0) \mid D_i(1) - D_i(0) = 1\big]}, \tag{2}$$

the LATE in levels normalized by the mean *untreated* outcome of the *same* population—the compliers—whose treatment effect the numerator measures. We call θ the complier-adjusted proportional effect (CAPE).

2.2 Why simple ratios miss

In practice, researchers scale the LATE by one of three base rates: the full-sample mean $\mathbb{E}[Y]$, the untreated-subsample mean, or the constant term of the second-stage regression. None equals the denominator of (2) in general. The full-sample mean includes treated outcomes and hence part of the effect being scaled. The untreated mean pools never-takers (and, in RDD contexts, control-side compliers) whose untreated outcomes may differ systematically from compliers'. The regression constant is a covariate-evaluated blend of complier and always-taker untreated levels. All three are valid only under a homogeneity condition—that non-compliers' untreated outcomes mirror compliers'—which the selection logic motivating the IV typically contradicts: units select into always- or never-taking precisely on characteristics correlated with untreated outcomes. The direction of the resulting error is not signable a priori; our replications in Section 5 find errors in both directions.

2.3 Identification: the transformed-outcome identity

Although compliers cannot be pointed to in the data—membership depends on counterfactual treatment status—their untreated mean is identified. Consider the effect of Z_i on the transformed outcome $Y_i(1 - D_i)$. When Z_i moves from 0 to 1, this transformed outcome changes by

$$Y_i\big(D_i(1), 1\big)\big(1 - D_i(1)\big) - Y_i\big(D_i(0), 0\big)\big(1 - D_i(0)\big). \tag{3}$$

Under Assumptions 1 and 2, expression (3) simplifies by compliance type $(D_i(0), D_i(1))$: for always-takers $(1, 1)$ it equals $Y_i(1) \cdot 0 - Y_i(1) \cdot 0 = 0$; for never-takers $(0, 0)$ it equals $Y_i(0) - Y_i(0) = 0$; for compliers $(0, 1)$ it equals $-Y_i(0)$; defiers are ruled out by Assumption 5. The key observation is that (3) involves only $Y_i(0)$, never $Y_i(1)$: whenever $D_i = 1$, the factor $(1 - D_i)$ zeroes out the term. Taking expectations and dividing by the first stage,

$$\frac{\mathbb{E}\big[\, Y_i(D_i(0), 0)(1 - D_i(0)) - Y_i(D_i(1), 1)(1 - D_i(1)) \,\big]}{\mathbb{E}[D_i(1) - D_i(0)]} = \mathbb{E}\big[\, Y_i(0) \mid D_i(1) - D_i(0) = 1 \,\big]. \tag{4}$$

Combining (1) and (4), θ is identified, and the estimator is a ratio of two 2SLS coefficients:

$$\widehat{\theta} \;=\; \frac{\hat{\beta}^{Y}_{\text{2SLS}}}{-\,\hat{\beta}^{Y(1-D)}_{\text{2SLS}}}, \tag{5}$$

where the denominator regression differs from the researcher's original 2SLS *only* in its dependent variable and therefore mechanically inherits the instrument, covariates, fixed effects, running-variable controls, bandwidth, weights, and clustering of the original design. This is

what makes the correction a drop-in addition to any existing replication package.

2.4 When is θ the right target?

Three outcome configurations arise in applications. *Binary* Y: the individual-level percentage change is undefined, and θ is the natural group-level parameter—the change in the complier probability relative to the complier base probability. *Non-negative* Y *with* $\Pr(Y = 0) > 0$ (patents, citations, income with unemployment): $\log(1 + Y)$ coefficients are not percentage effects and are sensitive to units (Chen and Roth, 2024). Chen and Roth's own IV application (Berkouwer–Dean) defines a 'Proportional LATE'—the LATE in levels over the control mean of instrument compliers—and estimates it with the identical $Y(1 - D)$ transformed-outcome 2SLS, citing Abadie (2002). They solve 'which population' in that application; our contribution is generality (denominator alignment beyond zeros) and the multi-/continuous-IV theory their footnotes 41 and 43 leave open. *Strictly positive* Y (employment of surviving firms): unit-level proportional effects exist, and Section 3 shows θ aggregates them with interpretable non-negative weights, whereas log-outcome IV estimates a different (log-scale) parameter for the same compliers; the two coincide only under homogeneity, and we recommend reporting both when feasible.

2.5 Relation to control complier means

In randomized experiments with one-sided noncompliance, the "control complier mean" (CCM) of Katz, Kling, and Liebman (2001) and Kling, Liebman, and Katz (2007)—used prominently by Chetty, Hendren, and Katz (2016) for Moving to Opportunity (MTO)—is computed as the mean outcome of treated compliers minus the treatment-on-treated estimate, a construction that exploits the absence of always-takers in the control group. CAPE nests the CCM and generalizes it: the transformed-outcome regression requires neither one-sidedness nor the absence of covariates, and extends to multiple treatment arms. In MTO, with mutually exclusive take-up indicators for the experimental and Section 8 vouchers, defining $\tilde{Y}_i = Y_i(1 - \text{TakeExp}_i - \text{TakeS8}_i)$ and instrumenting both take-up indicators with both assignment indicators recovers arm-specific CCMs from a single regression. We view MTO as evidence that a handful of top applied researchers already recognize the principle; our contribution is to make it operational in the quasi-experimental designs that dominate finance, where noncompliance is two-sided and instruments are often continuous.

3 Heterogeneity, Multiple Instruments, and Continuous Instruments

3.1 Heterogeneous effects with a single binary instrument

Proposition 1 (Weighted average and bounded range)**.** *Suppose $Y_i(0) > 0$ for all compliers, and let $\theta_i \equiv (Y_i(1) - Y_i(0))/Y_i(0)$. Then*

$$\theta \;=\; \sum_{i\in\mathcal{C}} w_i\,\theta_i, \qquad w_i = \frac{Y_i(0)}{\sum_{j\in\mathcal{C}} Y_j(0)} \;\geq\; 0, \qquad \sum_{i\in\mathcal{C}} w_i = 1,$$

where $\mathcal{C}$ is the set of compliers. Consequently, $\theta \in [\min_{i\in\mathcal{C}}\theta_i,\ \max_{i\in\mathcal{C}}\theta_i]$, and if $\theta_i = \bar{\theta}$ for all compliers, $\widehat{\theta} \xrightarrow{p} \bar{\theta}$.

The proof follows from writing the numerator of (2) as $\sum_{\mathcal{C}} Y_i(0)\theta_i/|\mathcal{C}|$. No analogous property holds when the denominator is a mean over a different population: the conventional ratio can lie outside the range of every unit-level proportional effect in the sample. For binary outcomes and outcomes with mass at zero, unit-level θ_i is ill-defined and θ should be read as the group-level parameter (2) directly.

3.2 Multiple binary instruments and weight matching

Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})' \in \{0,1\}^K$ with potential treatments $D_i(\mathbf{z})$. Assumptions 1–2 extend to the vector instrument, so $Y_i(\mathbf{D}, \mathbf{Z}) = Y_i(D_i)$. Assume the components $Z_{i1}, \ldots, Z_{iK}$ are mutually independent, jointly independent of potentials, with i.i.d. observations, and relevance ($\mathbb{E}[D_i(\mathbf{z}) - D_i(\mathbf{z}')] \neq 0$ for some $\mathbf{z}, \mathbf{z}'$).

Assumption 6 (Partial monotonicity)**.** *For each k and all i, $D_i(\mathbf{z})$ is weakly increasing in z_k holding $\mathbf{z}_{-k}$ fixed.*

Assumption 6 is the partial monotonicity condition of Mogstad, Torgovitsky, and Walters (2021), and the choice of monotonicity condition matters more with multiple instruments than is commonly appreciated. Condition 2 of Imbens and Angrist (1994) requires that for *every* pair of configurations $\mathbf{z}, \mathbf{w}$—including pairs differing in several components, such as $(1,0)$ versus $(0,1)$—either $D_i(\mathbf{z}) \geq D_i(\mathbf{w})$ for all i or the reverse for all i: all units must rank all instrument configurations identically. Mogstad, Torgovitsky, and Walters (2021) show that with more than one instrument this condition can hold only if choice behavior is effectively homogeneous. Partial monotonicity instead requires only that no single instrument can move any unit out of treatment holding the others fixed, while units remain

free to rank multi-coordinate changes differently. (The condition permits the direction of monotonicity to differ across components; we normalize every component to be increasing, re-signing instruments where necessary.) Define the one-coordinate contrasts $\Delta_{ik}(\mathbf{z}_{-k}) \equiv D_i(1, \mathbf{z}_{-k}) - D_i(0, \mathbf{z}_{-k}) \in \{0, 1\}$.

The unit-level identity behind (4) depends only on D_i being binary and the exclusion restriction: for any two configurations $\mathbf{z}, \mathbf{z}'$,

$$Y_i\big(D_i(\mathbf{z}')\big)\big(1 - D_i(\mathbf{z}')\big) - Y_i\big(D_i(\mathbf{z})\big)\big(1 - D_i(\mathbf{z})\big) = \begin{cases} -Y_i(0), & D_i(\mathbf{z}) = 0,\ D_i(\mathbf{z}') = 1, \\ Y_i(0), & D_i(\mathbf{z}) = 1,\ D_i(\mathbf{z}') = 0, \\ 0, & D_i(\mathbf{z}) = D_i(\mathbf{z}'), \end{cases} \tag{6}$$

with the middle (defiance) case ruled out by Assumption 6 for configurations differing in one component. For each component Z_k, mutual independence implies the component-specific Wald ratio

$$\frac{\mathrm{Cov}(Z_k, Y)}{\mathrm{Cov}(Z_k, D)} = \frac{\sum_{\mathbf{z}_{-k}} \Pr(\mathbf{Z}_{-k} = \mathbf{z}_{-k})\, \mathbb{E}\big[(Y_i(1) - Y_i(0))\Delta_{ik}(\mathbf{z}_{-k})\big]}{\sum_{\mathbf{z}_{-k}} \Pr(\mathbf{Z}_{-k} = \mathbf{z}_{-k})\, \mathbb{E}\big[\Delta_{ik}(\mathbf{z}_{-k})\big]},$$

a LATE for the units shifted by the kth instrument margin (the "compliers w.r.t. Z_k"). With $v_k = \mathrm{Var}(Z_k)$ and π_k the population first-stage coefficient on Z_k, 2SLS aggregates these ratios with weights $\lambda_k = \pi_k^2 v_k / \sum_\ell \pi_\ell^2 v_\ell \geq 0$, $\sum_k \lambda_k = 1$:

$$\hat{\beta}^{Y}_{\mathrm{2SLS}} \xrightarrow{p} \sum_{k=1}^{K} \lambda_k\, \mathbb{E}\big[\, Y_i(1) - Y_i(0) \mid \text{complier w.r.t. } Z_k \,\big]. \tag{7}$$

Mutual independence is what makes this decomposition positively weighted without further conditions. Mogstad, Torgovitsky, and Walters (2021) show that under partial monotonicity alone, 2SLS with *correlated* instruments—for instance, interactions constructed from a common component—can place negative weights on some margins, and they characterize verifiable conditions for non-negative weights in that case.

Proposition 2 (Weight matching)**.** *Under the assumptions of this subsection, 2SLS with outcome $Y(1 - D)$ and the same instruments satisfies*

$$\hat{\beta}^{Y(1-D)}_{2SLS} \xrightarrow{p} -\sum_{k=1}^{K} \lambda_k\, \mathbb{E}\big[\, Y_i(0) \mid \textit{complier w.r.t. } Z_k \,\big], \tag{8}$$

with weights λ_k identical *to those in (7), because the weights are determined entirely by the*

first stage, which the two regressions share. Consequently

$$\widehat{\theta} \xrightarrow{p} \frac{\sum_k \lambda_k \mathbb{E}\big[Y_i(1) - Y_i(0) \mid \text{complier w.r.t. } Z_k\big]}{\sum_k \lambda_k \mathbb{E}\big[Y_i(0) \mid \text{complier w.r.t. } Z_k\big]}, \tag{9}$$

a ratio of weighted averages sharing common non-negative weights. If the proportional effect is constant across complier subgroups, CAPE recovers it exactly; in general, CAPE lies within the range of subgroup-level proportional effects.

Proposition 2 is what licenses interpreting the ratio as a proportional effect in overidentified designs: numerator and denominator describe the same weighted population. Two remarks delimit the result. First, weight matching itself is mechanical and holds with full generality: for any dependent variable W, the 2SLS estimand is $\mathrm{Cov}(\mathbf{Z}_i'\boldsymbol{\pi}, W_i)/\mathrm{Cov}(\mathbf{Z}_i'\boldsymbol{\pi}, D_i)$, so whatever weighting over margins the first stage induces—including weighting schemes with negative components under correlated instruments, or under the non-saturated covariate specifications discussed in Section 3.4—applies identically to Y and to $Y(1-D)$. CAPE therefore always compares identically weighted aggregates, and is no more and no less credible than the 2SLS coefficient it rescales. Second, reading the ratio as a weighted-average proportional effect with the bounded-range property additionally requires the shared weights to be non-negative; mutual independence guarantees this, and correlated-instrument designs should verify it using the conditions of Mogstad, Torgovitsky, and Walters (2021). This second step, rather than complier-mean identification, is where the content of the extension lies: what must be established is not that each margin's untreated mean is identified, but that 2SLS aggregation preserves the pairing of numerator and denominator margins.

3.3 Continuous instruments

The same logic extends to a continuous instrument Z_i with support $\mathcal{Z} \subseteq \mathbb{R}$. Under independence, relevance, and monotonicity ($D_i(z)$ weakly increasing in z), each unit has a threshold z_i^* with $D_i(z) = \mathbf{1}(z \geq z_i^*)$; "compliers" become a continuum of margin-specific groups. By the Angrist and Imbens (1995) result, 2SLS identifies a weighted average of margin-specific effects with a weight function $\omega(z) \geq 0$ determined only by the first stage and the distribution of Z_i. Since the transformed-outcome regression shares that first stage, and its local effect at each margin is $-Y_i(0)$ by (6),

$$\widehat{\theta} \xrightarrow{p} \frac{\int_{\mathcal{Z}} \mathbb{E}\big[Y_i(1) - Y_i(0) \mid z_i^* = z\big]\, \omega(z)\, dz}{\int_{\mathcal{Z}} \mathbb{E}\big[Y_i(0) \mid z_i^* = z\big]\, \omega(z)\, dz}, \tag{10}$$

a weighted-average proportional effect with matched weights. Vectors of possibly continuous instruments reduce to the scalar effective instrument $\mathbf{Z}_i'\hat{\boldsymbol{\pi}}$ from the first stage. Weight matching continues to hold mechanically, because the numerator and denominator regressions share this index; interpreting the ratio as a positively weighted average over margins of the effective instrument, however, requires monotonicity of D_i in the index itself, which partial monotonicity alone does not deliver (Mogstad, Torgovitsky, and Walters, 2021). The vector case therefore shares the robustness—and the caveats—of the underlying 2SLS coefficient. Appendix A states the assumptions and derivations formally. The practical procedure never changes: (i) run 2SLS of Y on D with instruments $\mathbf{Z}$; (ii) run 2SLS of $Y(1-D)$ on D with the same instruments and specification; (iii) take the ratio (5).

