The Explanatory Power of Causal Effects

James Stratton ⓡ Nicolaj Thor*

June 24, 2026

Abstract

How much of the observed variation in outcome Y does variable X *causally* explain? We propose a causal R^2 (CR^2) to answer this question and quantify the importance of X in determining Y. CR^2 is the fit of a causal model, identified from observational plus experimental data, and captures the reduction in Y's variance from eliminating X's average causal effect. In applications, spring protection explains 31% of water quality variation among Kenyan springs; class size predicts 8% of variation in reading scores but explains only 3%; and sodium raises blood pressure similarly across genders, but explains far less among women.

*Emails: jstratton@g.harvard.edu and thor@brown.edu. We are grateful for comments from Adamson Bryant, Raj Chetty, Cole Davis, Cameron Deal, Joseph Eisenhauer, John Friedman, Peter Hull, Kosuke Imai, Larry Katz, Toru Kitagawa, Soonwoo Kwon, Florian Mudekereza, Sendhil Mullainathan, Mandy Pallais, Yechan Park, Jonathan Roth, Nico Rotundo, Jesse Shapiro, Elie Tamer, Winnie van Dijk, and audiences at Brown University, Harvard University, the University of Warwick, the Warren Alpert Medical School, and the 2025 American Causal Inference Conference. We are also grateful to have participated in the Alexander and Diviya Magaro Peer Pre-Review Program at Harvard's Institute for Quantitative Social Science.

1 Introduction

Social scientists often seek to explain observed variation in outcomes. What are the "causes of inequality in incomes" (Mincer 1958)? Why do some cities grow while others stall (Glaeser 2011)? And what are the "fundamental causes of the large differences in income per capita across countries" (Acemoglu, Johnson, and Robinson 2001)? Suppose a researcher, in answer to such a question, proposes that X is a main cause of observed variation in Y. Making and assessing this claim requires a measure of the share of variation in Y causally explained by X. We propose such a measure, and show its usefulness in economic settings.

The R^2 would suffice if by "explain" we meant "predict". Yet economists rarely use the R^2 because we typically seek causal, not predictive, explanations. Ordinary fit statistics such as the R^2 use only the joint distribution of Y and X and hence cannot untangle correlation from causation.

In contrast, causal inference measures how X *affects* Y (*e.g.*, how much an extra year of school raises wages), not whether *observed variation* in Y is explained by X (*e.g.*, the share of wage variation explained by schooling). These two concepts are related, but different. If X has no causal effect, it cannot causally explain variation in Y. Yet if X causally affects Y, it can explain a large or small share of variation in Y, depending on its variance and covariance with other variables affecting Y. To see the distinction in an extreme case, if all workers had the same education, education would explain no variation in wages, even with a large effect.[1]

We measure the share of variation in Y causally explained by X in three steps: first, estimate the causal effect of X on Y; second, compute the "best causal model" for Y, which minimizes population mean squared error (MSE), under the constraint that modelled differences in Y match X's causal effect; third, evaluate the fit of this model.

Our approach parallels the R^2. Recall that the (non-parametric) R^2 is the fit of the conditional expectation function, $m(x) \coloneqq \mathbb{E}[Y \mid X = x]$, the best predictive model for Y: $R^2 = 1 - \text{MSE}[m] \,/\, \text{Var}[Y]$. Instead, we seek a *causal* model, μ, which we define as a model constrained to match the average treatment effects of X. The best causal model $\mu^\star$ minimizes MSE subject to this constraint. This model captures the relation between X and Y after purging confounding: modelled differences in Y are caused

[1] As we later explain, even when X varies a lot in the population *and* has a large causal effect, it may account for little of the observed cross-sectional variation in Y.

by X and not by omitted variables. We define the **causal R-squared** (CR^2) as the fit of that model: $CR^2 = 1 - \text{MSE}[\mu^\star] / \text{Var}[Y]$.

We interpret CR^2 as the share of variance causally explained by X. It is also the reduction in outcome variance if X had no causal effect. It equals 0 if X has no causal effect or does not vary in the population, and 1 if X fully determines Y. It is unitless and bounded above by the predictive R^2, with equality if observables are independent of unobservables. Identification requires (i) an *experimental dataset* to identify causal effects, and (ii) an *observational dataset* to estimate the joint distribution of (Y, X). Our baseline setting for (i) is a randomized experiment, but the approach extends to quasi-experiments that recover an average treatment effect.

We illustrate CR^2 in three applications. Using data from Kremer et al. (2011), we find that variation in spring protection *causally explains* about a third of variation in water quality among Kenyan springs. In Project STAR, class size *predicts* 8% of variation in reading scores, but *causally explains* only 3%: the predictive power of class size mostly comes from omitted variables. Lastly, salt intake has similar causal effects on blood pressure for men and women, but causally explains 7% of variation among men, and under 1% in women.

These applications, we hope, illustrate the usefulness of CR^2 in quantifying a variable's role in explaining observed variation. The CR^2 also helps gauge the value of future research: if the explained share is low, our theory of the outcome is incomplete. CR^2 is primarily a measure of explanation, not a guide to policy counterfactuals. For instance, CR^2 assesses whether differences in education are a main cause of income gaps. However, a policymaker seeking to raise incomes might only care about the causal effect of education, even if it explains little *existing* variation (see Goldberger 1979). That said, we show that a policymaker interested in reducing income *inequality* may use CR^2 to assess how much inequality would fall if a given variable had, on average, no causal effect—or, equivalently, if a government policy compensated people for inequality caused by that variable.

Our approach relates most closely to the literature on "causes of effects", which asks whether a given unit's outcome would have differed under another treatment (*e.g.*, Pearl 1999; Dawid and Musio 2022); we instead ask what share of *population* variance comes from a given treatment. CR^2 is the performance of a model constrained to be causal, and so also recalls work on completeness, which compares theory-constrained

vs. unconstrained models (*e.g.*, Fudenberg et al. 2022). Recent work uses the predictive R^2 in novel ways: *e.g.*, to bound omitted variable (Oster 2019; Cinelli and Hazlett 2020, 2025) or external validity bias (Andrews and Oster 2019). We instead measure the fit of *causal* models.

Finally, Gelman and Imbens (2013) study "reverse causal questions", which involve assessing how well a model explains the outcome:

> [I]f we ask, Why do incumbents get more contributions than challengers...get some measure for candidate quality...and still see a large and statistically significant difference between the funds given to incumbents and challengers, then it seems we need more explanation. (p. 3)

The causal R^2 formalizes this idea: it measures how well a causal model explains the outcome, and how much remains unexplained.

Outline. Section 2 introduces the main ideas through an example. Section 3 defines CR^2. Section 4 discusses properties. Section 5 covers estimation. Section 6 presents applications.

2 An illustrative example

We first illustrate our approach in a simple example relating test scores to class sizes. Let Y_i stand for student i's test score, C_i class size, and U_i other factors affecting scores. We begin with a constant effects model:

$$Y_i = \alpha + \beta C_i + \gamma U_i, \tag{1}$$

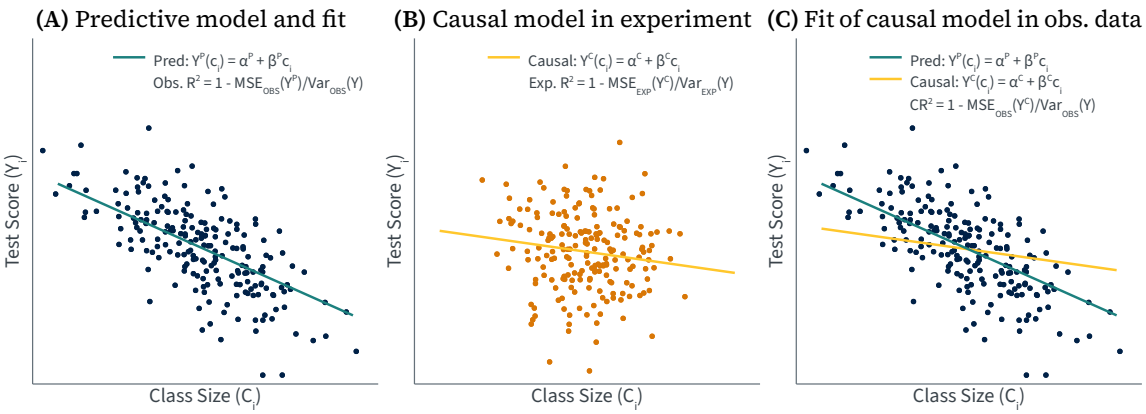
$$\begin{pmatrix} C_i \\ U_i \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \quad \mu = \begin{pmatrix} \mu_C \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{2}$$

where (1) is the structural equation for Y_i, and (2) is the joint distribution of C_i and U_i. Scores and class size are observable; U_i is not.

What share of variation in test scores is explained by class size? If we construe this question *predictively*, we can answer it with *observational data*. Figure 1(A) shows example realizations (Y_i, C_i). The line of best fit converges, as the sample grows, to the conditional expectation $Y^P(c_i) \coloneqq \mathbb{E}[Y_i \mid C_i = c_i] = \alpha^P + \beta^P c_i$, for $(\alpha^P, \beta^P) \coloneqq (\alpha - \gamma\rho\mu_C, \beta + \gamma\rho)$. As is well-known, Y^P is the best predictive model: it minimizes MSE. Its fit—the standard R^2—is the share of variation in scores that class size *predicts*.

Nonetheless, R^2 does not capture how much class size *causally explains*: R^2 can be high if class size has no causal effect ($\beta = 0$) but is strongly correlated with unobservable determinants of the outcome

Figure 1. Example: fit of predictive and causal models relating test scores to class size



Notes: This figure illustrates predictive and causal models relating test scores (Y_i) to class size (C_i). Panel (A) shows observational data. The line of best fit is the conditional expectation function (the best predictive model). Its fit is the predictive R^2. Panel (B) shows experimental data. The line of best fit is the average potential outcome function (the best causal model). Its fit in the experimental data is the "experimental R^2". Panel (C) adds the best causal model to Panel (A). Its fit in observational data is the causal R^2.

(ρ, γ large in magnitude). More broadly, since observational data do not identify C's causal effect, they cannot identify variance *causally* explained.

Suppose the analyst also conducts an experiment that draws students from the population, randomly assigns class sizes, and records the ensuing test scores. Figure 1(B) presents example realizations. Since class size is randomized, the best fit line converges to the average potential outcome $\mathbb{E}[Y_i(c_i)] = \alpha + \beta c_i$. We call this best fit line a *causal model* Y^C. It typically differs from the predictive model Y^P because of omitted variable bias: in Figure 1(B), Y^C is flatter than Y^P, consistent with positive selection into smaller classes ($\rho < 0$).

Since the best causal model recovers the causal effect β, it is tempting to interpret the R^2 in the experiment as the share of variation causally explained. Unfortunately, this "experimental R^2" reflects the share explained in the experiment, not the population. These quantities differ as experimental treatment shares are typically chosen to maximize precision, not to match the population (Neyman 1934; Athey and Imbens 2017). Even when they match the population, the *outcome distribution* differs: Y is a function of C and U, and $C \perp\!\!\!\perp U$ in the experiment, whereas they can covary in the population.

Thus, observational data identify the distribution of (Y, C), but not C's causal effect; experimental data identify the causal effect of C, but not its

explanatory power. We therefore measure the share of variance causally explained by class size by estimating its causal effect in experimental data, and then evaluating the fit of the corresponding causal model in observational data (Panel (C)). We call this measure the causal R-squared, $\mathrm{CR}^2 = 1 - \mathrm{MSE}\big[Y^C\big] / \mathrm{Var}[Y]$.

To interpret CR^2, take a student with score y_i and class size c_i. The causal model assigns a score $\alpha^C + \beta c_i$; the residual $y_i - (\alpha^C + \beta c_i)$ is unexplained by class size. If scores differed *only* because of class size ($\gamma = 0$), this residual would be zero for all students: class size would fully explain variation. In practice, other differences between students generate residual variation. Summing these squared residuals and dividing by the outcome variance gives the share of variance not causally explained; the complement is the share explained.

In this example, CR^2 has a simple form. Write $\eta_i \coloneqq \gamma u_i$:

$$\mathrm{CR}^2 = 1 - \frac{\mathrm{MSE}\big[Y^C\big]}{\mathrm{Var}[Y]} = 1 - \frac{\mathrm{Var}[\eta]}{\mathrm{Var}[Y]}.$$

Thus, the CR^2 is one minus the share of variance that is causally unexplained by class size. By contrast:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{MSE}\big[Y^P\big]}{\mathrm{Var}[Y]} = 1 - (1-\rho^2)\frac{\mathrm{Var}[\eta]}{\mathrm{Var}[Y]}$$
$$\text{Exp. } \mathrm{R}^2 = 1 - \frac{\mathrm{MSE}_{\text{exp}}[Y^C]}{\mathrm{Var}_{\text{exp}}[Y]} = 1 - \frac{\mathrm{Var}[\eta]}{\mathrm{Var}_{\text{exp}}[Y]}.$$

Relative to CR^2, the predictive R^2 is inflated by the correlation ρ between C and U: CR^2 purges this confounding from omitted variables. The experimental R^2 depends on the experimental variance of the outcome, and hence cannot be interpreted as the *population* share explained.

