Online Appendix

The Online Appendix has six sections, each self-contained. Online Appendix A discusses some alternative approaches to measure causally explained variation. Online Appendix B discusses bootstrapping. Online Appendix C presents simulations. Online Appendix D describes adapting the approach to include covariates. Online Appendix E shows that non-monotonicity is an inherent feature of "reasonable" measures of causally explained variation. Online Appendix F details the applications.

Online Appendix A. Alternative approaches to measure causally explained variation

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 A1 (Observational R^2). The observational R^2 violates properties (i) and (vi) of Proposition 1.

Example A2 (Experimental R^2). The experimental R^2 in general fails property (iii). To see this, note that even when X^O does not vary in the population, experimentally-induced variation can cause the experimental R^2 to be strictly positive.

Example A3 (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: $\frac{\text{Var}\left[Y^C_{\mathcal{F},X^O}(X^O)\right]}{\text{Var}[Y]}$. This measure would sometimes conclude that the observed features explain "more than 100%" of the variation in Y. To see this, consider

$$Y(X_{1,i}, X_{2,i}) = X_{1,i} - X_{2,i}, \tag{C1}$$

$$\begin{pmatrix} X_{1,i} \\ X_{2,i} \end{pmatrix} \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{C2}$$

for $\rho \in (-1, 1]$, where (C1) is the potential outcome function, and (C2) 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 $\text{Var}[Y^C(X_1)] = 1$. In contrast, $\text{Var}[Y] = 1 + 1 - 2\rho < 1$ for $\rho > 1/2$.

Example A4 (Coefficient in a regression of standardized Y on standardized

X^O.). 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 the population variances. Unfortunately, this coefficient cannot generally be interpreted as the share of variation causally explained by X^O, even in the special case in which there is a single unobserved feature X^O, and the true potential outcome function is linear. To see this, consider again the example from section 2. The standardized regression coefficient on C_i is $\frac{\beta}{\sqrt{\beta^2+\gamma^2+2\beta\gamma\rho}}$. This quantity may exceed one. For example, let $\beta = 1.5, \gamma = -1, \rho = 0.8$. Then, the standardized coefficient equals 1.63. Using this measure, we would conclude that 163% of the observed differences in test scores are explained by class size.

Online Appendix B. Details of bootstrapping

We 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 $\widehat{\text{CR}}^{2,(b)}$ based on $(Y^{(b)}, X^{O,(b)}, S^{(b)})$ in this appended dataset;

Define

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

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.

Online Appendix C. 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 and the feature variables are independent of one another. Say there are two features, only one of which is observed, with true data-generating process:

$$Y(X_1, X_2) = \beta_1 X_1 + X_2 \tag{C3}$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{C4}$$

where (C3) is the (linear) potential outcome function, and (C4) is the joint distribution of (X_1, X_2). We vary the causal effect of the observed feature (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 (C3) is linear, $\text{CR}^2(X_1) = \text{CR}^2_{\text{lin}}(X_1)$; since observed and unobserved features are independent, this also coincides with the predictive $\text{R}^2(X_1)$. 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. Our second simulation maintains the linear model, but allows for correlation between features. As before, there are two features, one of which is observed, with data-generating process:

$$Y(X_1, X_2) = X_2 \tag{C5}$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{C6}$$

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, $\mathrm{CR}^2(X_1) = 0$, 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. As before, there are two features, one of which is observed, but now we introduce a non-linearity into the true potential outcome function:

$$Y(X_1, X_2) = 0.2X_1 + \gamma X_1^2 + X_2 \tag{C7}$$

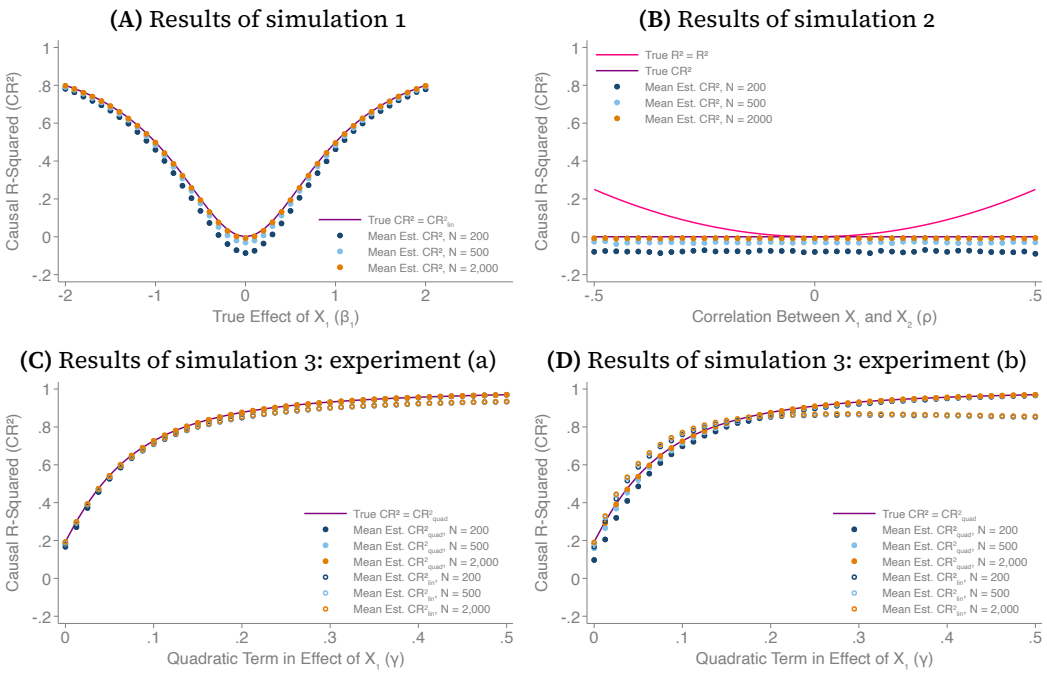
$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma), \quad \mu = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{C8}$$

We vary γ as part of the simulation. Since the true potential outcome function is quadratic, the non-parametric $\mathrm{CR}^2(X_1)$ coincides with $\mathrm{CR}^2_{\mathrm{quad}}(X_1)$, where quad denotes the class of quadratic models ($x_1 \to \mu + \nu x_1 + \pi x_1^2$). As γ increases, $\mathrm{CR}^2(X_1)$ increases, and approaches 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: intuitively, a local linear approximation around the mean of X_1 recovers an average effect "close to" the underlying quadratic effect.

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 a local average treatment effect *in the experiment*; 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.

Online Appendix D. 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 to compute the share of variation causally explained within a subpopulation defined by a covariate. 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). We do this 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 mean squared error 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 E. An impossibility result for monotonic measures of goodness-of-fit

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 formalise 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 X. Given a potential outcome function $Y(\cdot)$, a subset of features 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 $\rho_{P_X,Y}(\mathsf{X}) = \rho_{\hat{P}_X,\hat{Y}}(\mathsf{X})$.[22]

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

[22] 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.

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 standard 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. Say 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 distribution of features by P_X, and its marginal distributions 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 $\rho_{P_X^1,\hat{Y}}(\{X_1\}) = 1$. Since these two cases induce the same distribution of the outcome and observable features, 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, we have shown that no measure can satisfy all three axioms. □

Our conclusion from Proposition A1 is that, once we restrict attention to "reasonable" measures, the possibility of non-monotonicity is inevitable. Intuitively, when we seek to predict an outcome, knowing more features can only be helpful. When we seek to causally explain an outcome, observing an additional feature can "set us back", indicating that the feature suppresses rather than generates variance in the outcome.

Online Appendix F. Details of applications

F.1 Details of application 1

Data. The data are made available by Kremer et al. (2011) 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.

F.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, reading scores, 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 two-stage least-squares regression of the form

$$\text{Class size}_i \times c_i = \gamma + \delta \text{Class size}_i + \epsilon_i \tag{C9}$$

$$\text{Class size}_i = \zeta + \theta \text{Assigned small}_i + \eta_i, \tag{C10}$$

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 $\delta \neq 0$ 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 1 reports the two-stage least-squares coefficients from these regressions. Compliers are similar to non-compliers in terms of gender, race, and socioeconomic status, as proxied by free and reduced lunch receipt. This is consistent with

the observation in Krueger (1999) that the rate of compliance was high.

Appendix Table 1. Complier characteristics in Tennessee STAR

	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 two-stage least-squares regressions corresponding to equations (C9) and (C10) 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), c_i is an indicator variable for the student not being white. In column (3), c_i is the share of years in grades K–3 that the student has free or reduced lunch status.