The complier untreated mean has been recovered with a continuous instrument before, in the judge-leniency literature. Dahl, Kostøl, and Mogstad (2014, hereafter DKM) compute in Online Appendix B the potential DI participation of complier children had their parent been denied, combining compliance-type shares with conditional outcome means evaluated at the strict and lenient ends of the leniency distribution; Kleiner, Stoffman, and Yonker (2021) apply the same construction to report a peer-bankruptcy spillover relative to its complier base rate. Their formula is, after rearrangement, the transformed-outcome Wald ratio computed on an endpoint-binarized instrument: it is the identity (6) evaluated at two points of the support. The transformed-outcome 2SLS is the regression generalization: it uses all of the instrument's variation, requires no endpoint or percentile choice, accommodates the original design's covariates and clustering directly, and—by the weight-matching argument above—aggregates margin-specific untreated means with exactly the Angrist–Imbens (1995) weights of the published coefficient, rather than with weights defined by a chosen pair of endpoints.

To distinguish the two internally matched constructions, orient an endpoint pair (z_L, z_H) so that $\mathbb{E}[D \mid z_H] > \mathbb{E}[D \mid z_L]$. The endpoint complier base and endpoint Wald effect are

$$\mu_{0,EP} = \frac{\mathbb{E}[Y(1-D) \mid z_L] - \mathbb{E}[Y(1-D) \mid z_H]}{\mathbb{E}[D \mid z_H] - \mathbb{E}[D \mid z_L]}, \tag{11}$$

$$\tau_{EP} = \frac{\mathbb{E}[Y \mid z_H] - \mathbb{E}[Y \mid z_L]}{\mathbb{E}[D \mid z_H] - \mathbb{E}[D \mid z_L]}. \tag{12}$$

The internally consistent endpoint proportional effect is $\tau_{EP}/\mu_{0,EP}$. Dividing the full-support continuous-2SLS coefficient by $\mu_{0,EP}$ is instead a hybrid: it pairs a numerator and denominator that generally aggregate different complier margins. CAPE is the full-support analogue of the matched endpoint ratio, replacing both endpoint Wald contrasts with 2SLS coefficients from the same specification.

We now present a numerical example that isolates what weight matching adds alongside

the DKM endpoint construction. Let Z take values 0, $\frac{1}{2}$, and 1 with probabilities 0.05, 0.15, and 0.80, and let the population consist of two complier margins: 70 percent of units switch at threshold $\frac{1}{2}$, with untreated mean 0.05 and level effect 0.01 (a proportional effect of 20 percent), and 30 percent switch at threshold 1, with untreated mean 0.50 and level effect 0.15 (30 percent). Every unit's proportional effect is therefore either 20 or 30 percent. With the instrument entered linearly, 2SLS weights margin m by $\lambda_m \propto s_m \operatorname{Cov}(Z, \mathbf{1}\{Z \geq z_m\})$, where s_m is the margin's population share; here $\lambda = (0.505, 0.495)$, because the concentration of instrument mass at the top of the support raises the per-capita weight on the rare high-threshold margin by a factor of 2.3, almost exactly offsetting its small population share. The published coefficient is $0.505(0.01) + 0.495(0.15) = 0.079$, and the transformed-outcome regression delivers the identically weighted base $0.505(0.05) + 0.495(0.50) = 0.273$, so CAPE is 29.1 percent—inside $[20, 30]$, as the logic of Proposition 1 requires. The DKM endpoint construction averages the same two margins by population shares, giving a base of $0.70(0.05) + 0.30(0.50) = 0.185$. Paired with its endpoint-computed Wald numerator, $0.70(0.01) + 0.30(0.15) = 0.052$, this base gives 28.1 percent, again within range. Pairing the full-support 2SLS coefficient with the endpoint base instead gives $0.079/0.185 = 42.9$ percent because the numerator and denominator aggregate the two margins differently. Both underlying objects remain coherent for their respective targets: DKM describes the complier population defined by the selected endpoints, while CAPE supplies the denominator matched to the full-support 2SLS coefficient. A comparison at the 10th and 90th percentiles similarly defines a more specific endpoint population by omitting the tail margins. In DKM's application, the near-linear first stage and monotone estimated relationship make that endpoint construction well suited to the design. CAPE complements it by requiring no endpoint choice and by extending unchanged to multiple-instrument settings, such as Section 5.5, where no unique scalar pair of endpoints need be selected.

In the numerical example above, the weights follow from specializing Section 3.3: with $Y_i = Y_i(0) + \Delta_i D_i$, $D_i = \mathbf{1}\{Z_i \geq z_i^*\}$, and Z_i independent of potentials, $\operatorname{Cov}(Z, Y) = \sum_m s_m \Delta_m c_m$ and $\operatorname{Cov}(Z, D) = \sum_m s_m c_m$ with $c_m = \operatorname{Cov}(Z, \mathbf{1}\{Z \geq z_m\}) \geq 0$, so $\lambda_m = s_m c_m / \sum_\ell s_\ell c_\ell$. In the example, $c_1 = \mathbb{E}[Z\,\mathbf{1}\{Z \geq \frac{1}{2}\}] - E[Z]\Pr(Z \geq \frac{1}{2}) = 0.875 \times 0.05 = 0.0438$ and $c_2 = 0.80 - 0.875 \times 0.80 = 0.100$, giving $\lambda = (0.505, 0.495)$ against population shares $(0.70, 0.30)$. The transformed-outcome regression replaces Δ_m with $-\mathbb{E}[Y(0) \mid m]$ and leaves c_m, hence λ_m, unchanged.

In Section 5.10, we compute CAPE and the DKM estimators using data from two published papers with continuous instruments, and show that substantial differences also obtain in real-world data.

3.4 Covariates: a caveat and recommendations

Every replication in Section 5 includes covariates or fixed effects, as does essentially every published IV design. Blandhol, Bonney, Mogstad, and Torgovitsky (2022) show that with covariates, 2SLS retains a positively weighted LATE interpretation only under rich ("saturated") specifications; otherwise it may place negative weights on some margins. This caveat applies equally to the numerator and denominator of CAPE—both inherit whatever weighting the original specification induces, and the weight-matching of Proposition 2 continues to hold because both regressions share the identical first stage. In this sense CAPE is exactly as robust, and exactly as fragile, as the published 2SLS coefficient it rescales. We nonetheless recommend that researchers probe sensitivity with interacted or more saturated specifications where feasible, and we do so in our replications where the original papers report such variants.

3.5 Why not Poisson IV?

For count outcomes, an alternative is IV-Poisson or control-function PPML, which targets a constant-proportional-effect parameter under a multiplicative structural model. CAPE differs in two respects. First, it imposes no functional form: it stays entirely within the nonparametric LATE framework and requires nothing beyond the two linear IV regressions the researcher has already coded. Second, its estimand is the explicitly local parameter (2), matched to the population the design identifies, rather than a model-based proportional effect whose interpretation under heterogeneity and endogenous selection into compliance requires further assumptions. We view the approaches as complements; when a multiplicative model is credible, reporting both is informative.

3.6 Inference

$\widehat{\theta}$ is a ratio of two 2SLS coefficients estimated on the same sample with identical regressors and instruments; the regressions differ only in their dependent variable. Three procedures are available. When the complier base is well separated from zero, our default is a design-matched paired bootstrap; the joint delta method is a computationally convenient alternative. When the base is marginal, use Fieller confidence sets. All three require the *joint* sampling distribution of the two coefficients. Because $Y(1-D)$ coincides with Y for every untreated observation, the two estimates are strongly and mechanically correlated, and any procedure that treats the regressions as independent—separate bootstraps, or a ratio variance without the covariance term—is invalid.

The paired bootstrap covers the entire two-regression procedure: resample clusters to match the original design's clustering, re-estimate both regressions on each draw, and report the standard deviation or percentile interval of the ratio across 1,000 replications (Appendix B gives the algorithm). The joint delta method uses the covariance in closed form. Let $\hat{\mu}_C \equiv -\hat{\beta}^{Y(1-D)}_{\text{2SLS}}$ denote the estimated complier base rate, so that $\widehat{\theta} = \hat{\beta}^{Y}_{\text{2SLS}}/\hat{\mu}_C$, and let $\widehat{V}_{\beta\beta}$, $\widehat{V}_{\mu\mu}$, and $\widehat{V}_{\beta\mu}$ denote cluster-robust estimates of the asymptotic variances and covariance of $(\hat{\beta}^{Y}_{\text{2SLS}}, \hat{\mu}_C)$, obtained in a single step by stacking the two regressions into one system (Appendix B). Then

$$\widehat{\text{Var}}\big(\widehat{\theta}\big) \;=\; \frac{1}{\hat{\mu}_C^2}\Big(\widehat{V}_{\beta\beta} - 2\widehat{\theta}\,\widehat{V}_{\beta\mu} + \widehat{\theta}^2\,\widehat{V}_{\mu\mu}\Big). \tag{13}$$

Both the delta method and percentile-bootstrap intervals are reliable only when the complier base rate is well separated from zero. As the t-ratio $\hat{\mu}_C/\widehat{V}_{\mu\mu}^{1/2}$ falls, the sampling distribution of a ratio departs sharply from normality, and the delta method undercovers, while the percentile bootstrap maintains coverage only by producing uninformatively wide intervals.

The construction of Fieller (1954) remains valid in exactly those cases. Since $\theta = \beta^Y/\mu_C$ if and only if $\beta^Y - \theta\mu_C = 0$, a level-$(1-\alpha)$ confidence set collects every θ_0 not rejected by a t-test of that linear restriction:

$$\text{CS}_{1-\alpha} \;=\; \Big\{\theta_0 : \big(\hat{\beta}^{Y}_{\text{2SLS}} - \theta_0\,\hat{\mu}_C\big)^2 \;\le\; z^2_{1-\alpha/2}\Big(\widehat{V}_{\beta\beta} - 2\theta_0\,\widehat{V}_{\beta\mu} + \theta_0^2\,\widehat{V}_{\mu\mu}\Big)\Big\}, \tag{14}$$

where $z_{1-\alpha/2}$ is the standard normal critical value. This is a quadratic inequality in θ_0, solved in closed form in Appendix B. The set is a bounded interval exactly when $\hat{\mu}_C^2 > z^2_{1-\alpha/2}\widehat{V}_{\mu\mu}$—that is, when the base rate is statistically distinguishable from zero at level α—and is otherwise unbounded, which is the honest statement that the data cannot bound the proportional effect at that margin. The logic is that of weak-instrument-robust inference (Staiger and Stock, 1997), with the complier base rate in the role of the first stage. Hirschberg and Lye (2010) show that delta and Fieller intervals nearly coincide once the denominator's t-ratio exceeds roughly three, and diverge quickly below it.

Whatever interval accompanies CAPE, we recommend always reporting $\hat{\mu}_C$ itself, with its own standard error. A complier base rate that is small relative to its sampling error is a diagnostic in its own right: it says the data do not support *any* proportional-effect interpretation at that margin, adjusted or conventional. We note that this reporting standard exceeds current practice, in which published proportional effects almost never carry a standard error at all—the base rate is implicitly treated as a known constant even though it is estimated. Equations (13) and (14) apply verbatim to conventional ratios, so attaching an interval to the headline percentage is an improvement whether or not one adopts the complier-aligned denominator. Section 5.9 illustrates with sparse outcomes for which the conventional ratio

manufactures precise-looking percentages that this framework correctly flags as fragile.

4 The Prevalence of CAPE-Relevant Designs

To assess how often CAPE can be applied in curent empirical research, we need to know how often an IV or fuzzy-RDD design is used to identify the effect of a binary treatment. We start by surveying research articles published during 2015–2024 in the top three finance and top five economics journals. After excluding replies, comments, corrections, and editorials, the sample contains 6,717 articles, including 4,291 finance articles. We classify a paper as causal if it makes an explicit causal claim and uses an identification design intended to address endogeneity; OLS regressions do not qualify. A review of each article's first three pages identified 3,188 candidate papers, including 2,066 finance papers, after which a full-text review recorded each paper's distinct causal settings, treatments, and outcomes. We manually checked 148 papers, including 69 finance papers, against the source PDFs because descriptions of their methods, instruments, treatments, or compliance patterns were internally inconsistent; because settings were missing or malformed; or because a paper appeared to contain an IV or fuzzy-RDD design that the recorded method had missed. Correcting these cases leaves 3,164 papers with at least one usable causal setting, including 2,045 finance papers, which form the cohort reported in Table 1.

Of these, 946 papers, including 564 finance papers, contain an IV or fuzzy-RDD setting, and 273 papers, including 152 finance papers, use a binary treatment in the same setting. Among these 152 finance papers, 137 (90.1%) also study what we call a priority outcome in the same causal setting: a binary variable, rate or share, naturally bounded variable, count, or outcome with mass at zero. CAPE is particularly useful in these cases. Researchers often want to interpret effects on these outcomes in percentage terms, yet conventional scaling can use a base rate from the wrong population, while outcomes with zeros cannot be logged without changing the estimand.

The IV-plus-binary columns require both conditions to hold within the same causal setting: an excluded instrument or assignment must identify the effect of a binary treatment through a 2SLS or Wald design. This includes fuzzy RDD and binary, multiple, or continuous instruments. Direct-treatment RCTs, difference-in-differences designs, event studies, and sharp RDDs do not enter these columns unless they contain a separate qualifying IV setting; neither does an IV setting for a different, continuous treatment. Each paper is counted once per column.

Table 1: Causal-Inference Papers and the Design Scope of CAPE

Counts are distinct papers published during 2015–2024. *Causal* requires at least one usable causal setting under the definition in the text. *+ Finance* retains all papers in the three finance journals and finance-related papers in the five economics journals. *+ IV Setting* requires an IV or fuzzy-RDD setting, while *+ Binary Treatment* requires a binary-treatment setting. *+ Finance + Binary Treatment* combines the finance and binary-treatment conditions. *+ IV Setting + Binary Treatment* requires both conditions in the same setting. The *+ Finance + IV Setting + Binary Treatment* design-scope column also imposes finance. Each paper is counted once per column. The two IV-plus-binary design-scope columns identify papers whose design and treatment permit CAPE; they do not show whether a paper reports a proportional effect or uses a misaligned denominator. *Focal 2SLS magnitude comparison* identifies papers that compare at least one focal 2SLS, Wald, LATE, TOT, or CACE estimate with a mean, rate, or standard deviation of the same outcome. This includes comparisons without division and a standard outcome mean or standard deviation paired with the focal estimate in the same table even without prose discussion. *Correct complier population* identifies papers with at least one such comparison using the untreated/control-complier outcome level or rate, or the appropriate complier-outcome standard deviation. The final four columns report audit classifications.