3 Defining CR^2

We now formalize our approach and define CR^2 generally.

3.1 Setting

We use the standard Rubin Causal Model with two primitives: K feature variables (indexed by k) and a potential outcome function. Each *feature variable* X_k is a real-valued random variable taking values in $\mathcal{X}_k$. The feature vector is $X \coloneqq (X_1, \ldots, X_K)$ taking values in $\mathcal{X} \coloneqq \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$, with

joint law P_X. The feature variables and the *potential outcome function* $Y\colon \mathcal{X} \to \mathcal{Y}$ determine the real-valued *outcome* $Y := Y(X)$.

$Y(x_i)$ is the outcome for a *unit i* with realized features x_i. We treat the potential outcome function $Y(\cdot)$ as a *population* object, rather than defining a unit-specific $Y_i(\cdot)$. This is without loss, since X is defined to include all determinants of Y:[2] the effects of any X_k may be heterogeneous, but the sources of heterogeneity are other variables in X.

The potential outcome function $Y(\cdot)$ and the law of the features P_X together fix the joint law of the outcome and features $P_{Y,X}$. All expectations and variances are with regard to $P_{Y,X}$ unless otherwise specified.

Features may be observed or unobserved. Write the *observed features* as $X^O := (X_1, \dots, X_O)$ for $O \leq K$, and the *unobserved features* as $X^U := (X_{O+1}, \dots, X_K)$. Denote by P_{Y,X^O} the marginal distribution of (Y, X^O) induced by $P_{Y,X}$. We assume features are either observable *and* manipulable (X^O), or unobservable (X^U). In practice, researchers may have covariates: features which are observed, but not manipulated. We discuss incorporating covariates in Online Appendix B.

We assume the outcome and features have finite second moments, and the outcome has positive variance. For estimation, we assume finite fourth moments.

3.2 Predictive and causal models

We describe the relation between the outcome and observed features by a ***model***: a real-valued function of X^O. We define the ***risk*** of a model M as its quadratic loss, $\mathcal{R}(M) := \mathbb{E}\big[(M(X^O) - Y(X))^2\big]$. If we observe all features ($O = K$), a natural model is the potential outcome function $Y(\cdot)$, which achieves zero risk; we call $Y(\cdot)$ the *true model*. When no features are observed, the risk-minimizing model is $\mathbb{E}[Y]$, with risk $\mathrm{Var}[Y]$; we call this the *baseline model* $\bar{Y}$. When some, but not all, features are observed, the relevant model depends on whether the analyst's goal is prediction or causal explanation.

Definition 1. Given an outcome Y and observed features X^O, the ***best predictive model*** is $Y^P_{X^O} := \arg\min_M \mathcal{R}(M)$. Equivalently, $Y^P_{X^O}(X^O) = \mathbb{E}\big[Y \mid X^O\big]$ almost surely.

This best predictive model conflates causal and non-causal sources of association: $Y^P_{X^O}$ can change with X^O even when X^O has no causal effect.

[2] For a formal statement, see Online Appendix A.

A causal model purges this confounding. For any x^O, recall that the ***average potential outcome*** is $\mathbb{E}\big[Y(x^O)\big] \coloneqq \mathbb{E}_{P_{X^U}}\big[Y(x^O, X^U)\big]$. We say a model M is a ***causal model*** if, for any $x^O, \hat{x}^O \in \mathcal{X}^O$, we have $M(\hat{x}^O) - M(x^O) = \mathbb{E}\big[Y(\hat{x}^O)\big] - \mathbb{E}\big[Y(x^O)\big]$. A causal model is consistent with causal effects, in that modelled differences in Y match the average causal effect (ATE) of X^O.

Definition 2. Given an outcome Y and observed features X^O, the ***best causal model*** $Y^C_{X^O}$ is the causal model which minimizes risk:

$$Y^C_{X^O} \coloneqq \arg\min_M \mathcal{R}(M), \text{ s.t } M \text{ is a causal model.}$$

Equivalently, the best causal model is the average potential outcome function, re-centered to the expected value of the outcome:[3]

Lemma 1. *Every causal model M is the average potential outcome plus a constant: $M(x^O) = \mathbb{E}\big[Y(x^O)\big] + c$, for some constant $c \in \mathbb{R}$. The best causal model sets $c^* = \big(\mathbb{E}[Y] - \mathbb{E}\big[\mathbb{E}\big[Y(X^O)\big]\big]\big)$ such that the model is re-centered to the expected value of the outcome: $Y^C_{X^O}(x^O) = \mathbb{E}\big[Y(x^O)\big] + \big(\mathbb{E}[Y] - \mathbb{E}\big[\mathbb{E}\big[Y(X^O)\big]\big]\big)$.*

The best causal model solves the same risk-minimization problem as the best predictive model under the additional constraint that modelled differences in Y reflect X^O's causal effects, rather than mere association.

3.3 Fit of the best causal model: the CR^2

We now define goodness-of-fit.

Definition 3. The ***fit*** of a model M is its proportional reduction in risk relative to the baseline model:

$$G(M) = \frac{\mathcal{R}(\bar{Y}) - \mathcal{R}(M)}{\mathcal{R}(\bar{Y})} = \frac{\mathrm{Var}[Y] - \mathcal{R}(M)}{\mathrm{Var}[Y]}.$$

The baseline model has fit 0, the true model 1. The fit of the best predictive model is the familiar non-parametric R^2: $\mathrm{R}^2(X^O) \coloneqq G(Y^P_{X^O})$.

Definition 4. Given outcome Y and observed features X^O, the ***non-parametric causal*** R^2 is the fit of the best causal model:

$$\mathrm{CR}^2(X^O) \coloneqq G(Y^C_{X^O}).$$

[3] Re-centering ensures the best causal model is well-calibrated for the population, in that its expected value is the expected outcome.

We interpret CR^2 as the share of variance in Y causally explained by X^O. We show three different perspectives that make this interpretation precise.

First, this view mirrors the standard view of the R^2. The R^2, which is interpreted as the share of variance in Y predicted by X^O, is the complement of the fraction of variance unexplained (FVU): $\text{R}^2 = 1 - \text{Var}\big[Y - Y^P_{XO}\big] / \text{Var}[Y]$. Intuitively, for each unit (y_i, x_i^O), the best predictive model Y^P_{XO} leaves a residual $y_i - Y^P_{XO}(x_i^O)$; the R^2 compares the variance of these residuals with the variance of the outcome. The causal R^2 instead uses residuals from the best causal model. For each realization (y_i, x_i^O), the causal model leaves a residual $y_i - Y^C_{XO}(x_i^O)$ unexplained; CR^2 compares the variance of these residuals with the variance of the outcome.

Second, CR^2 measures the proportional reduction in outcome variance in a thought experiment that eliminates the average causal effect of X^O on Y, holding all else constant. In our view, that eliminated variance is causally explained by X^O: it does not exist but for the causal effect of X^O. Concretely, consider a government policy that compensates individuals for outcome differences arising from the causal effect of X^O: individual i with features x_i^O and pre-compensation outcome y_i has post-compensation outcome $y_i' = y_i - Y^C_{XO}(x_i^O)$. The ensuing outcome $Y' = Y - Y^C_{XO}$ has variance $\text{Var}\big[Y - Y^C_{XO}\big] = \mathbb{E}\big[(Y - Y^C_{XO})^2\big] - \mathbb{E}\big[Y - Y^C_{XO}\big]^2 = \mathcal{R}(Y^C_{XO})$. Hence, $\text{CR}^2 = [\text{Var}[Y] - \mathcal{R}(Y^C_{XO})]/\,\text{Var}[Y]$ yields the proportional reduction in variance from eliminating the average causal effect of X^O.

An alternative view decomposes Y as

$$Y = Y' + (Y^C_{XO}(X^O) - \mathbb{E}[Y]), \text{ for } Y' \coloneqq Y - (Y^C_{XO}(X^O) - \mathbb{E}[Y]).$$

The effect of X^O is captured in the second term $(Y^C_{XO}(x^O) - \mathbb{E}[Y])$; we interpret Y', which does not vary in expectation as we manipulate X^O, as the residual effect of other features. The causal R^2 is the percentage reduction in variance when moving from Y to Y': that is, the reduction in the outcome's variance when purged of the causal effects of the observed features.

Finally, say we ask how close X^O comes to fully causing variation in Y. One way to conceptualise this question is to ask what the outcome distribution would be if X^O were the only cause of Y: deviations from this benchmark must, at least partly, come from another cause.[4] If X^O

[4]One might instead consider also including heterogeneous causal effects of X^O.

were the *only* cause of Y, then unit i's outcome would equal the best causal model, $Y^C_{X^O}(x_i^O)$. The residual $y_i - Y^C_{X^O}(x_i^O)$ therefore measures the part of the outcome not explained by X^O, and CR^2 the explained share.

3.4 Parametric CR^2

In practice, researchers often estimate parametric models: for instance, the standard linear R^2 measures fit of a linear predictive model. We analogously define a parametric CR^2 for a function class $\mathcal{F} \subseteq L^2$ by first defining the best causal model under $\mathcal{F}$ as the $L^2(P_{X^O})$ projection of the unrestricted best causal model onto $\mathcal{F}$ (and likewise for the best predictive model under $\mathcal{F}$).[5] We assume these best models are unique. We call $\mathcal{F}$ ***well-specified*** if it includes $Y^C_{X^O}$, and ***misspecified*** otherwise.[6]

We define $\mathrm{R}^2_{\mathcal{F}}(X^O) \coloneqq G(Y^P_{\mathcal{F},X^O})$ as the fit of the best predictive model under $\mathcal{F}$, and $\mathrm{CR}^2_{\mathcal{F}} \coloneqq G(Y^C_{\mathcal{F},X^O})$ as the fit of the best causal model under $\mathcal{F}$.

When $\mathcal{F}$ is the class of linear functions, the best predictive model is the linear projection of Y on X^O, whose fit is the standard linear R^2. The best causal model is the (re-centered) linear projection of Y on X^O when X^O is randomly assigned according to its population marginal distribution. Define the ***linear causal*** R^2, $\mathrm{CR}^2_{\mathrm{lin}}(X^O)$, as the fit of this model. In the well-specified case, $\mathrm{CR}^2_{\mathrm{lin}}$ simply replaces the observational OLS coefficients $(\alpha^{\mathrm{OBS}}, \beta^{\mathrm{OBS}})$ in the predictive R^2 with their experimental counterparts $(\alpha^{\mathrm{EXP}}, \beta^{\mathrm{EXP}})$:

$$\mathrm{R}^2_{\mathrm{lin}}(X^O) = 1 - \frac{\mathbb{E}\big[(Y - \alpha^{\mathrm{OBS}} - (\beta^{\mathrm{OBS}})^\top X^O)^2\big]}{\mathrm{Var}[Y]},$$
$$\mathrm{CR}^2_{\mathrm{lin}}(X^O) = 1 - \frac{\mathbb{E}\big[(Y - \alpha^{\mathrm{EXP}} - (\beta^{\mathrm{EXP}})^\top X^O)^2\big]}{\mathrm{Var}[Y]}.$$

However, effects of X^O that vary by X^U are conceptually indistinguishable from effects of X^U that vary by X^O. We thus focus on the effect of X^O alone—a conservative view of X^O's causal power.

[5] We assume $\mathcal{F}$ includes all constant models, and that it contains two models M and M' which differ ($\mathbb{E}\big[(M(X_i^O) - M'(X_i^O))^2\big] > 0$), and that $\mathcal{F}$ is closed under addition of constants (if $\mathcal{F}$ includes a model M, it also includes $M + \alpha$ for any constant α).

[6] $\mathcal{F}$ is well-specified if it includes the best causal model $Y^C_{X^O}$ *given observables*, even if it omits relevant unobservables. For instance, say the effect of X^O on Y depends on unobserved X^U. If the average potential outcome given X^O is linear, the class $X^O \to \alpha + \beta X^O$ is well-specified, even though β averages over heterogeneity induced by X^U.

4 Properties of CR^2

We now discuss the properties of the causal R^2.

4.1 Basic properties

We begin with some basic properties needed to interpret CR^2 as the share of variance causally explained. Where the outcome is unclear, we specify it by writing $\mathrm{CR}^2_{\mathcal{F}}(X^O \to Y)$.

Proposition 1 (basic properties). *For any function class $\mathcal{F}$:*

(i) If X^O has no causal effect on the outcome (that is, $Y(x^O, x^U) = Y(x^{O\prime}, x^U)$ for all $(x^O, x^{O\prime}, x^U)$), then $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 0$.

(ii) If $\mathcal{F}$ is well-specified, and X^O fully determines the outcome (that is, $Y(x^O, x^U) = Y(x^O, x^{U\prime})$ for all $(x^O, x^U, x^{U\prime})$), then $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 1$.

(iii) If X^O does not vary, then $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 0$.