First stage. Next, we consider the first-stage relation between treatment assignment and class size. Appendix Table 2 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 2. 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 variable is the number of students in the child's class, varying the grade in which this number is measured.

Best predictive and causal models. Next, we estimate the best predictive and causal models. Appendix Table 3 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, which is consistent with omitted variables biasing the observational relation downward relative to the true causal effect.

Appendix Table 3. 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 4 summarizes our results.

Panel A shows the variance of test scores and the mean squared error 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 − MSE/ Var[Score]. These values are computed directly from Panel A. As shown in Figure 2(a), class size explains much more predictively than causally.

Panel C presents various hypothesis tests, computed by bootstrapping. We reject that the causal R^2 for reading is zero at the 1% level: that is, we reject the null hypothesis that class size causally explains no variation in reading scores. We cannot reject the corresponding null hypothesis for math scores at the 5% level, though the difference between reading and math scores is itself insignificant.

Appendix Table 4. 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]	142.34	111.74	81.21	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)	2.61	–	1.59	–
C. Hypothesis tests				
CR^2 for read = 0	0.009			
CR^2 for math = 0	0.059			
CR^2 for read = CR^2 for math	0.1540			

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 mean squared error 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 subsection 5.3.

F.3 Details of application 3

Data. The data are made available by Acemoglu, Johnson, and Robinson (2001) via Daron Acemoglu's website.

Processing and cleaning. Following Acemoglu, Johnson, and Robinson

(2001), we define income as the logarithm of per capita GDP in 1995, and average protection against expropriation as the average of 1985–1995 values of protection against expropriation assigned by Political Risk Services.

Complier characteristics. Appendix Table 5 replicates Appendix Table 1 for this application. There is some evidence that compliers are not geographically representative of the full sample: they are less likely to be drawn from Africa, and perhaps a little more northern. Thankfully, Acemoglu, Johnson, and Robinson (2001) show that their results are similar in Africa *vs.* outside.

Appendix Table 5. Complier characteristics in Acemoglu et al. (2001)

	Africa (1)	Asia (2)	Latitude (3)
Constant	17.02	-5.28	-4.00
	(4.224)	(2.800)	(1.170)
Coefficient	-2.65	0.83	0.62
	(0.644)	(0.427)	(0.178)
Observations	64	64	64

Notes: This table replicates Appendix Table 1 in the context of Acemoglu, Johnson, and Robinson (2001). In column (1), the characteristic is an indicator for the country being in Africa. In column (2), the characteristic is an indicator for the country being in Asia. In column (3), the characteristic is the country's latitude.

First stage. We estimate equation (9). Our results replicate column (9) in Table 3 of Acemoglu, Johnson, and Robinson (2001).

Best predictive and causal models. We then estimate the best predictive and causal models. Appendix Table 6 replicates Appendix Table 3 in the context of Acemoglu, Johnson, and Robinson (2001).[23] Protection against expropriation risk is positively related to log GDP per capita, both predictively and causally. The causal coefficient is somewhat larger than the observational coefficient, which Acemoglu, Johnson, and Robinson (2001) attribute to measurement error in expropriation risk.

[23]We include Appendix Table 6 for completeness, though both panels of the table are contained in Acemoglu, Johnson, and Robinson (2001): the slope coefficient in Panel A corresponds to the coefficient in column (2) of their Table 2, and the slope coefficient in Panel B corresponds to the coefficient in column (1) of their Table 4.

Appendix Table 6. Best predictive and causal models in Acemoglu et al. (2001)

	Log GDP per Capita (1)
A. Best predictive model	
Constant	4.66
	(0.409)
Protection against expropriation risk	0.52
	(0.061)
Observations	64
B. Best causal model	
Constant	1.91
	(1.011)
Protection against expropriation risk	0.94
	(0.154)
Observations	64

Notes: This table replicates Appendix Table 3 in the context of Acemoglu, Johnson, and Robinson (2001).

Fit of the best predictive and causal models. Finally, we compute the fit of the best predictive and causal models (Appendix Table 7). Panel A presents the variance of the outcome and mean squared error for the best predictive and causal models, *i.e.* the variance of the residuals. The best predictive model for GDP, as a linear function of expropriation risk, reduces the mean squared error by more than half; in consequence, the predictive R^2 is very high.[24] The best causal model reduces MSE much less.