								Focal 2SLS magnitude comparison		Correct complier population	
	Causal	+ Finance	+ IV Setting	+ Binary Treatment	+ Finance + Binary Treatment	+ IV Setting + Binary Treatment	+ Finance + IV Setting + Binary Treatment	All	Finance	All	Finance
The Journal of Finance	311	311	88	180	180	22	22	7	7	0	0
Journal of Financial Economics	749	749	187	481	481	50	50	14	14	1	1
The Review of Financial Studies	573	573	154	361	361	50	50	25	25	0	0
Finance total	1,633	1,633	429	1,022	1,022	122	122	46	46	1	1
American Economic Review	579	138	179	360	81	52	15	34	4	2	0
The Quarterly Journal of Economics	275	68	101	184	41	36	9	19	1	7	0
Journal of Political Economy	283	93	84	178	50	32	3	18	2	4	0
Econometrica	131	35	43	81	24	15	1	8	0	1	0
The Review of Economic Studies	263	78	110	126	38	16	2	8	1	2	0
Economics total	1,531	412	517	929	234	151	30	87	8	16	0
All journals	3,164	2,045	946	1,951	1,256	273	152	133	54	17	1

Note that these counts measure potential reach, not the prevalence or size of the base rate error. CAPE is useful in all contexts where an IV estimate can be interpreted relative to an outcome base rate, so we report all such cases. The survey does not assume that every qualifying paper makes such a comparison or uses a misaligned denominator. That requires a further audit, which we present below.

This audit shows that among the 273 binary-treatment papers in our survey (including 152 finance papers), 133 papers (54 in finance) compare a focal 2SLS coefficient to a base rate or an outcome standard deviation. Only 17 papers, including only one finance paper, use a benchmark aligned with the relevant complier population. These counts include comparisons without explicit scaling and cases in which a standard table mean or outcome standard deviation is paired with the focal estimate without any adjustments to ensure that these statistics come from complier-only samples.

Next, Table 2 shows that single-continuous or multiple IV cases account for 131 of the 380 eligible settings (34.5%). The continuous-instrument and multiple-IV extensions which CAPE permits therefore incrementally cover over one-third of the binary-treatment IV settings in the top finance and economics journals during our sample.

Table 2: Instrument Structure in Binary-Treatment IV Settings

The sample comprises 380 binary-treatment IV or fuzzy-RDD settings from 273 research articles (152 finance) published during 2015–2024 in the top three finance and top five economics journals. A *single binary* setting uses exactly one binary excluded instrument. A *multiple binary* setting uses at least two binary excluded instruments and no nonbinary instrument. A *single continuous* setting uses exactly one scalar nonbinary excluded instrument—including a price, propensity, distance, share, count, or scalar interaction—and no binary instrument. *Other* includes multiple-nonbinary, mixed binary–nonbinary, multidimensional, bundled-alternative, and insufficiently described instrument sets. Setting categories are mutually exclusive. The paper column counts distinct articles with at least one setting of each type; it may overlap across rows when an article contains settings of different types. Classifications use the corrected causal-setting extractions, with source checks where instrument support or multiplicity was ambiguous.

Instrument structure	Settings	Papers with at least one such setting
Single binary	185 (48.7%)	140 (51.3%)
Multiple binary	15 (3.9%)	13 (4.8%)
Single continuous	116 (30.5%)	80 (29.3%)
Other	64 (16.8%)	50 (18.3%)
Total	380	273

5 Does It Matter? Evidence from Nine Replications

This section digs deeper into contexts where CAPE provides the correct percentage effect. To do so, we first search for all top-3/top-5 published papers in our sample where we have all data required to replicate the paper's causal estimate and the base rate publicly available. We then replicate the key estimates in each paper and base rates used to scale the estimate (if any), as well as complier-only base rates relevant to CAPE to examine how much of a difference CAPE can make in real-world applications.

Table 3 summarizes nine papers, including four finance papers, through fourteen representative settings. Panel A contains ten non-log settings—seven binary outcomes, two continuous cumulative-growth outcomes, and one share outcome—while Panel B contains four log-transformed outcomes; together they span single binary, multiple binary, and continuous instruments. CAPE can be several times the conventional ratio or less than half of it; among settings compared on the same outcome scale, denominator corrections range from 0.6 to 248.7 percentage points in both directions. The log-versus-level settings make a distinct point: when zeros make a log coefficient unit-dependent, CAPE supplies a percentage effect on the raw outcome. The case studies that follow are organized by lesson, beginning with the finance applications, and each closes with the complier economics that explains the direction of the revision. Throughout, the original papers' level estimates and qualitative conclusions are untouched; our exercise concerns only the translation of levels into percentages.

All replicated specifications are variants of one estimating system:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + g(W_i) + \varepsilon_i, \\ D_i &= \gamma_0 + \boldsymbol{\gamma}_1' \mathbf{Z}_i + h(W_i) + u_i, \end{aligned} \tag{15}$$

where D_i is a binary treatment, $\mathbf{Z}_i$ collects excluded instruments (binary, multiple binary, or continuous; in fuzzy RDD, the above-cutoff indicator with g, h containing running-variable controls), and W_i collects covariates and fixed effects. The conventional proportional effect divides $\hat{\beta}_1$ by $\widehat{\mathbb{E}(Y)}$, $\widehat{\mathbb{E}(Y(0))}$, or $\hat{\beta}_0$; for $\log(1+Y)$ outcomes it is $e^{\hat{\beta}_1} - 1$. CAPE re-estimates (15) with dependent variable $Y_i(1 - D_i)$ and scales $\hat{\beta}_1$ by minus the resulting treatment coefficient.

Table 3: Published Estimates, Replications, and Proportional Effects

This table reports fourteen representative settings from nine papers, including four finance papers. Panel A contains ten non-log outcomes; Panel B contains four $\log(1+Y)$ outcomes. Paper 2SLS is the published coefficient; Rep. 2SLS is our estimate on that scale, with our standard error beneath where available. Panel B adds Raw 2SLS and Log-implied, $\exp(\hat{\beta}) - 1$. Simple scales the relevant raw coefficient by the conventional base described in each application; CAPE uses the estimated complier untreated mean. Gaps compare Simple with CAPE in Panel A and Log-implied with CAPE in Panel B. For the Farre-Mensa et al. growth rows, b is a change of $100b$ percentage points. CAPE uncertainty follows Section 3.6: parentheses are 1,000-resample joint-bootstrap standard errors for strong bases; brackets are 95% Fieller intervals from the joint-bootstrap covariance when bases are marginal or ratio draws unstable; and "no claim" is an all-real Fieller set. Sampat–Williams replicated 2SLS standard errors are not estimated, but its common first stage cancels from CAPE, so its intervals jointly resample 1,545 patent-application clusters. The daggered Acharya–Xu entry is the unconstrained 95% percentile-bootstrap interval; no draw is capped or trimmed, but its Fieller set is all real, so it is descriptive rather than a bounded proportional claim. The reconstructed NETS rows lack CAPE inference. Interval endpoints are rounded to percentage points. IV classifies the instrument set; blank paper fields denote another setting from the same paper and panel. Citation counts are from Google Scholar in August 2026.

Panel A: Non-log outcomes

					2SLS coefficient		Proportional change (%)		Gap versus CAPE		
Paper	#Cit.	Outcome	IV	Journal	Paper	Rep.	Simple	CAPE	Type	Gap (p.p.)	Rel. gap (%)
Solis (2017)	378	binary	single binary	JPE	0.155	0.1544 (0.0077)	48.79	57.30 (3.62)	Under	8.51	14.85
Farre-Mensa et al. (2020)	468	continuous	single continuous	JF	0.545	0.5224 (0.2969)	115.84	175.70	Under	59.85	34.07
		continuous	single continuous		0.795	0.7779 (0.4517)	103.23	144.29	Under	41.06	28.45
Stein and Yannelis (2020)	177	binary	single binary	RFS	0.0504	0.0513 (0.0155)	14.08	14.68 (6.11)	Under	0.61	4.14
Esposito et al. (2023)	122	binary	single binary	AER	0.076	0.0763 (0.0122)	384.99	168.42 (27.37)	Over	216.56	128.58
Angelucci et al. (2022)	90	binary	multiple binary	AER	0.622	0.6216 (0.1754)	475.45	226.77 [63, 1,913]	Over	248.68	109.66
Berry et al. (2020)	365	binary	single continuous	JPE	−0.072	−0.0720 (0.0350)	−49.50	−39.81 (16.04)	Over	9.69	24.33
Sampat and Williams (2019)	576	binary	single continuous	AER	−0.0141	−0.0141	−15.38	−23.47 [−60, 40]	Under	8.09	34.47
Asher and Novosad (2020)	952	binary	single binary	AER	0.129	0.1294 (0.0554)	109.87	152.51 [14, 20,098]	Under	42.65	27.96
		share	single binary		−0.092	−0.0917 (0.0434)	−19.27	−16.90 (7.22)	Over	2.37	14.05

Panel B: Log outcomes

					2SLS coefficient			Proportional change (%)			Gap versus CAPE		
Paper	#Cit.	Outcome	IV	Journal	Paper	Rep.	Raw	Log-implied	Simple	CAPE	Type	Gap (p.p.)	Rel. gap (%)
Farre-Mensa et al. (2020)	468	log(1+Y)	single continuous	JF	0.472	0.4716 (0.0493)	11.326	60.26	525.88	431.81 [177, 1,591]	Under	371.55	86.04
Stein and Yannelis (2020)	177	log(1+Y)	single binary	RFS	0.0362	0.0374 (0.0091)	0.4471	3.81	8.98	9.98 (2.90)	Under	6.17	61.83
Acharya and Xu (2017)	858	log(1+Y)	single binary	JFE	−0.3863	−0.3811 (0.1687)	−0.6810	−31.69	−112.59	−94.82 [−118, −62]†	Under	63.13	66.58
Sampat and Williams (2019)	576	log(1+Y)	single continuous	AER	−0.0230	−0.0230	−0.0571	−2.27	−39.64	−45.18 [−80, −9]	Under	42.91	94.98

5.1 Underestimation from richer always-takers: Solis (2017)

Solis (2017) studies the effect of credit access on college enrollment. College loans in Chile become available to students whose admission scores exceed 475; because students can retake the test, scoring below 475 on the first attempt does not preclude eventual eligibility, generating a fuzzy RDD. Using the first-attempt score as an IV for ever-eligibility, the paper estimates

$$\begin{aligned} \text{Elig}_i &= \gamma_0 + \gamma_1 \cdot \mathbf{1}(T_i \geq \tau) + f(T_i - \tau) + \eta_i, \\ Y_i &= \beta_0 + \beta_1 \cdot \text{Elig}_i + f(T_i - \tau) + \nu_i, \end{aligned} \tag{16}$$

with f piecewise linear. The baseline estimate $\hat{\beta}_1 = 0.155$ implies enrollment rises by 15.5 percentage points; the paper scales by the regression constant, $\hat{\beta}_1/\hat{\beta}_0 = 48.8\%$, and reports a relative increase of about 50 percent.[3]

The base rate $\hat{\beta}_0$ reflects the untreated enrollment probability of compliers *and* always-takers (there are no never-takers: students above 475 are eligible by definition). That scaling is valid only if always-takers' untreated enrollment mirrors compliers'. But Solis reports that test retakers—who predominantly overlap with always-takers—have better-educated parents and come disproportionately from higher-income quintiles. Their untreated enrollment probability should therefore be *higher* than compliers', making $\hat{\beta}_0$ too high and the reported 48.8% too low. Applying CAPE with the paper's functional form and bandwidth (Table 4), the complier base rate is 0.2695 versus the constant's 0.3165, and the proportional effect rises to 57.3%—a 17.4 percent upward revision. Despite compliers constituting 86.6% of the sample by the first stage, the always-taker adjustment is substantial. The same-IV income-quintile diagnostic in Table 5 shows only modest observable differences: 53.8% of compliers, versus 51.9% of the sample, are in the first quintile, while each remaining share is 0.5–0.8 percentage points lower. These coarse categories therefore point in the expected direction but cannot by themselves explain the larger base-rate difference.[4]

[3] Solis (2017), Table 4, Column 1, and the accompanying discussion on p. 582, which describes a 16 percentage point increase representing a relative increment of 50 percent.

[4] The available dataset includes no demographic variables beyond income quintiles, limiting further characterization.

Table 4: Ever College Enrollment and Ever Loan Eligibility in Solis (2017)

RD Coef. is $\hat{\beta}_1$ from Equation (16), where the dependent variable indicates ever enrolling in college during the sample period. Base rate for the simple ratio is the constant $\hat{\beta}_0$. Base rate for CAPE sets the dependent variable to $Y_i(\text{Elig}_i - 1)$, maintaining the original functional form and bandwidth. Ratio is RD Coef. divided by the respective base rate.

	RD Coef.	Base Rate	Ratio (%)
Simple	0.1544	0.3165	48.79
CAPE	0.1544	0.2695	57.30

Table 5: Income Quintiles for Compliers in Solis (2017)

Probability of belonging to each income quintile in the IV sample and among same-IV compliers (students eligible if and only if the first-year score clears 475). For each indicator X, its complier mean is minus the eligibility coefficient from the exact Solis 2SLS after replacing enrollment with $X(1 - \text{Elig})$; the sample, 44-point bandwidth, piecewise-linear score controls, treatment, and cutoff instrument are unchanged. The fifth quintile is excluded because the estimation sample contains only the first four; both displayed sets of shares sum to one.

	$\mathbf{1}(q_1 = 1)$	$\mathbf{1}(q_2 = 1)$	$\mathbf{1}(q_3 = 1)$	$\mathbf{1}(q_4 = 1)$
Sample Mean	0.5194	0.2178	0.1451	0.1177
Compliers Mean	0.5381	0.2103	0.1387	0.1129

5.2 Growth and count outcomes at a local margin: Farre-Mensa, Hegde, and Ljungqvist (2020)

Farre-Mensa et al. (2020) estimate the private value of a startup's first patent, exploiting quasi-random assignment of applications to USPTO examiners of differing leniency within art-unit-by-application-year cells. Their abstract describes startups winning the patent "lottery" as having "55% higher employment growth and 80% higher sales growth five years later." The empirical results and Table IV give the precise interpretation: patent approval raises five-year employment growth by 54.5 percentage points and sales growth by 79.5 percentage points. Because the outcomes are growth rates, the abstract's percent wording conflates a percentage-point change with a relative percentage change, which requires a denominator. The paper itself emphasizes that leniency identifies effects for "middling" applications—an explicitly local margin. If those compliers have low untreated growth, broad untreated means understate proportional effects at the relevant margin; and for count outcomes with many zeros (follow-on patents, citations), $\log(1 + Y)$ magnitudes are not

percentage effects.

Table 6 reports CAPE for the NETS growth outcomes (original Table IV) and non-NETS outcomes (original Tables V, VII, and VIII). The public archive omits the original patent-to-NETS link; we reconstruct it by fuzzy name/location matching, so Table IV coefficients differ slightly from the original (e.g., five-year employment growth 0.522 versus 0.545). Three findings stand out. First, growth effects at the leniency margin are larger than our constructed Raw-simple benchmarks: five-year employment growth is 175.7% of the complier base (0.297), compared with 115.8% using the broad untreated mean, and five-year sales growth is 144.3%, compared with 103.2%. These denominator-scaled ratios are distinct from the original 54.5- and 79.5-percentage-point coefficients; the abstract rounds those coefficients to 55 and 80 but labels them with percent signs, even though they are not relative changes computed against a base rate. Second, for the headline citations outcome, the published log-implied 60.3% and the CAPE of 431.8% measure different parameters: the level LATE (11.3 citations) is 4.3 times the complier base of 2.62 citations. Third, none of this weakens the paper's conclusions; it implies that the marginal applications whose fate examiners decide come from startups with relatively low untreated growth and follow-on innovation—which is itself economically interesting, and consistent with the paper's own emphasis on the "middling" margin.