(iv) If $\mathcal{F}$ is well-specified, and X^U does not vary, then $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 1$.

(v) Say $Y' = Y + \varepsilon$, for ε unobserved mean-zero random noise independent of, and causally unaffected by, X. Then $|\mathrm{CR}^2_{\mathcal{F}}(X^O \to Y)| \geq |\mathrm{CR}^2_{\mathcal{F}}(X^O \to Y')|$.

(vi) $\mathrm{CR}^2_{\mathcal{F}}$ is not symmetric: for a single observed feature X^O, in general, $\mathrm{CR}^2_{\mathcal{F}}(X^O \to Y) \neq \mathrm{CR}^2_{\mathcal{F}}(Y \to X^O)$.

(vii) CR^2 is invariant to affine transformations of the outcome, and to strictly monotone transformations of the features. Similarly, $\mathrm{CR}^2_{\mathrm{lin}}$ is invariant to affine transformations of the outcome and features.

Parts (i)–(iv) are limiting cases. If X^O does not affect Y, or does not vary, it explains no variation: $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 0$. If instead X^U does not vary or X^O fully determines the outcome—such that no variation in Y remains after setting X^O—then $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 1$, provided $\mathcal{F}$ is well-specified.

Part (v) is a comparative static: weakening the relation between the outcome and observed features by adding noise attenuates explanatory power. Part (vi) reflects that causal relations are not symmetric: X^O may cause Y, but not *vice versa*. Part (vii) says that CR^2 is unit-invariant: education explains the same share of variation in wages, no matter whether wages are measured in dollars or cents.

We consider these basic properties necessary for a measure of causally explained variation. Some alternatives do not satisfy them. For instance, the predictive R^2 generally violates (i) and (vi). Similarly, normalized effect sizes, such as Cohen's (1988) d, cannot be interpreted as the share of variation causally explained—they may, for example,

conclude that a variable explains over 100% of variation. Online Appendix C discusses other seemingly intuitive measures, and shows they lack one or more of the basic properties.

4.2 Comparison of R^2 and CR^2

The best predictive and causal models solve the same risk-minimization problem, but the best causal model is constrained to match causal effects; as a result, predictive R^2 weakly exceeds causal R^2.

Proposition 2 (relation between predictive and causal R^2). *For any function class $\mathcal{F}$, $CR^2_{\mathcal{F}}(X^O) \leq R^2_{\mathcal{F}}(X^O)$, with equality when $X^O \perp\!\!\!\perp X^U$.*

This result has practical use: even without knowing the causal effect of a variable, we can conclude that it has little causal explanatory power from the fact that it has little predictive power. The predictive and causal R^2 coincide when observables are independent of unobservables: without confounding, the best predictive model equals the best causal model, and hence they have the same fit.

Unlike the predictive R^2, CR^2 may be negative and non-monotonic.

Proposition 3 (possibility of negativity and non-monotonicity). *For any function class $\mathcal{F}$:*

(i) *$R^2_{\mathcal{F}}$ is bounded between 0 and 1, whereas $CR^2_{\mathcal{F}}$ is bounded above by 1, but may be negative.*

(ii) *$R^2_{\mathcal{F}}$ increases as more features are observed, whereas $CR^2_{\mathcal{F}}$ may fall.*

Intuitively, CR^2 is negative when the observables *suppress*, rather than create, variance in Y. Consider the introductory example. Suppose smaller classes causally increase scores, but struggling students sort into small classes, strongly enough that class size and scores are *positively* correlated. Then the best causal model, which predicts a negative relation, fits the population data *worse* than the baseline constant model, so CR^2 is negative. A similar phenomenon arises in Kitagawa–Oaxaca–Blinder decompositions: the share of the gender wage gap "explained" by college is negative because women complete college at higher rates, while college predicts higher wages (Blau and Kahn 2017). Intuitively, college *suppresses* the wage gap—it contributes negatively. Non-monotonicity has the same intuition: when no features are observed, CR^2 is zero, but it may be negative when some features are

added.[7]

4.3 Properties of $\mathrm{CR}^2_{\mathrm{lin}}$

The linear CR^2 is of applied interest because researchers often use linear models. It has a simple expression in terms of summary statistics.

Proposition 4 (summary statistics expression). *If $\mathcal{F}_{\mathrm{lin}}$ is well-specified:*

$$\mathrm{CR}^2_{\mathrm{lin}}(X^O) = \mathrm{R}^2_{\mathrm{lin}}(X^O) - \frac{1}{\mathrm{Var}[Y]}(\beta^P_{XO} - \beta^C_{XO})^\top \Sigma_{XO}(\beta^P_{XO} - \beta^C_{XO}),$$

where β^P_{XO} (β^C_{XO}) are the coefficients from an OLS regression of Y on X^O in observational (experimental) data, and Σ_{XO} is the observational covariance matrix of observed features.

With a single observed feature, $\mathrm{CR}^2_{\mathrm{lin}}(X^O) = \mathrm{R}^2_{\mathrm{lin}}(X^O) - \frac{\mathrm{Var}[X^O]}{\mathrm{Var}[Y]}(\beta^P_{XO} - \beta^C_{XO})^2$. That is, the difference between the predictive and causal linear R^2 is the squared difference between observational and experimental slopes, standardized to observational data. Hence, if the model is well-specified, $\mathrm{CR}^2_{\mathrm{lin}}$ can be computed from summary statistics without underlying microdata.

5 Identification, estimation, and inference

We now turn to estimation.

5.1 Data setting

The CR^2 is defined from the joint distribution of (Y, X^O), and the causal effects of X^O. The joint distribution is easy to obtain from observational data, but causal effects are not. For simplicity, our baseline is identification of causal effects via a randomized experiment; section 5.3 discusses quasi-experiments.

The analyst has two samples. The first is an observational sample consisting of N_O units (Y_i, X^O_i) drawn i.i.d. from the population. This sample thus identifies P_{Y,X^O}. To identify the causal effects of X^O, the analyst relies on an experiment: she draws an i.i.d. sample of size N_E from the population, independently randomizes each unit across T

[7] Online Appendix C shows there are no other monotone measures of the share of variation causally explained. Non-monotonicity is a constitutive property.

treatment arms, and observes the resulting outcome. Treatment arm t is assigned with probability $p_t > 0$, and sets the observed features to $x_t^O \in \mathcal{X}^O$.

The total sample size is $N = N_O + N_E$. Formally, the data is drawn from a superpopulation that samples the observational distribution with probability p and the experimental distribution with probability $1 - p$.

This data combination arises in several settings. The analyst may conduct an experiment, and separately draw an observational sample from the same population: over 30% of experimental papers in journals of the American Economic Association from 2015 to 2025 include observational evidence (Epanomeritakis and Viviano 2025). Alternatively, the experiment may include a "status quo" arm, which may be treated as observational data. Finally, in some cases a researcher has observational data with as-if random variation in the features in a subsample; this subsample can be used as the experimental data, and the broader sample can be used as the observational data.

5.2 Identification

We now turn to identifying CR^2. An experiment is ***full-rank*** if $(1, x_t^O)_{t=1}^T$ has full rank, and ***full-support*** if every $x^O \in \operatorname{supp} P_{X^O}$ appears as some treatment arm x_t^O.

Proposition 5 (identification). *Fix an outcome Y and observables X^O. Let s_i indicate i's sample.*

(i) For any $\mathcal{F}$, $\mathrm{CR}^2_{\mathcal{F}}(X^O)$ is not identified by an observational ($p = 1$) or experimental ($p = 0$) sample alone.

(ii) The non-parametric causal R^2, $\mathrm{CR}^2(X^O)$, is identified by a combination of observational and experimental samples ($p \in (0, 1)$), if and only if the experiment is full-support.

(iii) Under a well-specified linear model, the linear causal R^2, $\mathrm{CR}^2_{\mathrm{lin}}(X^O)$, is identified by a combination of observational and experimental samples ($p \in (0, 1)$) if and only if the experiment is full-rank.

(iv) Under the conditions for identification in (ii) or (iii), the following

plug-in estimator is consistent as N grows, holding p fixed:

$$\widehat{\mathrm{CR}}^2_{\mathcal{F}}(X^O) \coloneqq 1 - \frac{\widehat{\mathcal{R}}\left(\hat{Y}^C_{\mathcal{F},X^O}\right)}{\widehat{\mathrm{Var}}[Y]} \quad \textit{where}$$

$$\widehat{\mathrm{Var}}[Y] \coloneqq \frac{1}{N_O}\sum_{i:s_i=O}\left(y_i - \bar{y}_O\right)^2, \qquad \bar{y}_O \coloneqq \frac{1}{N_O}\sum_{i:s_i=O} y_i,$$

$$\hat{m}^C \coloneqq \arg\min_{M\in\mathcal{F}} \frac{1}{N_E}\sum_{i:s_i=E}\left(y_i - M(x_i^O)\right)^2, \qquad \hat{c}^* \coloneqq \bar{y}_O - \frac{1}{N_O}\sum_{i:s_i=O}\hat{m}^C(x_i^O),$$

$$\widehat{\mathcal{R}}(M) \coloneqq \frac{1}{N_O}\sum_{i:s_i=O}\left(y_i - M(x_i^O)\right)^2, \qquad \hat{Y}^C_{\mathcal{F},X^O} \coloneqq \hat{m}^C + \hat{c}^*.$$

Since the non-parametric CR^2 imposes no structure on treatment effects, identification requires manipulating X^O over its full support. This is plausible only if $\mathrm{supp}(X^O)$ is small. The linear CR^2 requires only that the experiment independently manipulates each feature.

For $\mathcal{F}$ linear, the plug-in estimator amounts to (i) regressing Y on X^O in the experimental data to obtain $\hat{m}^C$, the best linear approximation to the potential outcome function; (ii) re-centering; and (iii) computing the MSE relative to outcome variance in the observational data. Though consistent, $\widehat{\mathrm{CR}}^2_{\mathcal{F}}$ is generally downward-biased in small samples, since estimation error in the best causal model tends to inflate the model's risk.[8] Simulations in Online Appendix D show this bias vanishes reasonably quickly as N grows. Inference follows from the Delta method or bootstrap by standard arguments (see Online Appendix D).

5.3 External validity

We have assumed the experiment randomly samples from the population. In practice, some experiments and quasi-experimental designs identify a local average treatment effect (LATE) only among a subpopulation (*e.g.*, a standard instrumental variables design with a binary treatment identifies the LATE among compliers). Identifying CR^2 then requires a transportability assumption: the causal effects in the experiment must coincide with those in the population.

This is the familiar problem of generalizing a LATE or TOT to an ATE, and standard tools apply. One approach is to compare observable characteristics of the experimental and target populations. In a randomized experiment, the experimental subpopulation is observed

[8] In contrast, R^2 is upward-biased due to overfitting.

directly; in instrumental variables designs, Abadie's (2003) weighting theorem identifies the distribution of complier covariates. If these distributions are similar, we may have more confidence in generalizing the causal model. If they differ, the analyst can compute covariate-specific LATEs and reweight them to recover the ATE, under the assumption that effect heterogeneity is fully captured by observables (Angrist and Fernández-Val 2013; Hartman et al. 2015). The analyst may also assess sensitivity to omitted moderators of treatment effects (Huang 2024). Finally, sometimes, structural latent-index models can extrapolate causal effects beyond compliers (*e.g.*, Heckman, Tobias, and Vytlacil 2001, 2003; Heckman 2010; Angrist 2004).

6 Applications

We illustrate CR^2 in three settings, leaving details to Online Appendix E.

6.1 Water quality and spring protection

Our first application draws on the RCT in Kremer et al. (2011) evaluating the effects of spring protection on water quality, defined as the *Escherichia coli* level. Spring protection covers a spring's source in concrete so water flows through a pipe rather than the ground. We ask what share of variation in water quality across Kenyan springs is causally explained by variation in spring protection.

The experimental sample consists of springs in the authors' experiment. The observational sample consists of nearby springs not in the experiment. Protection status is binary. We assess the predictive relation in observational data through a regression at the spring (s) level:

$$\ln \text{E coli}_s = \alpha^P + \beta^P \text{Protection}_s + \epsilon_s^P, \tag{3}$$

We assess the causal relation through a regression in experimental data:

$$\ln \text{E coli}_s = \alpha^C + \beta^C \text{Protection}_s + \epsilon_s^C. \tag{4}$$

The estimated causal effect determines our best causal model,

$$\widehat{\ln \text{E coli}_s}^{C}(\text{Protection}_s) = \hat{c}^* + \beta^C \text{Protection}_s, \tag{5}$$

where $\hat{c}^*$ re-centers the causal model to the outcome mean. Appendix

Table 1(A) presents estimated values $(\hat{\alpha}^P, \hat{\beta}^P)$ and $(\hat{c}^*, \hat{\beta}^C)$, which are reasonably similar; indeed, we cannot reject $\hat{\beta}^C = \hat{\beta}^P$ at the 95% level.

The best predictive model reduces MSE by around one-sixth; the best causal model by only a marginally smaller amount (Appendix Table 1(B)). In consequence the R^2 and CR^2 are very similar, at roughly 15%, whereas the experimental R^2 is about 5 percentage points smaller (Figure 2(A)).

