Identifying the Cumulative Causal Effect of a Non-Binary Treatment from a Binary Instrument*

Jacob Goldin Vedant Vohra

August 10, 2023

Abstract

In many settings, the causal effect of a treatment on an outcome depends on the intensity with which the treatment is administered. We study identification of a treatment's cumulative effect — that is, the causal effect of moving from the treatment's minimum intensity to its maximum intensity. With arbitrary heterogeneity across units, we show that the standard LATE assumptions (Angrist and Imbens, 1995) do not constrain this parameter, even among compliers. We consider a range of additional assumptions and show how they can deliver informative bounds. We illustrate our approach with two applications, involving the effect of (1) health insurance on emergency department usage, and (2) attendance in an after-school program on student learning. (*JEL* C01, C21, C26)

*Goldin: University of Chicago and NBER. Vohra: University of California San Diego. We are grateful to Guido Imbens and Kaspar Wuthrich for helpful comments.

1 Introduction

In many settings, the causal effect of a treatment on an outcome depends on the intensity with which the treatment is administered. For example, the effect of a drug depends on its dosage. The effect of education on wages depends on the number of years of completed schooling. When a binary instrument is used to estimate the effect of a non-binary treatment with variable intensity, the two-stage least-squares estimator identifies the Average Causal Response (ACR), a weighted average of causal effects of a unit change in treatment intensity, where the weights depend on the fraction of compliers induced to cross the various treatment intensity levels (Angrist and Imbens, 1995).

In this paper, we study identification of a treatment's cumulative complier effect (CCE) — that is, the causal effect of moving from the treatment's minimum intensity to its maximum intensity, averaged among those individuals induced to a higher treatment intensity because of the instrument. Frequently, this parameter is an important one for policymakers to understand, such as when a new program or policy is being rolled out to a previously unexposed group, or when a researcher investigates whether a treatment exhibits diminishing marginal returns in its intensity. In such cases, the CCE reveals the effect of providing a "full dose" of the program to individuals who would otherwise not receive it, rather than—as with the ACR—a weighted average of dose-specific causal effects based on the particulars of the intervention being evaluated.

The main challenge in identifying the CCE is understanding the dose-response relationship between the treatment and the outcome – i.e., the causal effect of an additional unit of treatment intensity at each level of treatment (referred to as, unit causal effects). Under the standard instrumental variable (IV) assumptions of relevance, independence, and monotonicity, the two-stage least-squares estimator identifies only a specific weighted average of these unit causal effects (i.e., the ACR). Below, we highlight how extrapolating from the ACR to the CCE is subject to two forms of potential bias. First, the ACR over-weights the unit causal effects among compliers whose treatment intensity is more greatly affected by the instrument. Second, the ACR over-weights the unit causal effect for ranges of the dose response function through which compliers are more likely

to pass. Thus, depending on the specific instrument being analyzed, the ACR may yield a very different picture of a treatment's effects than the CCE.

We provide two main results. First, we consider identification of the CCE under the standard IV assumptions of relevance, independence, and monotonicity. Without imposing additional structure, we show that the CCE is entirely unconstrained, in that the data do not permit the researcher to rule out any possible value for the CCE.

Second, we consider identification of the CCE under additional identifying assumptions. The key assumption we consider requires a researcher to abstract from heterogeneity in the unit causal effects within the population of compliers. This assumption is restrictive, but may be a plausible approximation in research settings of interest, as we illustrate through two applications. We show how a researcher may use this assumption, along with additional structure motivated by the setting at hand, to partially identify the CCE as the solution to a constrained linear optimization problem. For example, a researcher may impose sign restrictions on the unit causal effects or on the concavity of the dose-response function, as in Goldin, Lurie and McCubbin (2021), or on the margin through which an instrument affects participation in a treatment, as in Rose and Shem-Tov (2021*a*). Here, we develop the conditions under which this approach identifies bounds on the CCE.

We apply our approach to study two randomized programs involving the free provision of health insurance and an after-school instruction program, respectively. For policymakers considering whether to expand these pilot programs, the CCE is a particularly relevant parameter; it describes the effect of fully providing the program to individuals who would not otherwise be able to participate. With respect to the health insurance expansion we study, for example, we focus on the effect of providing a full year of health insurance coverage to an individual who would other go uninsured. As we discuss in more detail below, this parameter could differ from the ACR in practice because the experimental intervention yielded a heterogeneous first-stage effect on insurance take-up and the relationship between months of insurance and the outcome being studied may be non-linear.