Table 6: Patent Grants, Follow-on Innovation, and Financing: Farre-Mensa, Hegde, and Ljungqvist (2020)

CAPE calculations for Farre-Mensa, Hegde, and Ljungqvist (2020). The treatment is approval of the startup's first patent application, instrumented by examiner leniency; specifications include art-unit-by-application-year and headquarters-state fixed effects, clustered by art unit. Table IV rows use our reconstruction of the NETS match; small coefficient differences from the original reflect data and fuzzy-matching differences. Type is the dependent-variable scale; log1p denotes $\log(1+Y)$. Paper 2SLS is the original coefficient and Rep. 2SLS is our estimate on that scale. For growth rows, a coefficient b is a change of $100b$ percentage points in cumulative growth; thus Paper 2SLS values 0.545 and 0.795 equal 54.5 and 79.5 percentage points. Raw 2SLS displays the separately estimated raw-outcome coefficient for log1p rows; for other rows Rep. 2SLS is already on the raw scale. Log-implied is $\exp(\hat{\beta}) - 1$ and applies only to log1p rows. Raw simple divides the raw coefficient by the untreated mean in the IV sample. CAPE base is $\mathbb{E}(Y(0) \mid \text{complier})$, and CAPE divides the same raw coefficient by that base. Gap is the absolute difference between CAPE and Log-implied for log1p rows and between CAPE and Raw simple otherwise; "–" denotes a non-applicable entry.

Tbl	Sample	Outcome	Type	Paper 2SLS	Rep. 2SLS	Raw 2SLS	Log-implied	Raw simple	CAPE base	CAPE	Gap (p.p.)
IV	NETS	Emp growth 1y	growth	0.061	0.079	–	–	67.5	0.072	108.9	41.4
IV	NETS	Emp growth 2y	growth	0.228	0.182	–	–	82.5	0.124	146.7	64.2
IV	NETS	Emp growth 3y	growth	0.333	0.307	–	–	98.4	0.134	229.7	131.3
IV	NETS	Emp growth 4y	growth	0.489	0.422	–	–	117.5	0.176	240.4	122.9
IV	NETS	Emp growth 5y	growth	0.545	0.522	–	–	115.8	0.297	175.7	59.9
IV	NETS	Sales growth 1y	growth	0.096	0.108	–	–	76.5	0.089	121.0	44.5
IV	NETS	Sales growth 2y	growth	0.276	0.209	–	–	74.2	0.189	110.6	36.4
IV	NETS	Sales growth 3y	growth	0.512	0.499	–	–	116.3	0.190	262.2	145.9
IV	NETS	Sales growth 4y	growth	0.796	0.555	–	–	101.2	0.407	136.1	34.9
IV	NETS	Sales growth 5y	growth	0.795	0.778	–	–	103.2	0.539	144.3	41.1
V	All	Applications	log1p	0.448	0.448	2.748	56.5	233.8	1.009	272.4	215.9
V	All	Grants	log1p	0.353	0.353	1.564	42.4	322.6	0.436	358.6	316.2
V	All	Approval rate	rate	0.244	0.244	–	–	51.1	0.468	52.2	1.1
V	All	Citations	log1p	0.472	0.472	11.326	60.3	525.9	2.623	431.8	371.5
V	All	Cites/patent	log1p	0.285	0.285	1.071	32.9	69.1	1.511	70.9	38.0
VII	All	VC 1y	binary	0.021	0.021	–	–	65.0	0.025	83.1	18.1
VII	All	VC 2y	binary	0.029	0.029	–	–	64.4	0.041	71.1	6.8
VII	All	VC 3y	binary	0.031	0.031	–	–	60.9	0.047	66.1	5.2
VII	All	VC 4y	binary	0.034	0.034	–	–	63.6	0.046	74.7	11.1
VII	All	VC 5y	binary	0.036	0.036	–	–	65.3	0.047	76.8	11.5
VII	All	IPO	binary	0.010	0.010	–	–	174.6	0.004	246.8	72.2
VIII	All	Collateral 1y	binary	0.004	0.004	–	–	41.9	0.011	34.4	7.5
VIII	All	Collateral 2y	binary	0.018	0.018	–	–	111.7	0.016	107.7	4.0
VIII	All	Collateral 3y	binary	0.029	0.029	–	–	140.3	0.021	137.0	3.3
VIII	All	Collateral 4y	binary	0.039	0.039	–	–	165.3	0.022	180.4	15.0
VIII	All	Collateral 5y	binary	0.058	0.058	–	–	223.5	0.021	281.1	57.6
VIII	All	Collateral any	binary	0.086	0.086	–	–	292.0	0.024	358.4	66.4
VIII	No VC	Collateral 1y	binary	−0.002	−0.002	–	–	−28.3	0.010	−18.4	9.8
VIII	No VC	Collateral 2y	binary	0.011	0.011	–	–	90.3	0.013	82.2	8.1
VIII	No VC	Collateral 3y	binary	0.019	0.019	–	–	118.7	0.015	124.6	5.9
VIII	No VC	Collateral 4y	binary	0.026	0.026	–	–	147.1	0.016	162.8	15.7
VIII	No VC	Collateral 5y	binary	0.044	0.044	–	–	226.1	0.016	269.1	43.0
VIII	No VC	Collateral any	binary	0.059	0.059	–	–	271.9	0.020	303.7	31.9

5.3 Unit dependence and a validation: Stein and Yannelis (2020)

Stein and Yannelis (2020) study the effects of access to the Freedman's Savings Bank on human capital, wealth, and labor outcomes among Southern Blacks in the 1860s, instrumenting account ownership with a branch-in-county indicator.[5] The first stage implies a complier share of 17.1%, so complier/non-complier differences could in principle matter greatly.

Table 7 delivers two lessons. First, a validation: for the binary outcome "Works," the conventional ratio is 14.08% and CAPE is 14.68%. Despite the small complier share, compliers and non-compliers evidently have comparable untreated labor-force participation, and the published magnitude survives the audit essentially intact. CAPE is a diagnostic, not an automatic revision. Second, unit dependence: for income, the paper's $\log(1 + \text{Occscore})$ coefficient implies 3.81%; but because Occscore has mass at zero (individuals without income data are coded as zero), this number depends on measurement units. Were Occscore measured in thousands rather than hundreds of 1950 dollars, the implied effect would be 2.12%; in tens of thousands, 0.4%. On the raw scale, the conventional ratio is 8.98% and CAPE is 9.98%, so outcome scaling accounts for most of the contrast with 3.81%; both raw-scale ratios are invariant to units.

Table 7: Account Holding and Labor Market Outcomes in Stein and Yannelis (2020)

IV estimates using Branch in County as the only instrument for having an account. Works indicates labor-force participation. Occscore is the IPUMS-assigned median income of the person's occupation in 1950 (hundreds of 1950 dollars); Income is $\ln(1 + \text{Occscore})$. Paper 2SLS is the published estimate and Rep. 2SLS is our estimate on the same outcome scale. Log-implied is $\exp(\hat{\beta}^{2SLS}) - 1$ for Income. Raw simple divides the Works or raw-Occscore coefficient by its model-implied conventional untreated base; CAPE divides the same raw coefficient by the estimated complier untreated mean. All regressions follow the original controls and fixed effects.

Statistic	Works	Income	Raw Occscore
Paper 2SLS	0.0504	0.0362	–
Rep. 2SLS	0.0513	0.0374	0.4471
Log-implied ratio	–	3.81%	–
Raw simple ratio	14.08%	–	8.98%
CAPE	14.68%	–	9.98%

[5] These estimates correspond to Stein and Yannelis (2020) Table A6, Row 4, where Branch in County is the sole instrument. Their baseline also includes Branch Distance; as Celerier and Tak (2022) note, the paper appears to use modern county boundaries, which affects distance calculations, so we focus on the branch-in-county instrument. Our summary statistics for Branch Distance align closely with Celerier and Tak (2022).

5.4 Overestimation from mechanically high-base-rate compliers: Esposito et al. (2023)

Esposito et al. (2023) study how screenings of *The Birth of a Nation* shifted attitudes toward national reunification, using navy enlistment as a behavioral outcome and instrumenting county-level screenings with screenings of *The Million Dollar Mystery*, a politically neutral film with a similar diffusion technology. Exposure raises the enlistment indicator by 7.6 percentage points, which the paper compares with a sample mean of about 3.39 percent, suggesting an effect of 223.9%.

The appropriate base rate is the untreated enlistment probability of compliers—county-months whose screening status is moved by the instrument. Compliers are only 37.6% of the IV sample, and the paper's own heterogeneity analysis shows they are more urban. Because the outcome is an indicator for *at least one* enlistment in a county-month, larger and more urban counties mechanically have higher base rates absent treatment, so $\mathbb{E}(Y(0) \mid \text{complier})$ should exceed broad benchmarks. Table 8 replicates the paper's 2SLS estimates and computes three ratios. In the baseline, the complier base rate is 0.0453, versus 0.0341 for the sample mean and 0.0198 for the estimated untreated mean; CAPE is 168.4%, which is 216.6 points below the simple ratio and 55.5 points below the sample-mean ratio, and the pattern holds across specifications. The same-IV diagnostics in Table 9 show substantially larger and more urban compliers in every specification, consistent with the two films spreading through the same theater network. Compliers also have lower illiteracy in every column, although the difference narrows with state-by-month fixed effects; the corresponding Black-population comparison changes sign and is not robust. The stable size and urbanity differences support the mechanical base-rate channel and show why the CAPE adjustment matters.

Table 8: *The Birth of a Nation* and Enlistment in Esposito et al. (2023)

This table replicates the 2SLS estimates from Table 4 in Esposito et al. (2023). The dependent variable indicates at least one navy enlistment in county c in month-year t; the treatment indicates a screening of *The Birth of a Nation*, instrumented by a screening of *The Million Dollar Mystery*. Columns 1–2 cover January 1913–August 1918; columns 3–4 stop at March 1917. All regressions include county and month-year fixed effects; columns 2 and 4 add state-month fixed effects. $\widehat{\mathbb{E}(Y)}$ is the sample mean; $\widehat{\mathbb{E}(Y(0))}$ the estimated untreated mean for all observations; $\widehat{\mathbb{E}(Y(0) \mid C)}$ the estimated untreated mean for compliers. The three ratios divide $\hat{\beta}^{2SLS}$ by these base rates.

	Jan 1913–Aug 1918		Jan 1913–Mar 1917	
	Baseline	State×month FE	Baseline	State×month FE
$\hat{\beta}^{2SLS}$	0.0763	0.0603	0.0306	0.0274
$\widehat{\mathbb{E}(Y)}$	0.0341	0.0341	0.0101	0.0101
$\widehat{\mathbb{E}(Y(0))}$	0.0198	0.0228	0.0069	0.0073
$\widehat{\mathbb{E}(Y(0) \mid C)}$	0.0453	0.0482	0.0224	0.0207
Sample Mean Ratio	223.92%	176.76%	302.14%	270.75%
Simple Ratio	384.99%	263.92%	440.76%	377.01%
CAPE Ratio	168.42%	125.13%	136.80%	132.28%

Table 9: Demographic Characteristics of Compliers in Esposito et al. (2023)

Full-sample means and same-IV complier means for 1910 county characteristics. For each X, a complier mean is minus the screening coefficient after replacing enlistment with $X(1-D)$ in the corresponding Table 8 2SLS. Thus every column preserves the period, county and month-year fixed effects, optional state-by-month fixed effects, *Birth of a Nation* treatment, *Million Dollar Mystery* instrument, and county clustering. High Share City indicates an above-median share living in cities of at least 25,000; Low Share Illiteracy and Low Share Black indicate below-median shares.

	Jan 1913–Aug 1918			Jan 1913–Mar 1917		
	Sample	Baseline	State×month	Sample	Baseline	State×month
Total Population	51,503.67	117,886.21	95,512.76	51,503.67	134,798.07	109,690.47
Male Population	26,470.98	59,836.19	48,462.24	26,470.98	68,267.32	55,506.34
High Share City	0.1371	0.3520	0.3226	0.1371	0.3905	0.3668
Low Share Illiteracy	0.5000	0.6798	0.5378	0.5000	0.6922	0.5752
Low Share Black	0.5000	0.6102	0.4561	0.5000	0.6058	0.4745

5.5 Multiple binary instruments: Angelucci, Meraglia, and Voigtländer (2022)

Angelucci, Meraglia, and Voigtländer (2022) study how municipal self-governance (Farm Grants) in medieval English boroughs shaped parliamentary representation, instrumenting

Farm Grant status with royal borough status interacted with navigable-river, sea-coast, and Roman-road indicators, controlling for levels. The paper reports that Farm Grant boroughs were about 40 percentage points (OLS coefficients in their Table 2, columns 1-2) more likely to be summoned to Parliament, relative to a mean of 23 percent. The paper's 2SLS estimates in columns 6–8 range from 62.2 to 73.7 percentage points; these are the specifications used in our CAPE analysis below. The IV estimate is identified from boroughs whose grant status is shifted by royal-by-trade geography. If such marginal towns were commercially important enough to have had some route to representation even without a grant, broad benchmarks scale the LATE by too low a base.

This is what we find. Table 10 reports, for the paper's three 2SLS specifications, sample-mean ratios of 275–327%, simple ratios of 475–643%, and CAPE ratios of 227–399%: the untreated representation rate for compliers (0.274 in the baseline) is roughly twice the broad untreated benchmark (0.131). The effect remains economically large—the paper's message stands—but the proportional magnitude is roughly halved. Because the design has multiple binary instruments, standard scalar κ comparisons do not apply. The three instruments share the royal-status component and are therefore correlated rather than mutually independent; as discussed in Section 3.2, weight matching between numerator and denominator holds regardless, because both regressions share the paper's first stage, but non-negativity of the shared weights is not automatic in interacted designs, so the CAPE ratio here should be read—exactly like the published 2SLS coefficient—as a comparison of identically weighted aggregates under the original specification. Table 11 uses our same-IV transformed-outcome diagnostic (Section **??**), which shows the complier margin is concentrated among older, institutionally established boroughs: 93.2% of compliers had borough status by 1200 versus 38.8% of the sample, and 35.2% were Domesday boroughs versus 19.3%. Compliers are also more likely to enjoy freedom from tolls and murage or pavage rights by 1348. This application also illustrates the inference caveat of Section 3.6: the baseline complier base is 0.274 with a robust standard error of 0.117, and the 95% Fieller confidence interval for CAPE is [63%, 1,913%]. In 38.1% of the 1,000 paired-bootstrap draws, the estimated base is negative or lies within two bootstrap standard errors of zero, the diagnostic proposed in Appendix B; accordingly, the ordinary ratio-bootstrap standard deviation is unstable and we report the Fieller interval instead.