Kremer et al. (2011) document substantial error in measuring *E. coli* levels. As is well-known, measurement error attenuates R^2, and, with repeated independent measures of Y, one can correct for classical measurement error by dividing the "raw R^2" by the test-retest reliability to compute a "signal R^2" (Spearman 1904). Online Appendix D shows similar logic applies to CR^2. We thus correct R^2 and CR^2 by dividing by the estimated test-retest reliability of 0.46 (Kremer et al. 2011). The third and fourth bars in Figure 2(A) display this signal CR^2, indicating that spring protection explains about one third of variation in water quality, both causally and predictively.

We emphasize the difference between our question and the original paper. Kremer et al. (2011) report the causal effect of spring protection on water quality, which speaks to the gains from hypothetically expanding protection. We instead ask what share of *observed* differences in water quality are explained by differences in spring protection, which speaks to its empirical importance as an explanation of differences in water quality. A natural question is why people have more access to clean water in some parts of the country than others. Our large estimated CR^2 indicates that spring protection is a main cause of these differences.

6.2 Test scores and class size

Our second application draws on Project STAR, which randomized over 10,000 elementary school students to classes of different size and has been widely used to estimate causal effects of class size (Krueger 1999; Chetty et al. 2011; Dynarski, Hyman, and Schanzenbach 2013). We ask what share of variation in test scores is causally explained by class size.

Our experimental sample consists of STAR data made publicly available by Achilles et al. (2008). Following prior literature, we assess the

causal effect of class size in a 2SLS regression:

$$\text{ClassSize}_i = \pi^C + \rho^C \text{Small}_i + \nu_i^C, \tag{6}$$

$$\text{Score}_{i,s} = \alpha_s^C + \beta_s^C \widehat{\text{ClassSize}}_i + \varepsilon_{i,s}^C, \tag{7}$$

where $\text{Score}_{i,s}$ is student i's mean score in subject s (reading or math) over grades 1–3, ClassSize_i is mean class size in grades 1–3, and Small_i is an indicator for experimental assignment to a small class. The second-stage coefficient $\hat{\beta}^C$ is the causal effect of class size, which pins down the set of causal models. As discussed in section 5.3, this causal effect is among *compliers*, and so could differ from the population average effect. In practice, compliance is very high (Krueger 1999), and compliers mirror the broader population on observables (Appendix Tables 2 and 3).

Achilles et al. (2008) also publish an observational sample of students in Tennessee schools matched to STAR schools that did not participate in the experiment. We assess the predictive relation between test scores and class size using the OLS regression

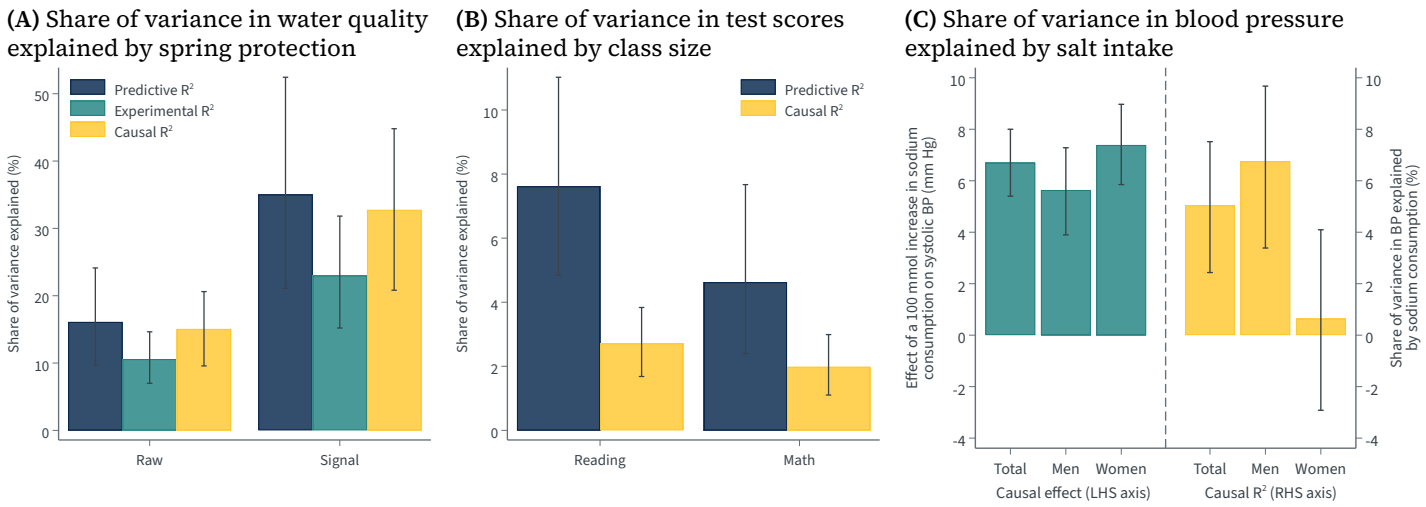
$$\text{Score}_{i,s} = \alpha_s^P + \beta_s^P \text{ClassSize}_i + \epsilon_{i,s}^P. \tag{8}$$

The resulting coefficients, $(\hat{\alpha}_s^P, \hat{\beta}_s^P)$, define our estimated best predictive model. The predictive slope is about five times the causal effect (Appendix Table 4), indicating substantial omitted variable bias in (8).

Finally, we assess the fit of the models (Appendix Table 5). Figure 2(B) shows that around 8% (5%) of variation in reading (math) scores is predicted by variation in class size, whereas only about 3% (2%) is causally explained. Thus, the predictive power of class size primarily reflects omitted variable bias: class size is not a major cause of test score variation, at least in this population.

Prior work has used STAR data to estimate the causal effect of class size, which answers the forward-looking policy question of how outcomes would change if we changed class sizes. The CR^2 instead measures how much class size can explain *existing* differences in outcomes. Researchers and policymakers have tried to understand the sources of differences in test scores among students (Dynarski and Michelmore 2017). The small CR^2 indicates that class size is not a major cause.

Figure 2. Causal R^2 in Applications



Notes: This figure shows the results in our applications. Panel (A) reports the share of variation in water quality in Kenyan springs causally explained by spring protection; Panel (B) test scores by class size; Panel (C) systolic blood pressure by salt intake, along with the causal effect of salt on blood pressure. Each panel shows 95% confidence intervals computed using 2,000 bootstrapped samples.

6.3 Blood pressure and sodium

Finally, we assess the share of variation in blood pressure causally explained by variation in sodium consumption, which is considered a leading cause of high blood pressure (Institute of Medicine 2010). Our experimental sample consists of participants in the DASH-Sodium experiment, an RCT evaluating the effects of sodium intake on blood pressure. Sacks et al. (2001) estimate the regression

$$\text{BP}_i = \alpha^C + \beta^C \text{Sodium}_i + \epsilon_i^C, \tag{9}$$

where BP_i is person i's systolic blood pressure. The estimate $\hat{\beta}^C$ yields our best causal model (after recentering). The authors also run separate regressions by gender, finding a larger effect among women.

To evaluate fit, we use observational data from the UK National Diet and Nutrition Survey, a long-running, nationally-representative survey which collected urinary sodium data (NDNS 2021). We follow a public health literature treating the U.S. and U.K. populations as transportable in this context (Scientific Advisory Committee on Nutrition 2003; Jones et al. 2020).

We find that sodium intake causally explains 5.1% of variation in

systolic blood pressure overall, and 6.8% in men, but only 0.8% among women ($p < 0.01$ for the difference) (Appendix Table 6). This is striking because the causal effect of sodium on blood pressure is, if anything, slightly larger for women. Sodium explains a much smaller share among women for three reasons. First, the variance of sodium is about 40% larger among men, so similar causal effects generate more variation for men. Second, the variance of blood pressure among men is about 20% smaller: there is less total variation to explain, so the share generated by sodium is larger. Third, among men, the experimental and observational regression coefficients are similar: there is little omitted variable bias. Among women, the causal effect is around 80% larger than the observational coefficient: sodium is *negatively* related to unobserved determinants of blood pressure. In consequence, sodium intake explains little among women, but a meaningful amount among men.

This discussion highlights the difference between the standard causal effects exercise—determining the effect of a change in sodium on blood pressure—and our goal: measuring the share of *observed* blood pressure variation causally explained by sodium intake. The former question is relevant to a doctor helping patients lower blood pressure; such a doctor should give reasonably similar advice to men and women, since the causal effects are similar. A doctor asking *why* a particular patient has high blood pressure, though should be less inclined to suspect sodium among women. Moreover, the low share explained, especially among women, suggests sodium is not a primary source of variation in blood pressure and that other determinants warrant further study.

Appendix

Proof of Lemma 1. The first sentence is direct from Definition 2. For the second sentence, since $\mathbb{E}\left[(\mathbb{E}[Y(X^O)] + c - Y)^2\right]$ is convex in c, the risk-minimizing constant is $c^* = \mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y(X^O)]]$. □

Proof of Proposition 1. For (i), the best causal model is $Y^C_{\mathcal{F},X^O}(x^O) = \mathbb{E}[Y]$, so $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 0$. For (ii), since $Y(x^O, x^U) = Y(x^O, x^{U'})$ for any $(x^O, x^U, x^{U'})$, write $Y(x^O, x^U) = Y(x^O)$. The model $Y^C_{X^O}(x^O) = Y(x^O)$ maximizes fit (its fit is 1), and is a causal model; since $\mathcal{F}$ is well-specified, it falls within $\mathcal{F}$. Hence, the best causal model achieves a fit of 1.

For (iii), the best causal model is $\mathbb{E}[Y]$, which achieves risk $\mathrm{Var}[Y]$, so $\mathrm{CR}^2_{\mathcal{F}}(X^O) = 0$. The argument for (ii) also shows (iv).

For (v), see the discussion of measurement error in Online Appendix D. For (vi), an example suffices: see Example A1. For the first sentence of (vii), since the transformation applied to Y is affine, write the transformed outcome as $y' = \alpha + \beta y$, for $\beta \neq 0$, and each transformed feature X_k as g_k, for $g_k(\cdot)$ strictly monotone. The best causal model for Y' is ${Y'}^C_{X^O}(x^{O'}) = \alpha + \beta Y^C_{X^O}(g^{-1}(x^{O'}))$, where g^{-1} is the element-wise inverse of $(g_k)_{k=1}^O$, and hence risk is β^2 times the risk of the best causal model for Y. Since variance is also inflated by β^2, the ratio of the risk to variance is unchanged, and hence CR^2 is unchanged. The second sentence follows similar steps, additionally using the fact g^{-1} is affine. □

Proof of Proposition 2. Part (i) follows directly from the fact that Definition 2 adds a constraint relative to Definition 1. For (ii), under independence, $\mathrm{ATE}(x_1^O \to x_2^O) = \mathbb{E}[Y \mid X^O = x_2^O] - \mathbb{E}[Y \mid X^O = x_1^O]$, so the best causal and predictive models coincide, and hence have the same fit. □

Proof of Proposition 3. The statements about R^2 are well-known. Since risk is non-negative, CR^2 is bounded above by 1. See Example A2 for an illustration of negativity and non-monotonicity. □

Proof of Proposition 4. Beginning with the second term on the RHS:

$$\begin{aligned}(\tilde{\beta}^{\mathrm{P}}_{XO} - \tilde{\beta}^{\mathrm{C}}_{XO})'\rho_{XO}(\tilde{\beta}^{\mathrm{P}}_{XO} - \tilde{\beta}^{\mathrm{C}}_{XO}) &= (\tilde{\beta}^{\mathrm{P}}_{XO})'\rho_{XO}\tilde{\beta}^{\mathrm{P}}_{XO} - 2(\tilde{\beta}^{\mathrm{P}}_{XO})'\rho_{XO}\tilde{\beta}^{\mathrm{C}}_{XO} + (\tilde{\beta}^{\mathrm{C}}_{XO})'\rho_{XO}\tilde{\beta}^{\mathrm{C}}_{XO} \\ &= \mathrm{R}^2_{\mathrm{lin}}(X^O) - \mathrm{CR}^2_{\mathrm{lin}}(X^O),\end{aligned}$$

where the first line follows from the fact ρ_{XO} is a correlation matrix (and so symmetric), and the second line applies both parts of Lemma A1. □

Proof of Proposition 5. For (i), $Y^C_{\mathcal{F},X^O}$ cannot be identified from the observational sample, nor $R(Y^C_{\mathcal{F},X^O})$ from the experimental sample. For (ii), for any M, $\mathcal{R}(M)$ is identified by the observational subsample. If the experiment is full-support, for each $x^O \in \mathcal{X}^O$, we can identify $Y^C_{X^O}(x^O)$ pointwise in the experimental subsample, and hence identify the function $Y^C_{X^O}$. Otherwise, we cannot identify $Y^C_{X^O}(x^O)$ for at least some values of x^O.