Panel B presents the corresponding share of variance predictively and causally explained. The point estimates indicate that institutions predict more than half of the variation in national income, and causally explain around one fifth. Panel C shows, however, that this estimated CR^2 is noisy: for instance, we cannot reject the hypothesis that institutions explain no variation in income. This noise comes from using few countries in the analysis.

[24]Note that the measurement error in expropriation risk, which Acemoglu, Johnson, and Robinson (2001) posit, would imply that the true predictive R^2 is even higher.

Appendix Table 7. Share of variance in GDP causally explained by institutions

	Value (1)
A. Outcome variance and mean squared error	
Var[Log GDP]	1.07
MSE[Best predictive model]	0.49
MSE[Best causal model]	0.87
B. Share of variance explained	
% of var. predictively explained in pop. (obs. R^2)	54.01%
% of var. causally explained in pop. (CR^2)	18.69%
C. Hypothesis tests	
$CR^2 = 0$	0.17

Notes: This table replicates Appendix Table 4 in the context of Acemoglu, Johnson, and Robinson (2001).

Comparison of institutions and childhood malaria as causes of cross-country income differences. To show how to use CR^2 to compare different causes of variation, we use data on the effect of malaria in childhood on national income from Bleakley (2010).[25] We first express the treatment effect in Bleakley (2010) as a treatment effect on adulthood income per unit of childhood malaria incidence. In historical US data, Bleakley (2010) finds an income effect of 0.16 log points for cohorts moving from the 95th to 5th percentile of pre-eradication malaria intensity (Table 5, averaging results using the Occupational Income Score and Duncan's Index). To put this effect into units of incidence, we calculate the difference in pre-eradication malaria childhood incidence between 95th and 5th percentile areas. Maxcy (1923) indicates that annual malaria mortality is 17.8 per 10,000 people in the 95th percentile and zero in the 5th percentile; the case fatality rate is 0.5%. As such, we estimate the difference in total population incidence between the 95th and 5th percentile to be 35.6%. Following GiveWell (2023), we adjust this population incidence to childhood incidence using the fact that childhood incidence in the U.S. is about 1.35 times population incidence, *i.e.* we multiply population incidence by 1.35. The treatment effect in units of childhood malaria incidence is thus 0.33 log points, or a 39.5% increase in adulthood income from eradication.

Second, we gather data on childhood malaria incidence for the sample in Acemoglu, Johnson, and Robinson (2001). The GDP per capita data in that

[25]We base our calculation partly on the methodology in GiveWell (2023).

paper is measured in 1995. Ideally, we would use incidence data from around 1960: cohorts that were adults in 1995. However, historical global data is sparse. We therefore rely on the earliest data from the Malaria Atlas Project (2026), year 2000.[26] While the malaria parasite prevalent in the Americas was probably mostly *Plasmodium vivax* (Bleakley 2010), much Malaria around the world today is *Plasmodium falciparum*. The data includes both *P. vivax* and *P. falciparum* incidence,[27] and we apply the causal effect of Bleakley (2010) to both. We also make a transportability assumption. We use the effect in Bleakley (2010) estimated in the US historically, and apply it to other countries in 2000. We also multiply the incidence data by 1.35 as described above to approximate childhood incidence from total population incidence. We then calculate the CR^2 of malaria in childhood.

F.4 Details of application 4

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.[28] These people were randomised 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 (11) using Figure 1A of Sacks et al. (2001). We estimate the best predictive model in the observational data.

[26]The distribution of Malaria has changed somewhat between 1960 and 2000 (Hay et al. 2004).

[27]The variable we construct is the sum of the incidence rates for both parasite types (the number of newly diagnosed malaria cases per 1,000 population, in a given year).

[28]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.

Fit of the best predictive and causal models. Appendix Table 8 summarizes our results. Panel A shows that, both pooling genders and for men and women separately, the best predictive and causal models reduce mean squared error 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.

Appendix Table 8. 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 explained in pop. (obs. R^2)	5.16%	6.99%	1.78%
% of var. causally explained in pop. (CR^2)	5.11%	6.89%	0.76%

Notes: This table replicates Appendix Table 4 for our application to the share of variation in blood pressure explained by salt intake.