Table 10: Farm Grants and Parliamentary Representation in Angelucci, Meraglia, and Voigtländer (2022)

Replication of 2SLS estimates from Table 2, columns 6–8, of the original paper; the published coefficients are 0.622, 0.669, and 0.737. The dependent variable indicates summons to Parliament by 1348; the treatment indicates a Farm Grant by 1348; instruments are royal borough status interacted with navigable river, sea coast, and Roman road, controlling for levels. Column 2 adds county fixed effects and soil quality; column 3 uses entropy weights for mesne boroughs matching royal boroughs on trade geography. The sample mean ratio divides $\hat{\beta}^{2SLS}$ by the unweighted mean dependent variable; the simple ratio by $\widehat{\mathbb{E}(Y(0))}$; the CAPE ratio by $\widehat{\mathbb{E}(Y(0) \mid \text{complier})}$, estimated by replacing the dependent variable with $Y(1-D)$ in the same 2SLS.

	Baseline 2SLS	County FE + soil	Entropy weights
$\hat{\beta}^{2SLS}$	0.6216	0.6693	0.7387
Robust s.e.	(0.1754)	(0.1651)	(0.1751)
$\widehat{\mathbb{E}(Y)}$	0.2259	0.2259	0.2259
$\widehat{\mathbb{E}(Y(0))}$	0.1307	0.1235	0.1149
$\widehat{\mathbb{E}(Y(0) \mid \text{complier})}$	0.2741	0.2544	0.1850
Sample mean ratio	275.23%	296.32%	327.07%
Simple ratio	475.45%	542.11%	643.02%
CAPE ratio	226.77%	263.13%	399.30%

Table 11: Complier Characteristics in Angelucci, Meraglia, and Voigtländer (2022)

Sample means versus same-IV complier means for pre-1348 borough characteristics. For each characteristic X, the complier mean is minus the 2SLS coefficient on Farm Grant after replacing the dependent variable with $X(1-D)$. Column "Complier 6" uses the baseline specification; "Complier 7" adds county fixed effects and soil quality. The diagnostic is aligned with the multiple-instrument 2SLS margin and is distinct from scalar Abadie κ comparisons.

Characteristic	Sample	Complier 6	Complier 7
Borough status by 1200	0.388	0.932	0.844
Borough status by 1250	0.645	0.946	0.854
Domesday borough	0.193	0.352	0.353
Freedom from tolls by 1348	0.215	0.268	0.256
Murage or pavage by 1348	0.182	0.257	0.288

5.6 Continuous instrument: Berry, Fischer, and Guiteras (2020)

Berry, Fischer, and Guiteras (2020) study demand for and health effects of household water filters in Northern Ghana, instrumenting filter purchase with randomized price (take-it-or-

leave-it offers for TIOLI subjects, Becker–DeGroot–Marschak draws for BDM subjects). In the short run, the paper reports that the filter reduces child diarrhea by about 7 percentage points relative to a baseline rate of 21%—a reduction of about one-third; at one year, point estimates are positive, and the paper reports no evidence of benefits.

Table 12 computes CAPE for one short-run (pooled, one-month) and one long-run (TIOLI, one-year) specification. The two columns do not describe one complier group observed twice: samples, mechanisms, and first stages differ, so the price-sensitive margin differs. In the short run, $\hat{\beta}^{2SLS} = -0.072$ and the simple ratio (-49.5%) *overstates* the proportional benefit, because price-induced compliers have a higher untreated diarrhea rate (0.181 versus 0.145); CAPE is -39.8%. In the long run, the point estimates imply that the simple ratio (102.5%) understates CAPE (162.7%) because one-year TIOLI compliers have a lower untreated rate (0.135 versus 0.215). That proportional estimate is too imprecise for reliable inference, however: the complier base is 0.135 with a compound-clustered standard error of 0.070, and its confidence interval includes zero. The diagnostics in Table 13—computed with the continuous-IV transformed-outcome analogue, since scalar κ-weights do not apply—rationalize the point-estimate movements: short-run compliers have more young children and more baseline diarrhea; long-run TIOLI compliers treat water more effectively, have better water sources, more schooling and assets, and lower measured E. coli. Regression to the mean and temporary immunity after recent infection can further lower the one-year untreated base among households with high baseline diarrhea. None of this overturns the paper's central finding that effects vary with willingness to pay; it shows that the long-run proportional magnitude is too imprecise to summarize by a finite confidence interval.

Table 12: Water Filters and Child Diarrhea in Berry, Fischer, and Guiteras (2020)

CAPE calculations for two Table 2 specifications of the original paper. Dependent variable: indicator for diarrhea among children 0–5 in the previous two weeks. Treatment: household filter purchase. Instrument: randomized offer price (TIOLI offer or BDM draw). Short run: one-month combined sample with controls and village fixed effects; long run: one-year TIOLI sample with controls and village fixed effects; standard errors clustered by compound. The simple ratio divides $\hat{\beta}^{2SLS}$ by the untreated outcome mean in the IV sample; CAPE by $\widehat{\mathbb{E}(Y(0) \mid \text{complier})}$ from the transformed-outcome 2SLS.

	Short-run, pooled	Long-run, TIOLI
$\hat{\beta}^{2SLS}$	−0.0720	0.2201
Clustered s.e.	(0.0350)	(0.1001)
$\widehat{\mathbb{E}(Y(0))}$	0.1454	0.2149
$\widehat{\mathbb{E}(Y(0) \mid \text{complier})}$	0.1807	0.1353
Simple ratio	−49.50%	102.45%
CAPE ratio	−39.81%	162.73%

Table 13: Continuous-IV Complier Diagnostics in Berry, Fischer, and Guiteras (2020)

Sample versus same-IV complier means for the two specifications of Table 12. Because the instrument is randomized price, complier means are the continuous-IV analogue: minus the 2SLS coefficient after replacing the dependent variable with $X(1 - D)$. Characteristics are measured at baseline; E. coli and turbidity are z-scores with lower values indicating cleaner water.

	Short-run, pooled		Long-run, TIOLI	
Characteristic	Sample	Complier	Sample	Complier
Children 0–5 with baseline diarrhea	0.340	0.382	0.380	0.493
Children 0–5 in household	1.641	1.676	1.553	1.669
Currently treats water effectively	0.111	0.108	0.086	0.107
All-year improved water source	0.168	0.186	0.169	0.195
Assets index	0.159	0.134	0.094	0.121
Respondent attended school	0.095	0.108	0.147	0.181
E. coli count, standardized	−0.089	−0.162	−0.179	−0.367
Turbidity, standardized	−0.076	−0.195	−0.326	−0.166

5.7 A Unit-invariant Estimand for Logs with zeros: Acharya and Xu (2017)

Acharya and Xu (2017) examine the effect of public listing on innovation, exploiting the jump in delisting probability at NASDAQ's continued-listing thresholds. Their estimated coefficient implies, as we compute, that delisted firms in externally finance-dependent industries produce about 32% fewer patents than public counterparts three years post-delisting.[6]

This magnitude is problematic because many sample firms have zero patents: Guo, Pérez-Castrillo, and Toldrà-Simats (2019) report over a quarter of U.S. public firm-years at zero, and Acharya and Xu additionally set missing patent counts to zero. As Chen and Roth (2024) show, average effects on $\log(1+Y)$ then have no percentage interpretation and are unit-dependent. The natural target is instead θ of equation (2): the level LATE scaled by the complier control mean. Table 14 replicates the paper's fuzzy-RDD specification (our $\hat{\beta}_1 = -0.3811$ versus -0.3863 in the original), and Table 15 separates the scales. In the baseline, the log coefficient implies -31.7%; the raw-count 2SLS estimate divided by a conventional constructed untreated mean is -112.6%; and the CAPE point estimate is -94.8%. Thus most of the contrast with -31.7% comes from moving to the raw outcome, while aligning the raw denominator with compliers moderates the magnitude by 17.8 percentage points.

The uncapped bootstrap yields a 95% percentile interval of $[-118.0, -61.7]\%$ for the baseline CAPE. The estimated base is 0.718 with a bootstrap standard error of 0.461, and 66.3% of draws have a negative base or a base within two bootstrap standard errors of zero; the corresponding Fieller set is all real. The -94.8% point estimate is therefore large in magnitude but should be treated as descriptive, not as a bounded proportional claim.

[6] The paper's reported "47% fewer patents" derives from $e^{0.3863} - 1 = 0.4715$ using the absolute value of $\hat{\beta}_1 = -0.3863$ (dependent variable: $\ln(1 + \text{patents})$); we compute $e^{-0.3863} - 1 = -0.3204$.

Table 14: NASDAQ Delisting and Patent Applications in Acharya and Xu (2017)

Our fuzzy-RDD estimates of the effect of delisting on patent applications three years ahead for externally finance-dependent firms. The source paper's Table 4, Panel A coefficient for $t+3$ patents is -0.3863 (0.1304); column 1 is the direct baseline comparator, while columns 2–4 are our specification variants. The dependent variable is $\ln(1+\text{patents}_{t+3})$. The running variable $c_{i,t}$ is the minimum of the maximum of three normalized core NASDAQ requirements and one normalized non-core requirement; $z_{i,t} = \mathbf{1}(c_{i,t} \leq 0)$ instruments Delisting. Covariates are ln(Assets), Tangible, Cash, Age, Capex, and ROA. Columns 2 and 4 add the interaction of $z_{i,t}$ and $c_{i,t}$. * $p < .1$, ** $p < .05$, *** $p < .01$.

	(1)	(2)	(3)	(4)
Rep. 2SLS: Delisting	-0.3811^{**}	-0.3976^{**}	-0.2638^{*}	-0.2683^{*}
	(0.1687)	(0.1809)	(0.1517)	(0.1611)
Covariates	Y	Y		
Interaction		Y		Y
N	2422	2422	2478	2478

Table 15: CAPE Estimates for Acharya and Xu (2017)

Estimates for the four specifications of Table 14. Log-implied is $\exp(\hat{\beta}_{\log(1+Y)}) - 1$. Raw 2SLS uses the patent count. Raw simple divides that coefficient by the model-implied conventional untreated mean, $\widehat{\mathbb{E}(Y)} - \hat{\beta}_Y \widehat{\mathbb{E}(D)}$, in the same estimation sample. CAPE base is minus the 2SLS coefficient with dependent variable $\text{Patent}_{t+3}(1 - \text{Delisting})$, and CAPE divides Raw 2SLS by this complier untreated mean. The last column reports the unconstrained 95% percentile-bootstrap interval from 1,000 observation resamples; no ratio draw is capped or trimmed.

Covariates	Interaction	Log-implied	Raw 2SLS	Raw simple	CAPE base	CAPE	95% CI
Y		−31.69%	−0.6810	−112.59%	0.7182	−94.82%	[−118.01, −61.73]%
Y	Y	−32.81%	−0.7057	−115.53%	0.7464	−94.54%	[−130.43, −64.81]%
		−23.19%	−0.3476	−65.76%	0.3735	−93.07%	[−173.80, −27.30]%
	Y	−23.53%	−0.3389	−64.36%	0.3667	−92.41%	[−187.30, −53.41]%

5.8 Outcome scaling under a continuous IV: Sampat and Williams (2019)

Sampat and Williams (2019) study whether human-gene patents reduce follow-on innovation. Their main design instruments the binary grant of a gene-patent application with examiner leniency, a continuous grant-rate measure, and uses art-unit-by-application-year fixed effects. This application therefore shows directly that CAPE is not confined to binary instruments.

The scientific-publication result separates two scaling problems. Because 91.1% of observations have no scientific publication, the Table 3 coefficient of -0.0230 on $\log(1 +$

publications) mechanically implies −2.3%, but that transformation does not define a unit-invariant percentage effect. Re-estimating the same specification in raw counts yields a level LATE of −0.0571: scaling by the sample mean gives −39.6%, whereas the complier untreated mean of 0.126 gives CAPE of −45.2%. Most of the contrast with −2.3% is therefore an outcome-scale correction; aligning the denominator adds 5.5 percentage points on the raw scale.

Table 16 reports the three log or binary outcomes for which data are available in the public archive, all of which we replicate. The binary any-publication effect moves from −21.0% against its raw sample mean to a CAPE of −27.3%, while diagnostic-test availability moves from −15.4% to −23.5%; the two clinical-trial outcomes in the paper's Table 3 rely on proprietary Pharmaprojects data and are omitted.

Although this is a two-sample IV, the common first stage cancels exactly from the CAPE ratio. We therefore jointly resample the 1,545 patent-application clusters in the outcome sample and construct Fieller intervals from the bootstrap covariance of the two reduced forms. The 95% intervals are $[-79.8, -8.5]$% for scientific publications and $[-60.3, 40.2]$% for diagnostic-test availability; Table 3 reports these intervals because the ratio draws are tail-sensitive, especially for the diagnostic-test outcome.

Table 16: Gene Patents and Follow-on Innovation in Sampat and Williams (2019)

The three log or binary outcomes from Table 3 of Sampat and Williams (2019) that are available in the public archive. Patent grant is binary and examiner leniency is continuous. Paper 2SLS includes the published standard error; Rep. 2SLS is our point estimate on the same scale. Raw 2SLS uses the untransformed count for scientific publications and is identical to Rep. 2SLS for binary outcomes. The current two-sample script does not estimate standard errors for the replicated or raw coefficients, so these columns report points only. The common first stage cancels from CAPE; Table 3 reports outcome-sample cluster-bootstrap Fieller intervals for its two selected Sampat–Williams rows, and the analogous any-publication interval is $[-58.5, -1.3]$%. Log-implied is $\exp(\hat{\beta}) - 1$ for the log outcome. Raw simple divides Raw 2SLS by the raw-outcome sample mean, whereas CAPE divides it by the estimated complier untreated mean; "–" means structurally inapplicable.

Outcome	Scale	Paper 2SLS (s.e.)	Rep. 2SLS	Raw 2SLS	Log-implied	Raw mean	Raw simple	CAPE base	CAPE
Scientific publications	$\log(1+Y)$	−0.0230 (0.0102)	−0.0230	−0.0571	−2.27%	0.1440	−39.64%	0.1264	−45.18%
Any scientific publication	binary	−0.0187 (0.0089)	−0.0187	−0.0187	–	0.0888	−21.01%	0.0683	−27.33%
Diagnostic test available	binary	−0.0141 (0.0123)	−0.0141	−0.0141	–	0.0918	−15.38%	0.0601	−23.47%

5.9 Outcome-specific denominator revisions: Asher and Novosad (2020)

Asher and Novosad (2020) estimate the effects of India's PMGSY rural-road program with a fuzzy RDD. Receipt of a new road is instrumented by whether village population lies

above the relevant eligibility threshold. Because the paper studies many outcomes under the same design, it shows that denominator alignment is outcome-specific: CAPE can materially revise one magnitude and validate another without changing either level estimate.

The Table 4 transport outcomes give the larger revisions. Government-bus availability rises by 0.129, or 109.9% of the below-threshold mean; its complier untreated mean is lower, producing a CAPE of 152.5%. Private-bus availability moves from 55.9% to 162.2%, but its complier base is 0.070 with a standard error of 0.059. That second point ratio is therefore a useful warning about denominator uncertainty, not a stable headline magnitude.

The two Table 5 occupational shares provide the validation cases. The agricultural share moves from −19.3% to −16.9%, and the manual-labor share from 16.0% to 18.4%; both gaps are about 2.4 percentage points. Thus the paper's main reallocation pattern—workers move out of agriculture and into manual labor—is essentially unchanged after aligning the denominator, whereas transport access is more sensitive to which untreated population supplies the base rate.