For (iii), as before, $\mathcal{R}(M)$ is identified by the observational subsample. Under our assumption that the linear model is well-specified, the best causal model is $Y^C_{X^O}(x^O) = \alpha + \sum_{k=1}^{O} x^O \beta^O$. If the experiment is full-rank, regressing Y on X^O in the experimental subsample recovers (α, β), and hence $Y^C_{X^O}$. If the experiment is not full-rank, it is easy to see that at least one β^k must not be identified.

Part (iv) follows immediately from noting that (i) $\hat{Y}^C_{\mathcal{F},X^O}$ converges to $Y^C_{\mathcal{F},X^O}$, (ii) for any M, $\hat{\mathcal{R}}(M)$ converges to $\mathcal{R}(M)$, (iii) $\widehat{\text{Var}}[Y]$ converges to $\text{Var}[Y]$, and then applying Slutsky's Theorem. □

References

Abadie, Alberto. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113, no. 2 (April): 231–263. ISSN: 0304-4076. https://doi.org/10.1016/S0304-4076(02)00201-4.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91, no. 5 (December): 1369–1401. ISSN: 0002-8282. https://doi.org/10.1257/aer.91.5.1369.

Achilles, C. M., Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. 2008. (Tennessee's Student Teacher Achievement Ratio (STAR) project. V. 1. Dataverse: Project STAR (https://dataverse.harvard.edu/dataverse/star). UNF:3:Ji2Q+9HCCZ-Abw3csOdMNdA==. License: CC0 1.0. Accessed January 26, 2026). https://doi.org/10.7910/DVN/SIWH9F.

Andrews, Isaiah, and Emily Oster. 2019. "A simple approximation for evaluating external validity bias." *Economics Letters* 178 (May): 58–62. ISSN: 0165-1765. https://doi.org/10.1016/j.econlet.2019.02.020.

Angrist, Joshua D. 2004. "Treatment effect heterogeneity in theory and practice." *The economic journal* 114 (494): C52–C83.

Angrist, Joshua D., and Iván Fernández-Val. 2013. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." In *Advances in Economics and Econometrics: Tenth World Congress,* edited by Daron Acemoglu, Manuel Arellano, and Eddie Dekel, 401–434. Econometric Society Monographs. Cambridge University Press.

Athey, Susan, and Guido W. Imbens. 2017. "The Econometrics of Randomized Experiments." In *Handbook of Field Experiments,* edited by Abhijit Vinayak Banerjee and Esther Duflo, 1:73–140. Handbook of Economic Field Experiments. Elsevier. https://doi.org/10.1016/bs.hefe.2016.10.003.

Blau, Francine D, and Lawrence M Kahn. 2017. "The gender wage gap: Extent, trends, and explanations." *Journal of Economic Literature* 55 (3): 789–865.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly Journal of Economics* 126 (4): 1593–1660.

Cinelli, Carlos, and Chad Hazlett. 2020. "Making sense of sensitivity: Extending omitted variable bias." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82 (1): 39–67.

———. 2025. "An omitted variable bias framework for sensitivity analysis of instrumental variables." *Biometrika* 112 (2): asaf004.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dawid, A. Philip, and Monica Musio. 2022. "Effects of Causes and Causes of Effects." First published as a Review in Advance on November 4, 2021, *Annual Review of Statistics and Its Application* 9:261–287. https://doi.org/10.1146/annurev-statistics-070121-061120.

Dynarski, Susan, Joshua Hyman, and Diane Whitmore Schanzenbach. 2013. "Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion." *Journal of policy Analysis and management* 32 (4): 692–717.

Dynarski, Susan M., and Katherine Michelmore. 2017. "The gap within the gap." Brookings Institution. Evidence Speaks, April 13, 2017. Accessed January 26, 2026. https://www.brookings.edu/articles/the-gap-within-the-gap/.

Epanomeritakis, Aristotelis, and Davide Viviano. 2025. *Choosing What to Learn: Experimental Design when Combining Experimental with Observational Evidence.* arXiv: 2510.23434 [econ.EM].

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the completeness of economic models." *Journal of Political Economy* 130 (4): 956–990.

Gelman, Andrew, and Guido Imbens. 2013. *Why ask why? Forward causal inference and reverse causal questions.* Technical report. National Bureau of Economic Research.

Glaeser, Edward L. 2011. *Triumph of the City: How Our Greatest Invention Makes Us Richer, Smarter, Greener, Healthier, and Happier.* 352. First edition. New York: Penguin Press, February. ISBN: 978-1-59420-277-3.

Goldberger, Arthur S. 1979. "Heritability." *Economica* 46 (184): 327–347.

Griliches, Zvi. 1974. "Errors in variables and other unobservables." *Econometrica* 42 (6): 971–998. https://doi.org/10.2307/1914213.

Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. 2015. "From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects." *Journal of the Royal Statistical Society Series A: Statistics in Society* 178 (3): 757–778.

Heckman, James, Justin L Tobias, and Edward Vytlacil. 2001. "Four parameters of interest in the evaluation of social programs." *Southern Economic Journal* 68 (2): 210–223.

———. 2003. "Simple estimators for treatment parameters in a latent-variable framework." *Review of Economics and Statistics* 85 (3): 748–755.

Heckman, James J. 2010. "Building bridges between structural and program evaluation approaches to evaluating policy." *Journal of Economic Literature* 48 (2): 356–398.

Hernán, Miguel A., and James M. Robins. 2025. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Huang, Melody Y. 2024. "Sensitivity Analysis for the Generalization of Experimental Results." *Journal of the Royal Statistical Society Series A: Statistics in Society* 187 (4): 900–918. https://doi.org/10.1093/jrsssa/qnae012.

Hull, Peter. 2025. *"One Weird Trick" to Characterize Effective Populations in Design-Based Specifications.* Technical report. Working Paper, September 3, 2025.

Institute of Medicine. 2010. *A Population-Based Policy and Systems Change Approach to Prevent and Control Hypertension.* Washington, DC: The National Academies Press. ISBN: 978-0-309-14809-2. https://doi.org/10.17226/12819.

Joint National Committee on Prevention Detection Evaluation and Treatment of High Blood Pressure. 1997. "The Sixth Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure." *Archives of Internal Medicine* 157, no. 21 (November): 2413–2446. ISSN: 0003-9926. https://doi.org/10.1001/archinte.1997.00440420033005.

Jones, Nicholas R, Terry McCormack, Margaret Constanti, and Richard J McManus. 2020. "Diagnosis and management of hypertension in adults: NICE guideline update 2019." *The British Journal of General Practice* 70 (691): 90–91.

Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Peterson Zwane. 2011. "Spring cleaning: Rural water impacts, valuation, and property rights institutions." *The Quarterly Journal of Economics* 126 (1): 145–205.

Krueger, Alan B. 1999. "Experimental estimates of education production functions." *The Quarterly Journal of Economics* 114 (2): 497–532.

Mincer, Jacob. 1958. "Investment in Human Capital and Personal Income Distribution." *Journal of Political Economy* 66, no. 4 (August): 281–302. ISSN: 0022-3808. https://doi.org/10.1086/258055.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625. ISSN: 09528385.

Oster, Emily. 2019. "Unobservable selection and coefficient stability: Theory and evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204.

Pearl, Judea. 1999. "Probabilities of Causation: Three Counterfactual Interpretations and Their Identification." *Synthese* 121 (1-2): 93–149. https://doi.org/10.1023/A:1005233831499.

Sacks, Frank M., Laura P. Svetkey, William M. Vollmer, Lawrence J. Appel, George A. Bray, David Harsha, Eva Obarzanek, et al. 2001. "Effects on Blood Pressure of Reduced Dietary Sodium and the Dietary Approaches to Stop Hypertension (DASH) Diet." *New England Journal of Medicine* 344, no. 1 (January): 3–10. ISSN: 0028-4793, 1533-4406. https://doi.org/10.1056/NEJM200101043440101.

Scientific Advisory Committee on Nutrition. 2003. *Salt and Health.* Scientific Advisory Committee on Nutrition.

Spearman, C. 1904. "The Proof and Measurement of Association between Two Things." *The American Journal of Psychology* 15 (1): 72–101. ISSN: 00029556.

University Of Cambridge, MRC Epidemiology Unit and NatCen Social Research. 2021. (National Diet and Nutrition Survey [SN 6533]. V. 19th Edition. [data collection]; accessed January 26, 2026). https://doi.org/10.5255/UKDA-SN-6533-19.

Vytlacil, Edward. 2002. "Independence, monotonicity, and latent index models: An equivalence result." *Econometrica* 70 (1): 331–341.

Whelton, Paul K., Robert M. Carey, Wilbert S. Aronow, Donald E. Casey, Karen J. Collins, Cheryl Dennison Himmelfarb, Sondra M. DePalma, et al. 2018. "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines." *Hypertension* 71, no. 6 (June): e13–e115. ISSN: 0194-911X, 1524-4563. https://doi.org/10.1161/HYP.0000000000000065.

Online Appendix

The Online Appendix has five sections. Online Appendix A contains intermediate steps and examples used in the proofs in the main text. Online Appendix B describes adapting the approach to include covariates. Online Appendix C discusses some alternative approaches to measure causally explained variation, and shows that non-monotonicity is an inherent feature of "reasonable" measures of causally explained variation. Online Appendix D presents additional material relating to estimation. Online Appendix E details applications.

Online Appendix A. Intermediate steps and examples for proofs of main results

Notation for population-level potential outcome function. In the main text, we defined the potential outcome function at the population-level, rather than a unit-specific potential outcome function. Similar notation appears in, *e.g.*, Vytlacil (2002) and Hernán and Robins (2025). To see that this is without loss, adopt Vytlacil's (2002) notation: let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space, and for each $x \in \mathcal{X}$, let $Y_x(\omega)$ be the random variable corresponding to the unit-specific potential outcome under x. Now redefine X to include any latent factors or indeed the unit index itself: $\tilde{X}(\omega) = (X(\omega), \omega)$. Then we can define the population-level $Y(\cdot)$ by $Y(\tilde{X}(\omega)) = Y_{X(\omega)}(\omega)$ for all ω.

Intermediate lemma. We now present an intermediate result for the main proofs.

Lemma A1. *Fix an outcome Y and a vector of observed features X^O, both with finite first and second moments and non-zero variances. Denote by $\tilde{Y}$ the standardized value of Y, by $\tilde{X}_k$ the standardized value of observed feature X_k, and by $\tilde{X}^O$ the corresponding vector of standardized observed features. Denote by ρ_{XO} the correlation matrix of observed features.*

(i) $\mathrm{R}^2_{\mathrm{lin}}(X^O) = (\tilde{\beta}^P)' \rho_{XO} \tilde{\beta}^P$*, where $\tilde{\beta}^P$ is the vector of OLS coefficients from a regression of $\tilde{Y}$ on $\tilde{X}^O$ in the population.*

(ii) If $\mathcal{F}_{\mathrm{lin}}$ is well-specified, $\mathrm{CR}^2_{\mathrm{lin}}(X^O) = 2(\tilde{\beta}^{\mathrm{P}}_{XO})' \rho_{XO} \tilde{\beta}^{\mathrm{C}}_{XO} - (\tilde{\beta}^{\mathrm{C}}_{XO})' \rho_{XO} \tilde{\beta}^{\mathrm{C}}_{XO}$*, where $\tilde{\beta}^C$ is the vector of OLS coefficients from a regression of $\tilde{Y}$ on $\tilde{X}^O$ in an experiment in which X^O is randomly assigned.*

Part (i) is well-known, but we include a proof for completeness.

Proof of Lemma A1. It is well-known that $\mathrm{R}^2_{\mathrm{lin}}$ is invariant to affine transformations. Proposition 1(vii) shows that this is also true of $\mathrm{CR}^2_{\mathrm{lin}}$. Then

we can demonstrate the lemma by standardizing all variables: for any linear model $\tilde{M}(\tilde{x}^O) = \tilde{\alpha} + \tilde{\beta}'\tilde{x}^O$ for the outcome $\tilde{Y}$, with the property that the model is well-calibrated (in the sense that its expectation is equal to the expected outcome):

$$\begin{aligned} G(M) &= 1 - \frac{\mathbb{E}\big[(\tilde{Y} - M(\tilde{X}^O))^2\big]}{\mathrm{Var}\big[\tilde{Y}\big]} \\ &= 1 - [1 - 2\mathrm{Cov}[\tilde{Y}, \tilde{M}(\tilde{X}^O)] + \mathrm{Var}\big[M(X^O)\big]] \\ &= 2\mathrm{Cov}[\tilde{Y}, \tilde{\beta}'\tilde{X}^O] - \tilde{\beta}'\rho_{XO}\tilde{\beta}. \end{aligned}$$

For (i), take the best linear predictor, $Y^P_{\mathrm{lin},X^O}(\tilde{x}^O) = \tilde{\alpha}^P + \tilde{\beta}^P\tilde{x}^O$. Then, $\mathrm{Cov}[\tilde{Y}, (\tilde{\beta}^P)'\tilde{X}^O] = (\tilde{\beta}^P)'\rho_{XO}\tilde{\beta}^P$, and hence $\mathrm{R}^2_{\mathrm{lin}}(X^O) = (\tilde{\beta}^P)'\rho_{XO}\tilde{\beta}^P$. For (ii), take the best linear causal model, $Y^C(X^O) = \tilde{\alpha}^C + \tilde{\beta}^C\tilde{X}^O$. If the model is well-specified, $\mathrm{CR}^2_{\mathrm{lin}}(X^O) = 2\mathrm{Cov}[\tilde{Y}, (\tilde{\beta}^C)'\tilde{X}^O] - (\tilde{\beta}^C)'\rho_{XO}\tilde{\beta}^C$. □

Examples used in proofs in main text. We now describe several examples used in the proofs in the main texts.