Table 17: Selected Main-Table Outcomes in Asher and Novosad (2020)

Fuzzy-RDD estimates for four selected outcomes from Tables 4 and 5 of Asher and Novosad (2020). Road receipt is instrumented by being above the relevant population threshold. Paper 2SLS is the published coefficient; Rep. 2SLS is our estimate with its heteroskedasticity-robust standard error. Raw simple divides Rep. 2SLS by the below-threshold mean, and CAPE divides it by the estimated complier untreated mean. Gap is the absolute difference in percentage points. The CAPE-base standard error makes the private-bus denominator uncertainty visible.

Paper table	Outcome	Paper 2SLS	Rep. 2SLS (s.e.)	Ctrl. mean	CAPE base (s.e.)	Raw simple	CAPE	Gap
Table 4	Government bus available	0.129	0.1294 (0.0554)	0.1178	0.0849 (0.0428)	109.87%	152.51%	42.65
Table 4	Private bus available	0.114	0.1144 (0.0751)	0.2046	0.0705 (0.0588)	55.91%	162.25%	106.34
Table 5	Workers in agriculture	−0.092	−0.0917 (0.0434)	0.4756	0.5424 (0.0364)	−19.27%	−16.90%	2.37
Table 5	Workers in manual labor	0.072	0.0718 (0.0434)	0.4482	0.3896 (0.0363)	16.02%	18.43%	2.41

5.10 Comparing CAPE with DKM in replicated papers

Table 18 compares CAPE with the DKM endpoint construction in the continuous-instrument applications of Berry, Fischer, and Guiteras (2020) and Sampat and Williams (2019). The comparison keeps the two targets distinct. The matched endpoint ratio divides the DKM endpoint Wald effect by its endpoint base; CAPE divides the full-support 2SLS effect by the base recovered from a transformed-outcome 2SLS using the same full-support specification. For completeness, the table also reports the practitioner hybrid that divides the full-support 2SLS effect by the endpoint base. That hybrid mixes the two weighting schemes and is not a DKM estimate.

The Berry short-run specification is a validation case. Its matched endpoint ratio is −37.1

percent and CAPE is −39.8 percent, showing that the DKM endpoint construction is adequate in this design while CAPE reaches the same conclusion without selecting percentiles. The two well-measured Sampat–Williams outcomes differ more: the matched endpoint and CAPE ratios are −26.5 and −45.2 percent for publication counts and −15.6 and −27.3 percent for any publication. These differences do not make either construction internally incorrect; they show that the selected endpoint population and the full-support 2SLS weighting can describe economically different complier margins.

The remaining two rows are descriptive. The Berry long-run CAPE base is 0.135 with a standard error of 0.070, and the Sampat–Williams diagnostic-test Fieller interval includes zero. We therefore do not treat their point differences as evidence about the relative constructions. Across endpoint choices, the same qualitative distinction remains: Appendix Table A.1 reports p5–p95 and min–max results.

Table 18: Endpoint and Same-IV Complier Bases in Continuous-Instrument Applications

Berry, Fischer, and Guiteras (2020) study child diarrhea after filter purchase, instrumented by randomized price. The short-run row uses the combined one-month sample and the long-run row the one-year TIOLI sample; both retain the original controls and village fixed effects. The price endpoints are 1.5 and 8.0 in the short-run sample and 2.0 and 6.0 in the long-run sample, with lower price as the high-treatment endpoint. Sampat and Williams (2019) study follow-on outcomes through 2011 for human-gene patent applications filed in 2000–2005, with patent grant instrumented by examiner leniency and art-unit-by-application-year fixed effects. Its two-sample calculation retains the first-stage slope from the application sample, calibrates fixed-effect-specific treatment intercepts to the outcome population, and standardizes components over 284,368 common-support outcome rows. Its endpoints are 0.144 and 0.398. In both applications, the DKM-style calculation evaluates linear treatment, untreated-outcome, and treated-outcome components at the 10th and 90th percentiles and standardizes them over common rows. Matched endpoint divides the endpoint Wald effect by the DKM endpoint base. Practitioner hybrid divides the full-support 2SLS effect by that base and is not a DKM estimate. CAPE uses matched full-support 2SLS and transformed-outcome 2SLS coefficients. Publication count is in levels. Parentheses give the CAPE-base standard error; endpoint entries are point estimates, and no formal test of the endpoint–CAPE base difference is reported. Panel B is descriptive because its CAPE denominators are fragile. No fitted linear-probability component is clipped.

		Untreated complier base		Proportional effect		
Paper	Outcome	DKM endpoint	CAPE (s.e.)	Matched endpoint	Practitioner hybrid	CAPE
Panel A: Informative comparisons						
Berry et al. (2020)	Short-run diarrhea	0.1727	0.1807 (0.0294)	−37.13%	−41.67%	−39.81%
Sampat–Williams (2019)	Publication count	0.1899	0.1264 (0.0373)	−26.51%	−30.08%	−45.18%
	Any publication	0.1069	0.0683 (0.0190)	−15.55%	−17.46%	−27.33%
Panel B: Fragile denominators; descriptive point estimates only						
Berry et al. (2020)	Long-run diarrhea	0.0826	0.1353 (0.0702)	342.61%	266.36%	162.73%
Sampat–Williams (2019)	Diagnostic test	0.1036	0.0601 (0.0216)	−21.70%	−13.63%	−23.47%

6 Simulation Evidence

This section uses simulations for two purposes. First, we assess finite-sample behavior of the CAPE estimator $\widehat{\theta}$ and the coverage of the three inference procedures in Section 3.6—the paired percentile bootstrap, the joint delta method, and Fieller confidence sets—as a function of denominator strength. Second, we provide controlled numerical illustrations of two properties that motivate CAPE in applied work: (i) when outcomes have mass at zero, CAPE delivers a unit-invariant proportional-effect estimand while log-based "percent effects" can vary mechanically with outcome units; and (ii) in multi-instrument settings where partial monotonicity holds, CAPE converges to a ratio of like-weighted averages and (asymptotically) lies within the range of subgroup-level proportional effects, unlike conventional scalings.

The simulation designs are constructed to satisfy the maintained IV assumptions (SUTVA, exclusion, random assignment, and monotonicity) by construction using a single latent uniform per unit to couple potential treatment states. Full design details, grids, and tables are reported in Appendix C. In the main text we emphasize the central practical message: the relevant diagnostic for whether a proportional statement is well-supported is the t-statistic of the estimated complier base rate, $\widehat{\mu}_C \equiv -\widehat{\beta}_{2SLS}^{Y(1-D)}$. When this denominator is weak, delta-method intervals can severely undercover and percentile-bootstrap intervals can achieve nominal coverage only by becoming extremely wide; Fieller sets remain valid and explicitly signal weak denominators via unbounded confidence sets. Consistent with this, we recommend always reporting $\widehat{\mu}_C$ with its own standard error and using Fieller sets when $\widehat{\mu}_C$ is not statistically distinguishable from zero.

7 Practical Guidance

Our recommendations for practice are brief. First, alongside any percentage-change interpretation of an IV or fuzzy-RDD estimate, report the estimated complier base rate with its own standard error and CAPE; use delta-method or paired-bootstrap inference when the base is well separated from zero, Fieller confidence sets when it is marginal, and make no proportional claim when it is statistically indistinguishable from zero. The marginal cost is one regression per outcome. Second, when the outcome is weakly positive with mass at zero, do not interpret $\log(1+Y)$ coefficients as percentages; CAPE delivers the well-defined, unit-invariant parameter of Chen and Roth (2024) at the identified margin. Third, use the transformed-outcome complier diagnostics to tell readers who the compliers are; this converts a mechanical adjustment into economics and often signs the base-rate revision ex

ante. Fourth, treat small, imprecise, or negative estimated complier bases as red flags: they indicate that no proportional statement—conventional or adjusted—is supported at that margin. Finally, the simple ratio is likely adequate when compliance is near-universal or when complier covariates closely match the sample; CAPE verifies this cheaply rather than assuming it. Stata and R implementations accompany the paper.

8 Conclusion

Interpreting IV and fuzzy-RDD estimates as percentage changes requires a base rate for the population whose treatment effect is identified: compliers. We show this base rate is estimable with one additional regression—the researcher's own 2SLS with $Y(1-D)$ as the dependent variable—and that the resulting ratio has an internally coherent interpretation under single, multiple, and continuous instruments, because the numerator and denominator aggregate margin-specific effects with identical first-stage-determined weights. Our top-journal survey identifies 946 papers with an IV or fuzzy-RDD setting, including 564 finance papers; 273 of them, or 28.9%, use a binary treatment in the same setting, including 152 finance papers. Across nine papers, including four finance papers, and fourteen representative settings, we find like-for-like conventional magnitudes off by 0.6 to 249 percentage points in either direction; the log-outcome applications reveal a separate outcome-scale problem, even as every paper's qualitative conclusion survives. The economic-significance step of quasi-experimental research deserves the same care the profession now devotes to identification; this paper supplies the missing denominator.

References

[1] Abadie, Alberto, 2002, Bootstrap tests for distributional treatment effects in instrumental variable models, *Journal of the American Statistical Association* 97, 284–292.

[2] Abadie, Alberto, 2003, Semiparametric instrumental variable estimation of treatment response models, *Journal of Econometrics* 113, 231–263.

[3] Acharya, Viral, and Zhaoxia Xu, 2017, Financial dependence and innovation: The case of public versus private firms, *Journal of Financial Economics* 124, 223–243.

[4] Angelucci, Charles, Simone Meraglia, and Nico Voigtländer, 2022, How merchant towns shaped parliaments: From the Norman Conquest of England to the Great Reform Act, *American Economic Review* 112, 3441–3487.

[5] Angrist, Joshua D., and Guido W. Imbens, 1995, Two-stage least squares estimation of average causal effects in models with variable treatment intensity, *Journal of the American Statistical Association* 90, 431–442.

[6] Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin, 1996, Identification of causal effects using instrumental variables, *Journal of the American Statistical Association* 91, 444–455.

[7] Asher, Sam, and Paul Novosad, 2020, Rural roads and local economic development, *American Economic Review* 110, 797–823.

[8] Atanasov, Vladimir, and Bernard Black, 2016, Shock-based causal inference in corporate finance and accounting research, *Critical Finance Review* 5, 207–304.

[9] Baker, Andrew C., David F. Larcker, and Charles C. Y. Wang, 2022, How much should we trust staggered difference-in-differences estimates?, *Journal of Financial Economics* 144, 370–395.

[10] Berry, James, Greg Fischer, and Raymond Guiteras, 2020, Eliciting and utilizing willingness to pay: Evidence from field trials in Northern Ghana, *Journal of Political Economy* 128, 1436–1473.

[11] Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky, 2022, When is TSLS actually LATE?, NBER Working Paper No. 29709.

[12] Celerier, Claire, and Purnoor Tak, 2022, Exploiting minorities through advertising: Evidence from the Freedman's Savings Bank, Working paper.

[13] Chen, Jiafeng, and Jonathan Roth, 2024, Logs with zeros? Some problems and solutions, *The Quarterly Journal of Economics* 139, 891–936.

[14] Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz, 2016, The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment, *American Economic Review* 106, 855–902.

[15] Cohn, Jonathan B., Zack Liu, and Malcolm I. Wardlaw, 2022, Count (and count-like) data in finance, *Journal of Financial Economics* 146, 529–551.

[16] Dahl, Gordon B., Andreas Ravndal Kostøl, and Magne Mogstad, 2014, Family welfare cultures, *The Quarterly Journal of Economics* 129, 1711–1752.

[17] Esposito, Elena, Tiziano Rotesi, Alessandro Saia, and Mathias Thoenig, 2023, Reconciliation narratives: The Birth of a Nation after the US Civil War, *American Economic Review* 113, 1461–1504.

[18] Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljungqvist, 2020, What is a patent worth? Evidence from the U.S. patent "lottery," *The Journal of Finance* 75, 639–682.

[19] Fieller, Edgar C., 1954, Some problems in interval estimation, *Journal of the Royal Statistical Society, Series B* 16, 175–185.

[20] Guo, Bing, David Pérez-Castrillo, and Anna Toldrà-Simats, 2019, Firms' innovation strategy under the shadow of analyst coverage, *Journal of Financial Economics* 131, 456–483.

[21] Heath, Davidson, Matthew C. Ringgenberg, Mehrdad Samadi, and Ingrid M. Werner, 2023, Reusing natural experiments, *The Journal of Finance* 78, 2329–2364.

[22] Heckman, James J., and Edward Vytlacil, 2005, Structural equations, treatment effects, and econometric policy evaluation, *Econometrica* 73, 669–738.

[23] Hirschberg, Joe, and Jenny Lye, 2010, A geometric comparison of the delta and Fieller confidence intervals, *The American Statistician* 64, 234–241.

[24] Imbens, Guido W., and Joshua D. Angrist, 1994, Identification and estimation of local average treatment effects, *Econometrica* 62, 467–475.

[25] Imbens, Guido W., and Donald B. Rubin, 1997, Estimating outcome distributions for compliers in instrumental variables models, *The Review of Economic Studies* 64, 555–574.

[26] Jiang, Wei, 2017, Have instrumental variables brought us closer to the truth, *The Review of Corporate Finance Studies* 6, 127–140.

[27] Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman, 2001, Moving to Opportunity in Boston: Early results of a randomized mobility experiment, *The Quarterly Journal of Economics* 116, 607–654.

[28] Kleiner, Kristoph, Noah Stoffman, and Scott E. Yonker, 2021. Friends with bankruptcy protection benefits, *Journal of Financial Economics* 139(2), 578–605.

[29] Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, 2007, Experimental analysis of neighborhood effects, *Econometrica* 75, 83–119.

[30] Marbach, Moritz, and Dominik Hangartner, 2020, Profiling compliers and noncompliers for instrumental-variable analysis, *Political Analysis* 28, 435–444.

[31] Mitton, Todd, 2022, Methodological variation in empirical corporate finance, *The Review of Financial Studies* 35, 527–575.

[32] Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters, 2021, The causal interpretation of two-stage least squares with multiple instrumental variables, *American Economic Review* 111, 3663–3698.

[33] Mullahy, John, and Edward C. Norton, 2024, Why transform Y? The pitfalls of transformed regression models with a mass at zero, *Oxford Bulletin of Economics and Statistics* 86, 417–447.

[34] Mukherjee, Abhiroop, George Panayotov, Rik Sen, Harsha Dutta, and Pulak Ghosh, 2022, Measuring vaccine effectiveness from limited public health datasets: Framework and estimates from India's second COVID wave, *Science Advances* 8, eabn4274.

[35] Roberts, Michael R., and Toni M. Whited, 2013, Endogeneity in empirical corporate finance, in George M. Constantinides, Milton Harris, and René M. Stulz, eds.: *Handbook of the Economics of Finance*, Volume 2 (Elsevier, Amsterdam), 493–572.

[36] Sampat, Bhaven, and Heidi L. Williams, 2019, How do patents affect follow-on innovation? Evidence from the human genome, *American Economic Review* 109, 203–236.

[37] Solis, Alex, 2017, Credit access and college enrollment, *Journal of Political Economy* 125, 562–622.