Example A1 (Example for proof of Proposition 1(vi))**.** Say $Y_i(X_i) = X_i$ and $X_i \sim \mathcal{N}(0,1)$. That is, X causally affects Y, but not *vice versa*. By Proposition 1(i), $\mathrm{CR}^2(Y \to X) = 0$. By Proposition 1(iii), $\mathrm{CR}^2(X \to Y) = 1$.

Example A2 (Example for proof of Proposition 3)**.** Consider the example discussed in Section 2, with $\alpha = 0$. The best causal model (unrestricted or linear) is $Y^C(c_i) = \beta c_i$, with goodness-of-fit

$$G(Y^C_{XO}) = 1 - \frac{\gamma^2}{\beta^2 + \gamma^2 + 2\beta\gamma\rho} = \frac{\beta^2 + 2\beta\gamma\rho}{\beta^2 + \gamma^2 + 2\beta\gamma\rho}.$$

Choosing $\rho = -1$, the expression can be rewritten $G(Y^C_{XO}) = \frac{\beta(\beta-2\gamma)}{(\beta-\gamma)^2}$, which can be made arbitrarily negative by taking $\beta = \gamma + \epsilon$ for ϵ small. This illustrates the statement about causal R^2 in part (i).

Part (ii) can again be illustrated by example: add a third variable Z to the example above with no causal effect. The causal R^2 of Z is zero; the joint causal R^2 of Z and C is negative.

Online Appendix B. Incorporating covariates

In the main text, we equated observable and manipulable features. That is, we assumed that all observed features could also be changed in an experiment. This distinction is partly conceptual: since it is difficult to identify the causal effect of these features, it is also difficult to identify

the share of variation they causally explain.

However, we may incorporate these covariates as follows. Say a non-manipulable variable M is observed and treatment is randomly assigned within M. Then, a natural approach is to compute the best causal model $Y^C_{X^O,m}$ for each value m of the non-manipulable feature M, and then compute $\{\text{CR}^2_m\}_{m\in\text{supp}(M)}$. This allows the analyst to report the share of variation causally explained across subgroups (*e.g.*, among men *vs.* women), as in our application to blood pressure and salt intake.

Another approach, which we do not pursue, is to allow the causal model in the general population to vary by the level of the covariate. Suppose the analyst aims to assess the share of variation explained in the population overall by a causal model which allows the effects of the feature of interest to differ between subgroups. It is tempting to define the "combined" causal model $\tilde{Y}^C_{X^O}(x^O) = \sum_{m'\in\text{supp}(M)} \mathbb{1}\{m = m'\}Y^C_{X^O,m'}(x^O)$, which assigns to each unit the value given by the causal model for that unit's subgroup. One might then assess the fit of this model by evaluating the MSE of $\tilde{Y}^C_{X^O}$. To see why this approach is unreasonable, suppose for a moment that there is no causal effect of X^O on the outcome, but the subgroup M is highly predictive of the outcome; then, $\tilde{Y}^C_{X^O}$ can easily achieve a very high fit, despite X^O explaining none of the variation in Y. The difficulty is that, when fitting separate causal models in each sub-group and then combining them, the combined model allows for different "intercepts" for each sub-group, and hence overstates the share of variance explained.

Online Appendix C. Alternative approaches to measure causally explained variation

We first describe some examples of alternative approaches. We then present an impossibility result, which we interpreet as showing that non-monotonicity is a constitutive feature of measures of goodness-of-fit for causal models.

C.1 Examples of alternative approaches

We discuss some other approaches to measure the share of variance in an outcome that is causally explained by a variable. We note how these approaches sometimes fall short.

Example A3 (Observational R^2). The observational R^2 violates properties (i) and (vi) of Proposition 1.

Example A4 (Experimental R^2). The experimental R^2 in general fails property (iii): even when X^O does not vary in the population, experimentally-induced variation can cause the experimental R^2 to be positive.

Example A5 (Relative variance of average potential outcome). An alternative measure is the ratio of the variance of the best causal model to the total variance of Y: $\text{Var}\left[Y^C_{\mathcal{F},X^O}(X^O)\right]/\text{Var}[Y]$. This measure would sometimes conclude that the observed features explain "more than 100%" of the variation in Y. To see this, consider

$$Y(X_{1,i}, X_{2,i}) = X_{1,i} - X_{2,i}, \tag{1}$$

$$\begin{pmatrix} X_{1,i} \\ X_{2,i} \end{pmatrix} \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{2}$$

for $\rho \in (-1, 1)$, where (1) is the potential outcome function, and (2) is the joint distribution of features. Suppose X_1 is observed, but X_2 is unobserved. The best causal model is then $Y^C(X_1) = X_1$, and hence $\text{Var}\left[Y^C(X_1)\right] = 1$. In contrast, $\text{Var}[Y] = 1 + 1 - 2\rho < 1$ for $\rho > 1/2$.

Example A6 (Standardized experimental regression coefficient). An alternative measure is the coefficient in a regression in experimental data of standardized Y on standardized X^O, where the standardization is with respect to population variances. Again, this measure would sometimes conclude that observed features explain more than 100% of variation. Consider the example from section 2. The standardized regression coefficient on C_i is $\beta/\sqrt{\beta^2 + \gamma^2 + 2\beta\gamma\rho}$. For $\beta = 1.5, \gamma = -1, \rho = 0.8$, the standardized coefficient is 1.63, so we would conclude that 163% of observed differences in test scores are explained by class size.

C.2 An impossibility result for monotonic measures of goodness-of-fit

We now discuss a more general result. The main text discussed a particular feature of the CR^2: the possibility that the share of variance causally explained falls as more features are observed. We claimed that this is an inherent attribute of measures of variation causally explained. We now formalize this claim through an axiomatic analysis.

Denote the set of features by $\hat{\mathcal{X}}$. Denote its power set by $2^{\hat{\mathcal{X}}}$, with typical element $\mathtt{X}$. Given a potential outcome function $Y(\cdot)$, a subset of features $\mathtt{X}$, and a distribution of features P_X, define a ***general measure of variation causally explained*** as a function $\rho_{P_X,Y} : 2^{\hat{\mathcal{X}}} \to \mathbb{R}$ such that, if

(P_X, Y) and $(\hat{P}_X, \hat{Y})$ induce the same joint distribution of Y and the features in X, and induce the same average potential outcome as a function of the features in X, then $\rho_{P_X,Y}(\mathsf{X}) = \rho_{\hat{P}_X,\hat{Y}}(\mathsf{X})$.[9]

We begin by describing three axioms.

Axiom 1 (Completeness). *A general measure of variation causally explained satisfies* ***completeness*** *if the measure is equal to one whenever the full set of features is observed:* $\rho_{P_X,Y}(\hat{\mathcal{X}}) = 1$ *for any* P_X *and any* $Y(\cdot)$.

Axiom 2 (Limited information). *A general measure of variation causally explained satisfies* ***limited information*** *if the measure is strictly less than one whenever the observed data* rule out *the possibility that the full set of features has been observed. Fix the set of observed features* X. *Denote the corresponding average potential outcome function by* Y_{X}. *Suppose that the distribution of observed features, and* Y_{X}, *could not alone generate the population joint distribution of the outcome and features. Then* $\rho_{P_X,Y}(\mathsf{X}) < 1$.

Axiom 3 (Monotonicity). *A general measure of variation causally explained satisfies* ***monotonicity*** *if observing additional features causes the share of variation explained to weakly increase. Formally, for any* $\mathsf{X} \subseteq \mathsf{X}'$, $\rho_{P_X,Y}(\mathsf{X}') \geq \rho_{P_X,Y}(\mathsf{X})$.

Motivation for axioms. When we say that no reasonable measure satisfies no monotonicity, we mean that no measure satisfies monotonicity once we restrict to measures that satisfy completeness and limited information. We argue that any reasonable measure should have these properties. Say that a measure did *not* satisfy completeness. Then, even observing *all* of the sources of variation in the outcome, our measure would still indicate that we cannot explain all the variation. This seems unreasonable. Essentially the converse intuition holds for limited information: if a measure does not satisfy limited information, then it sometimes indicates that the observed features fully explain variation in the outcome, even though there is no potential outcomes function that is consistent with the view that there are no other determinants of the outcome. We consider both of these axioms essential for a reasonable measure of variation causally explained.

Logical independence. These axioms are logically independent: no pair of axioms implies the third axiom. To see this, note that the stan-

[9]This restriction is motivated by the fact that, even with observational data on Y and the variables in $\mathcal{X}$, and experimental data in which $\mathcal{X}$ is randomly assigned and the resulting values of Y are recorded, the analyst cannot identify anything that is not a function of the joint distribution or the average potential outcome.

dard measure of predictive R^2 satisfies completeness and monotonicity, but not limited information. The CR^2 satisfies completeness and limited information, but not monotonicity. On the other hand, limited information and monotonicity, but not completeness, are satisfied by a measure which trivially defines any set of features to explain none of the variation in the outcome. This shows that this description of axioms does not involve any redundancy.

Impossibility. We now state and prove an impossibility result.

Proposition A1. *No general measure of the share of variation causally explained satisfies completeness, limited information, and monotonicity.*

Proof of Proposition A1. We prove the result by way of a simple example. The potential outcome function is $Y(X_1, X_2, X_3) = \gamma X_1 + \beta X_2 + \beta X_3$, for $\gamma \neq 0, \beta \neq 0$, where X_2 and X_3 are perfectly negatively correlated, and X_1 is independent of X_2 (and hence X_3). Denote this population feature distribution by P_X, and its marginals by P_X^1, P_X^2, and P_X^3, respectively. Denote by $P_X^{1,2}$ the joint distribution of the first and second features.

Suppose by way of contradiction that there is such a measure, ρ. Completeness requires $\rho_{P_X,Y}(\{X_1\}) = 1$. To see this, say there is only one feature, X_1, with distribution P_X^1, and the potential outcome function is $\hat{Y}(X_1) = \gamma X_1$. Completeness requires $\rho_{P_X^1,Y}(\{X_1\}) = 1$. Since these two cases induce the same distribution of the outcome and observables, and have the same average potential outcome, we must also have $\rho_{P_X,Y}(\{X_1\}) = 1$.

The second step is to argue limited information requires $\rho_{P_X,Y}(\{X_1, X_2\}) < 1$. This is because, if (X_1, X_2) are observed, the average potential outcome function is $\gamma x_1 + \beta x_2 + \beta \mathbb{E}[X_3]$. This average potential outcome, and the population distribution of observed features, could not generate the population joint distribution of the outcome and features, since, in the population, Y and X_2 are independent. By consequence, limited information requires $\rho_{P_X,Y}(\{X_1, X_2\}) < 1$.

The third step is to note that applying monotonicity to $\rho_{P_X,Y}(\{X_1\}) = 1$ gives $\rho_{P_X,Y}(\{X_1, X_2\}) \geq 1$. Since the second and third steps contradict one another, no measure can satisfy all three axioms. □

Thus, among "reasonable" measures (in the sense of completeness and limited information), the possibility of non-monotonicity is inevitable. Intuitively, when we predict an outcome, knowing more features can only be helpful. When causally explaining an outcome, observing an additional feature can "set us back", indicating that the feature

suppresses rather than generates variance in the outcome.

Online Appendix D. Additional material on estimation and inference

This appendix presents several pieces of additional material on estimation and inference.

D.1 Simulations

Simulation 1: independent features in a well-specified linear model. We begin with the simplest non-trivial setting, in which the potential outcome function is linear, there are two features (one of which is observed), and the features variables are independent of one another:

$$Y(X_1, X_2) = \beta_1 X_1 + X_2 \tag{3}$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{4}$$

where (3) is the (linear) potential outcome function, and (4) is the joint distribution of (X_1, X_2). We vary β_1 as part of the simulation.

Appendix Figure 1(A) summarizes our results. We begin with the true $\text{CR}^2(X_1)$. Since (3) is linear, $\text{CR}^2(X_1) = \text{CR}^2_{\text{lin}}(X_1)$; since observed and unobserved features are independent, $\text{CR}^2(X_1) = \text{R}^2(X_1)$. When X_1 has no effect on the outcome ($\beta_1 = 0$), $\text{CR}^2(X_1) = 0$; as β_1 increases (in magnitude), $\text{CR}^2(X_1)$ rises, eventually approaching 1.