[38] Staiger, Douglas, and James H. Stock, 1997, Instrumental variables regression with weak instruments, *Econometrica* 65, 557–586.

[39] Stein, Luke C. D., and Constantine Yannelis, 2020, Financial inclusion, human capital, and wealth accumulation: Evidence from the Freedman's Savings Bank, *The Review of Financial Studies* 33, 5333–5377.

Appendix

A Extension to Continuous Instruments

This appendix formalizes Section 3.3. Let Z_i have support on an interval $\mathcal{Z} \subseteq \mathbb{R}$, with binary D_i and outcome Y_i. SUTVA and exclusion carry over, so $Y_i(\mathbf{D}, \mathbf{Z}) = Y_i(D_i)$.

Assumption 7 (Random instrument, continuous)**.** *Z_i is independent of potential outcomes $\{Y_i(0), Y_i(1)\}$ and potential treatment states $\{D_i(z) : z \in \mathcal{Z}\}$, and is i.i.d. across i.*

Assumption 8 (Relevance, continuous)**.** $\mathrm{Cov}(D_i, Z_i) \neq 0$.

Assumption 9 (Monotonicity, continuous)**.** *$D_i(z)$ is weakly increasing in z for all i.*

Under Assumption 9, each unit has a (possibly infinite) threshold $z_i^* \in [-\infty, +\infty]$ with $D_i(z) = \mathbf{1}(z \geq z_i^*)$; always-takers have $z_i^* = -\infty$, never-takers $z_i^* = +\infty$, and a unit is a *complier at margin z* if $z_i^* = z$.

Core identity. For any $z' > z$,

$$Y_i\big(D_i(z')\big)\big(1 - D_i(z')\big) - Y_i\big(D_i(z)\big)\big(1 - D_i(z)\big) = \begin{cases} -Y_i(0), & D_i(z) = 0,\ D_i(z') = 1, \\ 0, & D_i(z) = D_i(z'), \end{cases}$$

with defiance ruled out by Assumption 9. At every local margin, the causal effect on $Y(1-D)$ is $-Y_i(0)$ for units switching at that margin and zero otherwise.

2SLS aggregation. By Angrist and Imbens (1995), under Assumptions 7–9, 2SLS identifies a weighted average of margin-specific effects,

$$\hat{\beta}^{Y}_{\mathrm{2SLS}} \xrightarrow{p} \int_{\mathcal{Z}} \mathbb{E}\big[Y_i(1) - Y_i(0) \mid z_i^* = z\big]\, \omega(z)\, dz,$$

with $\omega(z) \geq 0$, $\int_{\mathcal{Z}} \omega(z)\, dz = 1$, and ω determined only by the first-stage function $\mathbb{E}[D_i \mid Z_i = z]$ and the distribution of Z_i—not by the outcome. Since the transformed-outcome regression shares the first stage and its local effect at each margin is $-Y_i(0)$,

$$\hat{\beta}^{Y(1-D)}_{\mathrm{2SLS}} \xrightarrow{p} -\int_{\mathcal{Z}} \mathbb{E}\big[Y_i(0) \mid z_i^* = z\big]\, \omega(z)\, dz,$$

with the identical weight function, and (10) follows by the continuous mapping theorem. With a continuous instrument, "compliers" are a continuum of margin-specific groups; CAPE is a weighted-average proportional effect across margins with common weights in numerator and denominator, and the bounded-range property is preserved. With a vector of possibly continuous instruments, 2SLS constructs the scalar effective instrument $\hat{D}_i = \mathbf{Z}_i'\hat{\boldsymbol{\pi}}$ from the first stage; the two regressions share this index, so weight matching holds mechanically, but a positively weighted margin-specific interpretation additionally requires monotonicity of $D_i(\mathbf{z})$ in the index, which partial monotonicity (Assumption 6) alone does not imply; see Mogstad, Torgovitsky, and Walters (2021) and Section 3.2.

A.1 Endpoint-choice sensitivity

Table A.1 varies the endpoint pair used in the DKM-style construction. The matched endpoint ratios remain coherent for each selected pair, but the selected pair changes the complier margins being summarized. CAPE has no corresponding endpoint choice because its numerator and denominator both use the full-support 2SLS weighting.

Table A.1: Endpoint-Choice Sensitivity in the Continuous-Instrument Applications

Berry, Fischer, and Guiteras (2020) study child diarrhea after filter purchase, instrumented by randomized price; the short-run row uses the combined one-month sample and the long-run row the one-year TIOLI sample. The component models retain the original controls and village fixed effects and are standardized over common IV rows. Sampat and Williams (2019) study follow-on outcomes through 2011 for human-gene patent applications filed in 2000–2005, with patent grant instrumented by examiner leniency and art-unit-by-application-year fixed effects. In that two-sample design, the construction retains the first-stage slope from the application sample, calibrates fixed-effect-specific treatment intercepts to the outcome population, and standardizes all components over the 284,368 outcome rows with common treated and untreated fixed-effect support. In each endpoint block, Base is $\mu_{0,EP}$ from equation (11); Matched is $\tau_{EP}/\mu_{0,EP}$, using the endpoint Wald numerator; and Hybrid is the published full-support 2SLS coefficient divided by the endpoint base. The hybrid is a possible practitioner adaptation, not a DKM estimate. CAPE uses the full-support 2SLS numerator and its same-IV transformed-outcome base. All endpoint entries are un-clipped point estimates; no formal test of an endpoint–CAPE difference is reported. The dagger marks rows whose CAPE denominator is inferentially fragile, as detailed in Table 18. For Berry short-run, the p5–p95 and min–max linear first stages imply inadmissible aggregate compliance-type shares and are extrapolative sensitivities. Min–max entries more generally should be read as descriptive.

		Full-support CAPE		p10–p90 endpoints			p5–p95 endpoints			Min–max endpoints		
Paper	Outcome	Base	CAPE	Base	Matched	Hybrid	Base	Matched	Hybrid	Base	Matched	Hybrid
Berry et al.	Short-run diarrhea	0.1807	−39.81%	0.1727	−37.13%	−41.67%	0.1806	−36.81%	−39.84%	0.1925	−36.37%	−37.38%
	Long-run diarrhea†	0.1353	162.73%	0.0826	342.61%	266.36%	0.0826	342.61%	266.36%	0.0826	342.61%	266.36%
Sampat–Williams	Publication count	0.1264	−45.18%	0.1899	−26.51%	−30.08%	0.1871	−25.52%	−30.52%	0.1782	−22.11%	−32.05%
	Any publication	0.0683	−27.33%	0.1069	−15.55%	−17.46%	0.1058	−14.50%	−17.64%	0.1022	−10.93%	−18.25%
	Diagnostic test†	0.0601	−23.47%	0.1036	−21.70%	−13.63%	0.1029	−21.32%	−13.72%	0.1008	−20.05%	−14.01%

Randomized price has discrete support, so we also estimated a saturated observed-price-cell version of the Berry endpoint construction. The short-run matched ratios are −45.1, 38.6, and −44.4 percent under p10–p90, p5–p95, and min–max, respectively, based on only 15–32 compounds in the relevant endpoint cells. The long-run endpoint pair is unchanged across the three rules and gives 174.0 percent, based on 44 and 48 endpoint-cell compounds. This sensitivity underscores that endpoint results can depend on both support and component-model choices; it does not alter the interpretation of any internally matched endpoint ratio.

B Inference Details

This appendix derives the joint distribution underlying Section 3.6, gives a cluster-robust estimator of the joint covariance, solves the Fieller set in closed form, and records the bootstrap algorithm.

Joint distribution and covariance estimation. Let $r_i = (D_i, W_i')'$ collect the second-stage regressors (treatment, covariates, and fixed effects, with the constant absorbed into W_i) and $m_i = (\mathbf{Z}_i', W_i')'$ the instrument vector; both are common to the two regressions. Write R and M for the stacked matrices, $\hat{R} = M(M'M)^{-1}M'R$ for the projected regressors, and $\hat{r}_i$ for row i of $\hat{R}$. The two estimators are $\hat{\beta}^Y = (\hat{R}'R)^{-1}\hat{R}'Y$ and $\hat{\beta}^{\tilde{Y}}$, defined identically with $\tilde{Y}_i = Y_i(1 - D_i)$, with residuals $\hat{\varepsilon}_i = Y_i - r_i'\hat{\beta}^Y$ and $\tilde{\varepsilon}_i = \tilde{Y}_i - r_i'\hat{\beta}^{\tilde{Y}}$. Because the two equations share regressors and instruments, the system point estimates coincide with equation-by-equation 2SLS; only the covariance requires a joint computation. Under standard 2SLS regularity conditions the stacked estimator is jointly asymptotically normal, and a cluster-robust estimator of the joint covariance of the two coefficient vectors is

$$\widehat{\mathbf{V}} \;=\; \Big(I_2 \otimes (\hat{R}'R)^{-1}\Big)\Big(\sum_{g=1}^{G} \hat{s}_g \hat{s}_g'\Big)\Big(I_2 \otimes (\hat{R}'R)^{-1}\Big), \qquad \hat{s}_g = \begin{pmatrix} \sum_{i \in g} \hat{r}_i\, \hat{\varepsilon}_i \\ \sum_{i \in g} \hat{r}_i\, \tilde{\varepsilon}_i \end{pmatrix}, \tag{17}$$

where g indexes the clusters of the original design (single observations, if the design does not cluster). The scalars required by (13) and (14) are read off $\widehat{\mathbf{V}}$: $\widehat{V}_{\beta\beta}$ and $\widehat{V}_{\mu\mu}$ are the diagonal entries corresponding to the two treatment coefficients, and $\widehat{V}_{\beta\mu}$ is the *negative* of their off-diagonal entry, the sign reflecting $\hat{\mu}_C = -\hat{\beta}^{\tilde{Y}}_{\text{2SLS}}$.

Implementation. The joint covariance requires no new software. Stack the data: duplicate the estimation sample; define an equation indicator taking the dependent variable to Y in the first copy and $Y(1 - D)$ in the second; interact the treatment, the excluded instruments, and every control and fixed effect with both values of the indicator; and estimate the fully interacted system by a single 2SLS, clustering at the original level with clusters

spanning both copies of each observation. The two treatment coefficients and their full joint covariance matrix are then delivered by one estimation. (With high-dimensional fixed effects this doubles the number of absorbed effects, which modern software handles.) Alternatively, the entries of $\widehat{\mathbf{V}}$ may be estimated from the paired bootstrap draws described below, so that a Fieller set can be constructed even when the analytic stacking is inconvenient. In two-sample designs, where the pipeline combines moments estimated in different samples, the closed forms above do not apply directly; there we bootstrap each sample independently and recompute the full pipeline on every draw. An exception arises when the same first-stage estimate enters the numerator and denominator multiplicatively and therefore cancels from CAPE. In that case, resampling the shared reduced-form sample is sufficient for CAPE inference, although separate component 2SLS and complier-base standard errors still require both samples.

Delta method. The map $\theta(\beta, \mu) = \beta/\mu$ is continuously differentiable at any $\mu \neq 0$ with gradient $(1/\mu, -\beta/\mu^2)$; applying the sandwich formula to the relevant 2×2 block of $\widehat{\mathbf{V}}$ yields (13). (Equivalently, when $\hat{\beta}^Y \neq 0$, $\widehat{\mathrm{Var}}(\widehat{\theta}) = \widehat{\theta}^2[\widehat{V}_{\beta\beta}/(\hat{\beta}^Y)^2 + \widehat{V}_{\mu\mu}/\hat{\mu}_C^2 - 2\widehat{V}_{\beta\mu}/(\hat{\beta}^Y\hat{\mu}_C)]$.) The approximation requires the true μ_C to be bounded away from zero, and its finite-sample quality is governed by the denominator's t-ratio: as $\hat{\mu}_C/\widehat{V}_{\mu\mu}^{1/2}$ falls, the exact distribution of the ratio becomes heavy-tailed and skewed, and normal-based intervals undercover.

Fieller construction. Define $\hat{\sigma}^2(\theta_0) = \widehat{V}_{\beta\beta} - 2\theta_0\widehat{V}_{\beta\mu} + \theta_0^2\widehat{V}_{\mu\mu}$, the estimated variance of the linear combination $\hat{\beta}^Y - \theta_0\hat{\mu}_C$. Under $H_0(\theta_0)\colon \beta^Y - \theta_0\mu_C = 0$, the statistic $T(\theta_0) = (\hat{\beta}^Y - \theta_0\hat{\mu}_C)/\hat{\sigma}(\theta_0)$ is asymptotically standard normal, since it tests a linear restriction on jointly asymptotically normal coefficients; with few clusters, critical values from the t_{G-1} distribution are the usual finite-sample refinement. The set (14) is $\{\theta_0 : T(\theta_0)^2 \leq z^2\}$ with $z = z_{1-\alpha/2}$; expanding,

$$a\,\theta_0^2 + b\,\theta_0 + c \;\leq\; 0, \qquad a = \hat{\mu}_C^2 - z^2\widehat{V}_{\mu\mu}, \quad b = 2\big(z^2\widehat{V}_{\beta\mu} - \hat{\beta}^Y\hat{\mu}_C\big), \quad c = (\hat{\beta}^Y)^2 - z^2\widehat{V}_{\beta\beta}. \tag{18}$$

Let $\Delta = b^2 - 4ac$. Three cases arise. (i) If $a > 0$, the set is a nonempty bounded interval, $\big[(-b - \sqrt{\Delta})/(2a),\ (-b + \sqrt{\Delta})/(2a)\big]$; it is nonempty because $\theta_0 = \widehat{\theta}$ always satisfies (14), the left side vanishing there. (ii) If $a < 0$, the set is the complement of a bounded open interval when $\Delta > 0$ and the entire real line when $\Delta \leq 0$. (iii) If $a = 0$, the set is a half-line. Case (i) obtains precisely when the complier base rate is significantly different from zero at level α. The unbounded sets in cases (ii)–(iii) are informative rather than pathological: they report that the sample cannot bound the proportional effect in at least one direction, a deficiency that no reformulation of the point estimate repairs. Because (14) inverts an asymptotically valid test of a linear restriction, its coverage does not deteriorate as $\mu_C \to 0$,

in contrast with (13) and with percentile bootstrap intervals; this is the ratio analogue of weak-instrument-robust inference (Staiger and Stock, 1997).

Paired bootstrap. For $b = 1, \ldots, B$ (we use $B = 1{,}000$): (i) resample G clusters with replacement, matching the original design's clustering (observations, if unclustered; in two-sample designs, each sample independently unless a common first stage cancels from the ratio); (ii) re-estimate the numerator and denominator components on the paired draw, holding the specification—instruments, controls, fixed effects, bandwidths, weights—fixed: use both 2SLS regressions ordinarily, or the two reduced forms when a common first stage cancels; (iii) record $\hat{\beta}^{Y,(b)}$, $\hat{\mu}_C^{(b)}$, and $\widehat{\theta}^{(b)} = \hat{\beta}^{Y,(b)}/\hat{\mu}_C^{(b)}$, or the corresponding reduced-form components and their identical ratio in the cancellation case. Report the standard deviation of $\widehat{\theta}^{(b)}$ or the percentile interval. Estimating both regressions on the same draw is what encodes the covariance term; drawing independent resamples for the two regressions is incorrect. As a diagnostic, we recommend reporting the share of draws in which $\hat{\mu}_C^{(b)}$ is negative or within two bootstrap standard errors of zero. When that share is nonnegligible, the empirical distribution of $\widehat{\theta}^{(b)}$ is heavy-tailed, percentile intervals become unstable—the application in Section 5.5 is an example—and the Fieller set should be reported instead. Section 6 examines the finite-sample coverage of all three procedures.

C Simulation Evidence: Designs and Additional Results

This appendix reports the simulation designs and additional results referenced in Section 6. Simulations serve three roles: (i) quantify finite-sample behavior of $\widehat{\theta}$ and coverage of the inference procedures in Section 3.6; (ii) illustrate numerically the properties that distinguish CAPE from conventional scalings in settings where those properties have content (multi-instrument bounded-range behavior and unit invariance with outcomes that have mass at zero); and (iii) check robustness of the implementation machinery in designs for which we do not derive closed forms. Throughout, population mappings that admit closed-form expressions in the paper are computed analytically (not by Monte Carlo); simulations are used to study sampling behavior and to illustrate comparative statics.

C.1 A monotone binary-IV data-generating process

To ensure that the maintained IV assumptions hold by construction, we generate binary potential treatment states using a single latent uniform U_i per unit. Let $Z_i \in \{0, 1\}$ be an instrument drawn independently of all latent variables, and let D_i and Y_i be binary. Define

potential treatment and potential outcome indicators by threshold crossing:

$$D_i(z) = \mathbf{1}\{\, U_i \leq \Lambda\big(a + \gamma_i + \delta_i z + \Omega_i\big)\,\}, \tag{19}$$

$$Y_i(d) = \mathbf{1}\{\, V_i \leq \Lambda\big(\alpha_i + \beta_i d + c\,\Omega_i\big)\,\}, \tag{20}$$

where $\Lambda(\cdot)$ is the logistic CDF, $U_i, V_i \overset{iid}{\sim} U(0,1)$ and are independent of each other and of $(\alpha_i, \beta_i^*, \gamma_i, \delta_i^*, \Omega_i)$. The latent indices are jointly normal:

$$(\alpha_i, \beta_i^*, \gamma_i, \delta_i^*, \Omega_i) \sim \mathcal{N}(0, \Sigma),$$

with Σ chosen to deliver target complier shares and target gaps described below. We set $\delta_i = \Phi(\delta_i^*) \geq 0$ and $\beta_i = \Phi(\beta_i^*) \geq 0$ to ensure non-negative first-stage and treatment responses.

The coupling via a single U_i is what enforces monotonicity unit-by-unit: since $\delta_i \geq 0$, we have $\Lambda(a + \gamma_i + \delta_i + \Omega_i) \geq \Lambda(a + \gamma_i + \Omega_i)$, hence $D_i(1) \geq D_i(0)$ for every unit. By contrast, drawing $D_i(0)$ and $D_i(1)$ as independent Bernoulli variables from their marginal probabilities would generate defiers with positive probability even when the first stage is positive in expectation. The outcome equation is written in terms of potential outcomes $Y_i(0)$ and $Y_i(1)$; the common V_i couples these in a transparent way within the latent-index representation.

C.2 Exercise B: finite-sample behavior and inference coverage

Design. Exercise B uses a single binary instrument with binary treatment and binary outcome. We assign compliance types using the coupled-uniform design above, with type shares and type-level means set directly at the following targets:

- Always-taker share $\pi_a = 0.15$.
- Complier share $\pi_c \in \{0.05, 0.10, 0.25, 0.50\}$.
- Complier untreated mean $m_c \in \{0.03, 0.10, 0.30\}$.
- Never-taker untreated gap $g_n = 0.5$ (a calibration target for the untreated mean difference between never-takers and compliers).
- True proportional-effect parameter $\theta = 0.5$.
- Sample sizes $N \in \{500, 2{,}000, 10{,}000\}$.

The executed cells are the direct-control special case of the latent-index design above: type shares and type-level untreated means are set directly at the targets (always-takers π_a; compliers π_c; never-takers the remainder; complier untreated mean m_c; never-taker mean $m_c(1+g_n)$; complier treated mean $m_c(1+\theta)$), rather than induced through Σ. This makes the targets exact by construction while preserving the threshold-crossing coupling—and hence monotonicity—unit by unit. The grid fully crosses π_c, m_c, and N (36 cells); the $\pi_c = 0.05$ cells double as a deliberately weak-first-stage arm. A clustered configuration supplements the grid: $G = 60$ clusters of $n_g = 50$ units, cluster-level assignment of Z, a cluster random effect shifting all outcome propensities, cluster resampling for the bootstrap, and cluster-robust variances for the delta and Fieller constructions (500 replications, 399 resamples). For each cell, we draw 2,000 Monte Carlo replications. Within each replication, we compute (i) the CAPE point estimate $\widehat{\theta}$ and (ii) three nominal 95% intervals from Section 3.6: a paired percentile bootstrap (unit-level resampling in the i.i.d. cells; cluster-level in the clustered configuration) , a joint delta-method interval using the stacked-system covariance, and a Fieller confidence set. For the bootstrap, we use $B = 999$ resamples per replication (odd B yields exact percentile ranks); the results are essentially unchanged with $B = 1{,}000$.

Implementation note. In this single-binary-instrument exercise we use a saturated/no-covariate specification, so each 2SLS coefficient coincides with the corresponding Wald ratio. We summarize "denominator strength" by the t-statistic for the estimated complier base rate,

$$t_{\text{den}} \equiv \frac{|\widehat{\mu}_C|}{\text{se}(\widehat{\mu}_C)}, \qquad \widehat{\mu}_C \equiv -\widehat{\beta}_{2SLS}^{Y(1-D)}.$$

Fieller sets are bounded precisely when $\widehat{\mu}_C$ is statistically distinguishable from zero at the stated level (using the same critical value as in the Fieller construction).

What we report. Table C.1 reports selected cells of the full grid, sorted by the median t_{den}. The full grid is provided in `results_exB.csv`. We report median percent bias (median of $100(\widehat{\theta} - \theta)/\theta$) rather than mean bias because the mean can be unstable or undefined in weak-denominator cells (the denominator can be very close to zero in finite samples). Coverage is the empirical fraction of replications in which the nominal 95% interval contains the true θ.

Table C.1: Finite-Sample Bias and 95% Coverage by Denominator Strength

Selected cells of the Exercise B grid (complete results in `results_exB.csv`), sorted by the median denominator t-statistic t_{den}. Med. bias is the median of $100(\widehat{\theta}-\theta)/\theta$. Coverage columns are empirical coverage of nominal 95% intervals: percentile paired bootstrap, delta method, and Fieller. "Unb. F." is the percentage of draws with an unbounded Fieller set (the draws in which the base rate is not statistically distinguishable from zero at the stated level). Median lengths are reported for the bootstrap and for bounded Fieller sets; $\theta = 0.5$ throughout. Entries "—" for Med. bias mark cells where more than half the draws have pathological denominators, so even the median is not meaningful; the clustered row omits lengths, which are not comparable across resampling schemes.

π_c	m_c	N	t_{den}	Med. bias (%)	Boot	Delta	Fieller	Unb. F. (%)	Len. boot	Len. F.
0.05	0.03	500	0.66	—	92.2	63.2	95.5	95.0	12.26	2.49
0.05	0.03	2,000	0.68	-287.9	92.8	62.6	94.5	94.2	12.75	2.50
0.05	0.10	2,000	0.74	-246.5	94.3	67.3	95.0	93.2	13.76	2.79
0.10	0.10	2,000	0.86	-168.8	96.0	77.1	94.0	89.8	18.37	3.49
0.25	0.03	2,000	1.04	-97.5	97.0	84.2	95.3	85.2	23.16	3.67
0.10	0.10	10,000	1.55	-27.0	97.0	87.9	94.5	68.3	20.27	4.06
0.25	0.10	2,000	1.81	-10.4	97.3	89.2	94.7	55.8	13.84	3.83
0.50	0.10	500	2.12	-5.8	96.9	90.8	94.0	43.9	7.53	3.90
0.50	0.03	2,000	2.25	+1.6	97.9	90.0	95.6	38.5	6.75	4.05
0.25	0.30	2,000	3.63	+1.9	97.2	93.3	95.4	4.4	2.15	2.09
0.25	0.10	10,000	4.15	-2.4	96.6	93.3	94.7	2.0	1.89	1.89
0.50	0.03	10,000	4.98	-0.3	95.8	95.0	96.1	0.1	1.59	1.59
0.50	0.10	10,000	9.39	+0.6	94.8	94.3	95.0	0.0	0.74	0.75
0.50	0.30	10,000	17.96	+1.0	95.2	95.8	95.4	0.0	0.34	0.34
Clustered, $G = 60$, $n_g = 50$			1.50	-43.3	96.2	85.0	94.2	67.0	—	—

Findings. Four patterns emerge. (i) *Coverage degrades first for the delta method:* as t_{den} falls, normal approximations to the ratio become poor and delta-method intervals can substantially undercover. (ii) *The percentile bootstrap can "cover by widening":* when the denominator is weak, percentile intervals may maintain near-nominal coverage only by becoming extremely wide, providing little usable information about proportional magnitudes. (iii) *Fieller sets remain valid and communicate weakness transparently:* in the same weak-denominator region, Fieller sets achieve stable coverage while returning unbounded sets with high frequency, which is the correct statement that the data do not bound the proportional effect. (iv) *Bias is well summarized by medians rather than means:* when t_{den} is sufficiently large, median bias is small; when it is small, ratio pathologies dominate and mean-based summaries become unstable.

C.3 Exercise C: two instruments, bounded-range behavior, and partial vs. joint monotonicity

Exercise C illustrates properties that are specific to overidentified designs. We generate two binary instruments (Z_1, Z_2) that are mutually independent and satisfy partial monotonicity component-wise (Assumption 6 in the paper). We construct two complier subgroups: units whose treatment is shifted only by Z_1 and units shifted only by Z_2, plus always-takers and never-takers. Outcomes are strictly positive, with $Y(0)$ Gamma-distributed and subgroup-specific proportional effects $Y(1) = (1 + \theta_k)Y(0)$. The executed calibration is: subgroup effects $\theta_1 = 1.5$ and $\theta_2 = 0.6$; class shares 0.20 (compliers to Z_1), 0.20 (compliers to Z_2), 0.50 (never-takers), 0.10 (always-takers); complier-class untreated means 1.0 and never-taker mean 3.0; population CAPE = 1.05, inside the subgroup range $[0.6, 1.5]$.

This design is set up so that *Imbens–Angrist joint monotonicity fails* across certain multi-coordinate changes (e.g., between $(1, 0)$ and $(0, 1)$ some units move into treatment while others move out), while partial monotonicity holds holding one instrument fixed. In this environment, CAPE remains interpretable as a ratio of like-weighted averages under the maintained conditions in Section 3.2, and asymptotically lies in the range of subgroup proportional effects. Conventional ratios that scale by sample means need not lie in this range and can be far outside it when non-compliers have very different untreated levels.

Table C.2 reports (i) the mean of each proportional estimator across draws and (ii) the fraction of draws in which each estimator lies in the subgroup-effect range; it also reports the numerical verification that the decomposition weights recovered from the Y and $Y(1 - D)$ regressions coincide draw by draw—with a shared first stage, $\hat{\beta}_{2SLS} = \sum_k \hat{\omega}_k \widehat{W}_k$ with $\hat{\omega}_k \propto \hat{\pi}_k \widehat{\mathrm{Cov}}(Z_k, D)$ is an exact in-sample identity, and the maximum discrepancy between the two weight sets across all draws and outcomes is at machine precision. Full results are in `results_exC.csv`.

Table C.2: Bounded-Range Behavior in a Two-Instrument Design

"In range" is the percentage of draws in which each ratio lies in the subgroup range $[\min_k \theta_k, \max_k \theta_k]$. CAPE is asymptotically range-contained under the conditions in Section 3.2; in finite samples, denominator noise can push any ratio outside the range.

	Mean of ratio			In range (%)		
N (draws)	CAPE	Untr.-mean	Full-mean	CAPE	Untr.-mean	Full-mean
4,000 (2,000 draws)	1.156	0.432	0.475	69.6	4.4	12.1
20,000 (1,000 draws)	1.070	0.433	0.476	97.4	0.0	0.2
Weight match, max discrepancy across all draws and outcomes: 1.2×10^{-14}						

C.4 Exercise D: mass at zero, continuous instrument, and unit invariance

Exercise D illustrates unit invariance in a continuous-instrument environment with outcomes that have mass at zero. We draw $Z \sim U(0,1)$ and generate a continuum of complier margins via thresholds z_i^* with $D_i(z) = \mathbf{1}\{z \geq z_i^*\}$. We allow margin heterogeneity in both untreated outcome levels and treatment effects, including a substantial probability mass at $Y(0) = 0$. The analytic target for CAPE is computed by numerical integration using the Angrist–Imbens (1995) weights (as in Section 3.3 and Appendix A of the paper) , so the simulation checks sampling behavior against the population target rather than against the estimator itself. In the executed design ($Z \sim U(0,1)$; complier thresholds $z^* \sim \text{Beta}(2,2)$; $\pi_a = 0.10$, $\pi_n = 0.30$; untreated zero probability $0.6 - 0.2z^*$, else Poisson with mean $1 + 2z^*$; treatment halves the zero probability and scales counts by 1.8), across 500 replications at $N = 20{,}000$, CAPE averages 1.684 (s.d. 0.124) against the analytic target of 1.675.

We also demonstrate unit dependence of log-based "percent effects" when zeros are present. For the same underlying outcome measured in different unit conventions (implemented as $\log(1 + cY)$ for c varying across scales), the implied effect $e^{\widehat{\beta}} - 1$ can change substantially with c: for $c \in \{0.1, 1, 10\}$ it averages 13.8%, 66.9%, and 176.4% respectively—a factor of nearly 13 across unit conventions.

The same design also verifies the same-IV complier-characteristics diagnostic at the continuous margin. For a predetermined characteristic constructed as $X_i = (z_i^*)^2$ among compliers (with constant values for always- and never-takers), the transformed-outcome regression with dependent variable $X_i(1 - D_i)$ returns 0.2854 (s.d. 0.009 across replications) against the analytic value $\int (z^*)^2 \, \omega(z^*) \, dz^* = 2/7 \approx 0.2857$ implied by the Angrist–Imbens weights. Because the target is computed from the closed-form weight formula rather than from the estimator itself, the check validates the diagnostic's weighting claim—that the recovered complier mean aggregates margins with exactly the first-stage weights—and not merely its internal consistency.

C.5 Optional robustness checks (designs without closed forms)

As a robustness check on implementation, we evaluate two designs for which we do not derive closed forms in the paper. First, under non-saturated covariate controls where compliance type is a deterministic function of a covariate (a hard case related to Blandhol et al. (2022)), we quantify the deviation between the transformed-outcome regression and the true complier untreated mean: across cells (complier shares 0.15 and 0.50; selection gaps +1.0 and −0.6; $N = 2$ million per cell), deviations range from −1.0% to +1.2% of the true base.