To examine the plug-in estimator's performance, we specify several additional parameters. The analyst collects a sample of size N, of which three-fifths of units are in the observational sample ($p = 0.6$). The experiment consists of two equally-probable treatment arms, assigning units to $X_1 \in \{0, 1\}$. The navy dots show the performance of the estimator when the full sample size is $N = 200$; we perform $1,000$ simulations and present the mean estimate. As expected, the estimator displays a downward bias which falls fairly quickly as the number of units increases, as the light blue ($N = 500$) and orange ($N = 2,000$) series show.

Simulation 2: correlated features in a well-specified linear model. As before, there are two features, one of which is observed, but we allow for correlation between the features:

$$Y(X_1, X_2) = X_2 \tag{5}$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{6}$$

Relative to the first simulation, we fix the first feature to have no causal effect ($\beta_1 = 0$), but vary the correlation between features (ρ).

Appendix Figure 1(B) summarizes our results. For any ρ, $\text{CR}^2(X_1) = 0$, since X_1 does not have any effect on Y. In contrast, the predictive R^2 is strictly positive when $\rho \neq 0$: X_1 does have some predictive power due to its correlation with X_2. As before, the plug-in estimator (using the parameters as for the first simulation) shows a finite-sample downward bias that vanishes reasonably quickly in the number of observations.

Simulation 3: correlated features in a misspecified model. Finally, we introduce misspecification through adding a non-linearity into the true potential outcome function:

$$Y(X_1, X_2) = 0.2X_1 + \gamma X_1^2 + X_2 \tag{7}$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \quad \mu = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{8}$$

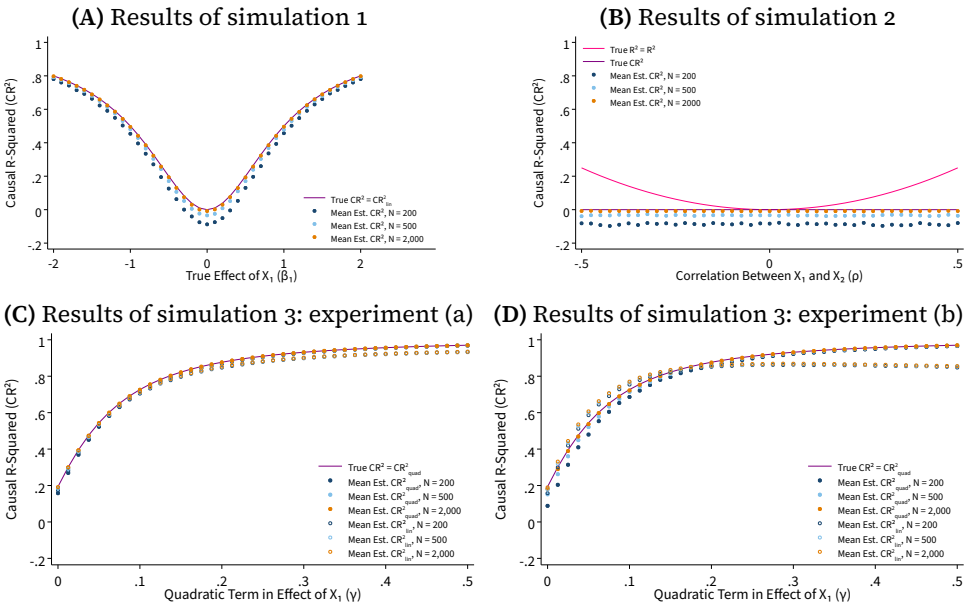
We vary γ as part of the simulation. Since the true potential outcome function is quadratic, the non-parametric $\text{CR}^2(X_1)$ coincides with $\text{CR}^2_{\text{quad}}(X_1)$, where quad denotes the class of quadratic models. As γ increases, $\text{CR}^2(X_1)$ increases, approaching 1 for γ large.

We now turn to the plug-in estimator. We consider two experiments, each with three treatment arms to allow for estimating a quadratic model. In experiment (a) (Panel (C)), the experimental sample is spit evenly between being assigned the mean of X_1, or one standard deviation above or below. Solid circles show mean estimates under a correctly-specified quadratic model; they exhibit a small, finite-sample downward bias that vanishes reasonably quickly as N grows. Hollow circles show mean estimates from a misspecified linear model. Here, the plug-in estimator need not converge to the true CR^2. In practice, the degree of divergence is modest.

In experiment (b) (Panel (D)), the experiment instead evenly divides units between being two, three, or four standard deviations above the mean of X_1. As before, the well-specified quadratic estimates (solid circles) converge to the true CR^2. Now, however, the misspecified linear estimates (in hollow circles) differ substantially from the true CR^2. Intuitively, the linear model recovers an experiment-weighted linear projection of a non-linear dose-response curve; as a result, the fit of

the causal model estimated from the experiment is worse when the treatment arms are further away from the mass of the feature in the population.

Appendix Figure 1. Results of simulations



Notes: This figure presents results from our simulations. Panel (A) displays results from simulation 1: the purple line displays the true CR^2, and dots show the mean estimated CR^2 in simulations of different sizes. Panel (B) replicates Panel (A) for Simulation 2. Panel (C) replicates Panel (A) for the first experiment in Simulation 3, and Panel (D) for the second experiment.

D.2 Details on the Delta method and bootstrapping

We first sketch how the Delta method can be used to compute standard errors in the case of the linear CR^2. Denote the estimated coefficients of the linear causal model by $\hat{\theta}$. Note that we can write the plug-in estimator for the causal R^2 as a simple function of other statistics:

$$\widehat{\mathrm{CR}}^2_{\mathrm{lin}}(\hat{\theta}) = 1 - \frac{\hat{\theta}_1}{\hat{\theta}_2}, \text{ for } \hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} := \begin{pmatrix} \widehat{\mathcal{R}}_{\mathrm{lin}} \\ \widehat{\mathrm{Var}}[Y]. \end{pmatrix}$$

This suggests a Delta method approach to inference. The main intermediate step is to show that $\hat{\theta}$ is itself asymptotically normal, and to derive its asymptotic covariance. This requires some care, since $\hat{\theta}_2$ depends only on the observational sample, but $\hat{\theta}_1$ depends on both the observational sample and the experimental sample through the coefficients of the linear causal model.

To do so, one can again apply the Delta method to find the asymptotic distribution of $\hat{\theta}$ by writing $\hat{\theta}$ as a function of a random vector of experimental and observational sample moments. By the Lindeberg-Lévy Central Limit Theorem, the distribution of the observational sample moments is asymptotically normal with mean zero and Σ_O the covariance matrix of the observational sample moments.

Similarly, since the experimental sample moments are simply the vector of coefficients from an OLS regression in the experimental sample, their asymptotic distribution is normal with mean zero and variance Σ_E, the standard asymptotic variance of the OLS estimator. Moreover, the observational and experimental sample moments are statistically independent since they are estimated in separate samples. Since both are asymptotically normal and independent, we can conclude that they are jointly asymptotically normal. In particular, the stacked vector of observational and experimental moments has asymptotic block-diagonal covariance matrix

$$\Sigma = \begin{pmatrix} \frac{1}{p}\Sigma_O & \mathbf{0} \\ \mathbf{0} & \frac{1}{1-p}\Sigma_E \end{pmatrix},$$

where the $\frac{1}{p}$ and $\frac{1}{1-p}$ terms account for the fact that the overall sample (of size N) is split between an observational dataset (with share p) and an experimental dataset (with share $1-p$).

We can then apply the Delta method to describe the asymptotic distribution of $\hat{\theta}$, and again apply the Delta method to find the asymptotic distribution of the linear causal R^2.

The Delta method is useful in the linear case, and especially when there is only one observed feature. A second, more general approach is to compute standard errors through bootstrapping.

Algorithm 1: Bootstrap Variance Estimation

Data: Y, X^O, S

Result: Bootstrapped Variance Estimator $\hat{V}$

for $i \leftarrow 1$ **to** B **do**

  Construct a bootstrap dataset $(Y^{(b)}, X^{O,(b)}, S^{(b)})$ by sampling N_O rows of (Y, X^O, S) with replacement from the observational sample, and N_E rows of (Y, X^O, S) with replacement from the experimental sample;

  Append these datasets together to create a dataset of size $N = N_O + N_E$;

  Compute the CR^2 estimator $\widehat{\mathrm{CR}}^{2,(b)}$ based on $(Y^{(b)}, X^{O,(b)}, S^{(b)})$ in this appended dataset;

Define

$$\hat{V} = \frac{1}{B}\sum_{b=1}^{B}\left[\left(\widehat{\mathrm{CR}}^{2,(b)} - \frac{1}{B}\sum_{b=1}^{B}\widehat{\mathrm{CR}}^{2,(b)}\right)^2\right];$$

This bootstrap is reasonably simple to implement, and performs well in our simulations, but has the disadvantage that is less transparent than the Delta method, and more computationally costly.

D.3 Measurement error

In practice, the analyst may measure the features or outcome with error. This appendix section discusses the role of measurement error. We begin with a proof of Proposition 1(v).

Proof of Proposition 1(v). Since ε is causally unaffected by X^O, and is mean-zero, its addition does not affect the best causal model. Denote this best causal model by M^C. Then:

$$\begin{aligned}
\left|CR^2_{\mathcal{F}}(X^O \to Y')\right| &= \left|\frac{\mathrm{Var}[Y'] - \mathbb{E}\left[(Y' - M^C(X^O))^2\right]}{\mathrm{Var}[Y']}\right| \\
&= \left|1 - \frac{\mathbb{E}\left[(Y - M^C(X^O))^2\right] + \mathrm{Var}[\varepsilon]}{\mathrm{Var}[Y] + \mathrm{Var}[\varepsilon]}\right| \\
&\leq \left|1 - \frac{\mathbb{E}\left[(Y - M^C(X^O))^2\right]}{\mathrm{Var}[Y]}\right| = \left|CR^2_{\mathcal{F}}(X^O \to Y)\right|.
\end{aligned}$$

□

We now go on to describe the role of measurement error more precisely in the linear case. Consider ***classical measurement error***: for $Z \in$

$\{X^O, Y\}$, the observed value is $Z' = Z + \varepsilon$ for ε mean zero and uncorrelated with (X^O, Y). There are four cases to consider, depending on whether noise appears in the outcome or feature, and whether it arises in experimental or observational data.[10]

Proposition A2 (measurement error). *Say there is a single observable feature. Denote by* $\mathrm{CR}^2_{\mathrm{CME}}$ *the linear* CR^2 *when there is classical measurement error (CME), and by* CR^2 *the true value.*

(i) CME in the outcome in the observational data attenuates CR^2*:* $\mathrm{CR}^2_{\mathrm{CME}} = \frac{\mathrm{Var}[Y]}{\mathrm{Var}[Y]+\mathrm{Var}[\varepsilon]}\mathrm{CR}^2$.

(ii) CME in the feature in the observational data reduces CR^2*:* $\mathrm{CR}^2_{\mathrm{CME}} = \mathrm{CR}^2 - \beta^2_{\mathrm{EXP}}\frac{\mathrm{Var}[\varepsilon]}{\mathrm{Var}[Y]}$*, for* β_{EXP} *the experimental regression coefficient on* X^O*.*

(iii) CME in the outcome in the experiment does not affect CR^2*.*

(iv) CME in the feature in the experiment can increase or reduce the causal R^2*, but* CR^2 *remains bounded above by* R^2*.*

Proof of Proposition A2. For simplicity, denote $X \coloneqq X^O$. Denote by Y^C_{lin} the best linear causal model for the relationship between X and Y. For part (i), noting that CME in the observational data does not affect the best causal model:

$$\begin{aligned}\mathrm{CR}^2_{\mathrm{CME}} &= 1 - \frac{\mathbb{E}\left[(Y' - Y^C_{\mathrm{lin}}(X))^2\right]}{\mathrm{Var}[Y']} \\ &= 1 - \frac{\mathbb{E}\left[(Y - Y^C_{\mathrm{lin}}(X))^2\right] + \mathrm{Var}[\epsilon]}{\mathrm{Var}[Y] + \mathrm{Var}[\epsilon]} = \mathrm{CR}^2\frac{\mathrm{Var}[Y]}{\mathrm{Var}[Y] + \mathrm{Var}[\epsilon]},\end{aligned}$$

where the second line makes use of the independence of ϵ. For part (ii), as before, CME in the observational data does not affect the best causal model. Denoting by β the slope coefficient from a regression of Y on X in the experimental data:

$$\begin{aligned}\mathrm{CR}^2_{\mathrm{CME}} &= 1 - \frac{\mathbb{E}\left[(\alpha + \beta(X + \epsilon) - Y)^2\right]}{\mathrm{Var}[Y]} \\ &= \frac{\mathrm{Var}[Y] - \mathbb{E}\left[(\alpha + \beta X - Y)^2\right] - \beta^2\,\mathrm{Var}[\epsilon]}{\mathrm{Var}[Y]} = \mathrm{CR}^2 - \beta^2\frac{\mathrm{Var}[\epsilon]}{\mathrm{Var}[Y]},\end{aligned}$$

where the first line defines (α, β) as the intercept and slope respectively in the best causal model, and the second line again uses the independence of ϵ.

Part (iii) follows from the well-known result that classical measure-

[10] Measurement error in the experimentally-assigned feature would occur, *e.g.*, if treatment status is sometimes recorded incorrectly.

ment error in the dependent variable does not bias OLS coefficients. For part (iv), denote by $\text{Var}_{EXP}[\cdot]$ the variance in the experiment. It is well-known that, under CME in the independent variable, the OLS slope coefficient will be attenuated as follows: $\beta_{\text{CME}} = \beta \frac{\text{Var}_{EXP}[X]}{\text{Var}_{EXP}[X]+\text{Var}_{EXP}[\varepsilon]}$. The result follows from subtracting CR^2 from CR^2_{CME} and rearranging. □

Proposition A2 can be used to sign the bias from measurement error. It also motivates correcting CR^2 for measurement error, under further assumptions. For instance, consider case (i), and suppose the analyst observes two equally noisy independent measurements of each unit's outcome, Y'_1 and Y'_2. It is well-known (Spearman 1904; Griliches 1974) that the "raw predictive R^2" (R^2_{raw}) between Y and X^O is attenuated relative to the "signal predictive R^2" ($\text{R}^2_{\text{signal}}$), and that the the analyst can recover $\text{R}^2_{\text{signal}}$ by dividing R^2_{raw} by the reliability $r \coloneqq \text{Corr}[Y'_{i,1}, Y'_{i,2}]$.[11] By similar logic, in the well-specified case, $\text{CR}^2_{\text{signal}} = \text{CR}^2_{\text{raw}}/r$; hence, an analyst with multiple outcome measures may correct for measurement error. We do this in our application to Kremer et al. (2011) in section 6.

Online Appendix E. Details of applications

E.1 Details of application 1

Data. Kremer et al. (2011) publish the data via Harvard Dataverse.

Processing and cleaning. We restrict both the experimental and observational samples to springs. With this restriction, there are 274 units in the observational sample, and 726 units in the experimental sample.

Best predictive and causal models. Panel (A) of Table 1 presents the estimated values $(\hat{\alpha}^P, \hat{\beta}^P)$ and $(\hat{c}^*, \hat{\beta}^C)$ in equations (3) and (4). Panels (B) and (C) assess the fit of these models.

We then assess the fit of these models (Panels B–C).

[11] This is because $r = \text{Var}[Y]/\text{Var}[Y']$, so $\text{R}^2_{\text{raw}} = \text{Var}[Y'] - \mathbb{E}[(Y' - Y^P(X^O))^2]/\text{Var}[Y'] = \text{R}^2_{\text{signal}} \text{Var}[Y]/\text{Var}[Y']$.

Appendix Table 1. Share of variance in E Coli causally explained by variance in spring protection

	Obs. data (1)	Exp. data (2)
A. Best predictive and causal models		
Constant	4.82	3.64
	(0.144)	(0.089)
Spring protection	-1.98	-1.47
	(0.273)	(0.158)
B. Outcome variance and mean squared error		
Var[ln E Coli]	4.88	4.46
MSE[Best predictive model for ln E Coli]	4.08	–
MSE[Best causal model for ln E Coli]	4.13	3.98
C. Share of variance explained		
% of var. predictively explained in pop. (obs. R^2)	16.42%	–
% of var. causally explained in exp. (exp. R^2)	–	10.71%
% of var. causally explained in pop. (CR^2)	15.38%	–

Notes: This table summarizes our analysis of variation in water quality (*E. coli* level) causally explained by spring protection. Panel (A) reports estimates of equations (3) and (4), with corresponding standard errors. Panel (B) reports the MSE of these models. Panel (C) reports the implied share of variance explained.

E.2 Details of application 2

Data. The data are made available in Achilles et al. (Tennessee's Student Teacher Achievement Ratio (STAR) project).

Processing and cleaning. For the experimental dataset, we restrict attention to students with non-missing schools, class sizes, and reading and math scores in each of grades K–3. For the observational dataset, we have access to students' information only in grades 1–3; we restrict attention to students with non-missing data in those years. For each student in the experimental (observational) data, we express reading and math scores in grades K-3 (1-3) in percentage points, and then take the unweighted mean over the grades for each of reading and math scores. We use this mean as our outcome variable. For each student in the experimental (observational) data, we compute the mean class size in grades K-3 (1-3), and use this mean as our feature variable of interest.

Complier characteristics. Following the approach described in Abadie (2003) and Hull (2025), we assess the characteristics of the complier group through a 2SLS regression of the form

$$\text{Class size}_i \times c_i = \gamma + \delta \text{Class size}_i + \epsilon_i \tag{9}$$

$$\text{Class size}_i = \zeta + \theta \text{Assigned small}_i + \eta_i, \tag{10}$$

where c_i is the characteristic of interest. In each case, we first demean the characteristic of interest so a test for the null hypothesis that $\delta = 0$ can be interpreted as a test for the null hypothesis that the mean value of the characteristic among compliers is the same as the mean value of the characteristic among non-compliers.

We run these regressions in the experimental sample. Appendix Table 2 reports 2SLS coefficients. Compliers are similar to non-compliers in terms of gender, race, and socioeconomic status, as proxied by free and reduced-price lunch receipt. This is consistent with the observation in Krueger (1999) that the rate of compliance was high.

Appendix Table 2. Tennessee STAR Complier Characteristics

	Male (1)	Minority Race (2)	FRL Recipient (3)
Constant	0.37	1.10	-1.94
	(1.369)	(1.233)	(1.201)
Coefficient	-0.02	-0.05	0.10
	(0.067)	(0.061)	(0.059)
Observations	2529	2529	2529

Notes: This table presents estimates of (γ, δ) from 2SLS regressions (9) and (10) for three student characteristics c_i. The first row shows the estimated value of γ, and the third row shows the estimated value of δ (with corresponding standard errors in the second and fourth rows, respectively). The fifth row shows the number of observations. In column (1), c_i is an indicator variable for the student being male; in column (2), an indicator for being non-white; in column (3), the share of years in grades K–3 that the student received free or reduced-price lunch.

First stage. Next, we consider the first-stage relation between treatment assignment and class size. Appendix Table 3 presents estimates of equation (6). We present estimated first stages for grades K–3, as well as for the average class size over these grades, which we use as our ultimate variable of interest. The effects of treatment assignment are large and highly significant, with F-statistics well above 1,000.

Appendix Table 3. First stage in Tennessee STAR

	Class size grade K (1)	Class size grade 1 (2)	Class size grade 2 (3)	Class size grade 3 (4)	Class size avg. K-3 (5)
Constant	22.24	22.40	22.19	22.25	22.27
	(0.046)	(0.062)	(0.068)	(0.078)	(0.045)
Assigned small	-7.28	-6.43	-6.45	-6.05	-6.55
	(0.082)	(0.111)	(0.121)	(0.140)	(0.081)
Observations	2771	2771	2771	2771	2771
F-statistic	7979.672	3336.140	2844.131	1860.067	6518.657

Notes: This table presents estimates of (π^C, ρ^C) from an OLS regression corresponding to equation (6) in the experimental data. The first row shows the estimated value of π^C, and the third row shows the estimated value of ρ^C (with corresponding standard errors in the second and fourth rows, respectively). The fifth row shows the number of observations. The sixth row shows the F-statistic. In each column, the outcome is the number of students in the child's class in the grade in the column title.

Best predictive and causal models. Next, we estimate the best predictive and causal models. Appendix Table 4 presents estimates of equations (8) (Panel A) and (7) (Panel B). In both the best predictive and best causal models, class size has a reasonably large and statistically significant effect on test scores, but the effect is substantially smaller in the best causal model, consistent with omitted variables biasing the observational relation downward relative to the causal effect.

Appendix Table 4. Best predictive and causal models in Tennessee STAR

	Reading (1)	Math (2)
A. Best predictive model (observational data)		
Constant	121.36	111.16
	(5.643)	(4.309)
Class size	-1.56	-0.91
	(0.243)	(0.186)
Observations	501	501
B. Best causal model (experimental data)		
Constant	93.10	94.86
	(1.411)	(1.137)
Class size	-0.32	-0.23
	(0.069)	(0.055)
Observations	2771	2771

Notes: This table presents the best predictive and causal models in the STAR setting. Column (1) presents results for reading scores, and column (2) for math scores. Panel A shows estimates of (α_s^P, β_s^P) from an OLS regression corresponding to equation (8) in the observational data, for subjects s corresponding to reading and math scores. The first row shows the estimated value of α_s^P, and the third row shows the estimated value of β_s^P (with corresponding standard errors in the second and fourth rows, respectively). The fifth row shows the number of observations. Both class size and test scores are averaged over grades. Panel B replicates Panel A, instead presenting estimates of (α_s^C, β_s^C) from a regression corresponding to equation (7) in the experimental data.

Fit of the best predictive and causal models. Finally, we assess the fit of the best predictive and causal models. Appendix Table 5 summarizes our results. Panel A shows the variance of test scores and the MSE of the best predictive and causal model, *i.e.* the variance of the residuals.

Panel B shows the corresponding shares of variance causally explained, that is, $1 - \text{MSE}/\text{Var[Score]}$. These values are computed directly from Panel A.

Appendix Table 5. Share of variance in test scores causally explained by class size

	Reading obs. data (1)	Reading exp. data (2)	Math obs. data (3)	Math exp. data (4)
A. Outcome variance and MSE				
Var[Score]	146.15	112.89	82.52	73.29
MSE[Best predictive model]	135.02	–	78.71	–
MSE[Best causal model]	142.18	111.74	80.88	72.56
B. Share of variance explained				
% of var. predictively explained in pop. (obs. R^2)	7.62	–	4.62	–
% of var. causally explained in exp. (exp. R^2)	–	1.02	–	1.00
% of var. causally explained in pop. (CR^2)	2.72	–	1.99	–
C. Hypothesis tests				
CR^2 for read = 0	0.000			
CR^2 for math = 0	0.000			
CR^2 for read = CR^2 for math	0.0805			

Notes: This table presents the share of variance in test scores causally explained by variance in class size. The first two columns present results for reading scores. Column (1) presents results in the observational data, and column (2) in the experimental data. Columns (3)-(4) replicate columns (1)-(2) for math scores. Panel A shows the outcome variance and MSE of the models. Panel B shows the share of variance explained. Panel C shows hypothesis tests involving the CR^2, computed using the bootstrapping described in Online Appendix D.

Panel C presents various hypothesis tests, computed by bootstrapping. We reject that the causal R^2 for reading or math is zero at the 1% level: that is, we reject the null hypothesis that class size causally

explains no variation in reading scores (and likewise for math scores). Although the point estimate for the CR^2 for math scores is somewhat lower than the estimate for reading scores, the difference is insignificant at 5% level.

E.3 Details of application 3

Data. The observational data are available from the UK Data Service under study number 6533.

For our experimental data, we use the DASH-Sodium experiment. We construct the estimated pooled causal effect of sodium using Figure 1A of Sacks et al. (2001), and the estimated causal effects by gender using Figure 2A of Sacks et al. (2001). The study recruited 412 U.S. participants with normal, high-normal, or high blood pressure.[12] These people were randomized into a control group and six treatment groups. Each treatment was a combination of 1) a typical US diet *vs.* a healthy diet and 2) low, intermediate, or high salt levels (50, 100, and 150 mmol sodium per day respectively). Study staff prepared the food, and participants got all their meals and snacks at an outpatient clinic. After a two-week run-in period during which everyone ate a high-sodium control diet, participants followed their assigned treatment diet for 30 days. At the end of the month, researchers measured participants' blood pressure, which is the main outcome of the study.

Best predictive and causal models. We estimate the best causal model in equation (9) using Figure 1A of Sacks et al. (2001). We estimate the best predictive model in the observational data.

Fit of the best predictive and causal models. Appendix Table 6 summarizes our results. Panel A shows that, both pooling genders and for men and women separately, the best predictive and causal models reduce MSE relative to the baseline variance of blood pressure. This reduction is much smaller for women. In consequence, the share of variance explained for women is substantially lower than for men.

[12] US guidelines on what blood pressure is normal or high has changed over time. At the time of the study, a blood pressure of 125 mm Hg systolic and 85 mm Hg diastolic was considered normal (Joint National Committee on Prevention Detection Evaluation and Treatment of High Blood Pressure 1997). Some current guidelines (e.g., Whelton et al. (2018)) would consider it high. The study required participants to have a diastolic blood pressure of 80–95 mm Hg and a systolic blood pressure of 120–160 mm Hg.

Appendix Table 6. Share of variance in blood pressure causally explained by salt intake

	Pooled (1)	Men (2)	Women (3)
A. Outcome variance and mean squared error			
Var[BP]	286.38	255.74	305.06
MSE[Best predictive model]	271.61	237.86	299.63
MSE[Best causal model]	271.75	238.13	302.73
B. Share of variance explained			
% of var. predictively exp. (obs. R^2)	5.16%	6.99%	1.78%
% of var. causally exp. (CR^2)	5.11%	6.89%	0.76%

Notes: This table replicates Appendix Table 5 for our application to the share of variation in blood pressure explained by salt intake.