In contrast to the large literature studying identification of the effects of binary treatments with binary instruments, there has been less work studying the use of binary instruments to identify the

effects of non-binary treatments. Angrist and Imbens (1995) provides conditions under which two stage least squares identifies the ACR, but as discussed above and in Heckman, Urzua and Vytlacil (2006), the ACR is tied to a specific instrument rather than the treatment, and therefore may not be well-suited to assessing specific alternative policy interventions. A related literature focuses on extrapolating policy-relevant parameters from IV research designs, but the methods studied in this literature require additional exogenous variation in the form of non-binary instruments (Heckman and Vytlacil, 2007; Imbens and Newey, 2009) or limit their focus to binary treatments (e.g., Mogstad and Torgovitsky, 2018). An exception is Torgovitsky (2015), which point-identifies the dose response function for a continuous treatment under the assumption of rank invariance for both the first-stage and outcome equations. Chernozhukov and Hansen (2005) also accommodates non-binary treatments under a rank invariance condition in the quantile IV setting.

Finally, a common practice by researchers in settings with non-binary treatments is to "binarize" the treatment by collapsing it into two categories; two recent papers study the assumptions underlying the validity of this approach (Andresen and Huber, 2021; Rose and Shem-Tov, 2021*b*). In contrast, the parameter we study is based on the dose-response relationship for the original (un-collapsed) non-binary treatment.

2 Setting and Notation

Consider a population, indexed by i, in which individuals are assigned a binary instrument, $Z_i \in \{0,1\}$, and one level of a discrete treatment, ranging in intensity from 0 to J. Let $D_i(Z) \in \{0,1,2,...,J\}$ denote i's treatment level under each value of the instrument, and let $Y_i(j)$ denote the outcome of interest that would be obtained if i were to receive treatment level j.

We also assume that the following conditions are satisfied.

Assumption 1: Relevance

$E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0] \neq 0$

Assumption 2: Independence

$\{Y_i(0), Y_i(1), ..., Y_i(J), D_i(0), D_i(1)\} \perp\!\!\!\perp Z_i$

Assumption 3: Monotonicity

$\mathbb{P}\left(D_i(1) \geq D_i(0)\right) = 1$

Angrist and Imbens (1995) show that under these assumptions, the standard two-stage least-squares estimator identifies the average causal response (ACR) of D_i on Y_i, i.e. a weighted average of the causal effect from a unit change in the treatment on the outcome, where the weights are the share of corresponding unit changes in the treatment intensity induced by the instrument. More precisely, the ACR corresponds to the right hand side of the following equation:

$$\frac{E[Y \mid Z=1] - E[Y \mid Z=0]}{E[D \mid Z=1] - E[D \mid Z=0]} = \sum_{j=1}^{J} w_j E[Y_i(j) - Y_i(j-1) \mid D_i(1) \geq j > D_i(0)] \tag{1}$$

where

$$w_j = \frac{P[D_i(1) \geq j > D_i(0)]}{\sum_{j=1}^{J} P[D_i(1) \geq j > D_i(0)]}.$$

To analyze the ACR, it will be convenient to introduce the following notation. Let β_{ij} denote the unit-causal effect for individual i at treatment intensity j:

$$\beta_{ij} = Y_i(j) - Y_i(j-1). \tag{2}$$

Similarly, let δ_{ij} indicate whether the instrument induces unit i to cross treatment intensity j, $\delta_{ij} = \mathbb{1}\{D_i(1) \geq j > D_i(0)\}$. In addition, let δ_i denote i's change in treatment intensity induced by the instrument, $\delta_i = \frac{1}{J} \sum_{j=1}^{J} \delta_{ij}$; let δ_j denote the share of units induced by the instrument to cross treatment intensity j, $\delta_j = \frac{1}{N} \sum_{i=1}^{N} \delta_{ij}$; and let $\overline{\delta}$ denote the total share of treatment level increases induced by the instrument, $\overline{\delta} = \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij}$.

Using this notation, we can express the ACR as the (scaled) sum of unit-causal effects induced by the intervention:

$$ACR = \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij} \beta_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij}}$$

Our goal is to shed light on the average cumulative effect of a treatment. We define the cumulative effect of a treatment on an outcome Y as the causal effect of a shift from $D = 0$ to $D = J$, or $Y_i(J) - Y_i(0)$. Our parameter of interest, the cumulative complier effect (CCE), is defined as the mean cumulative effect for the population of compliers:

$$CCE = E\left[Y_i(J) - Y_i(0) \mid D_i(1) > D_i(0)\right] \tag{3}$$

In the next section, we highlight challenges to identifying the CCE under Assumptions 1-3. However, these assumptions do permit us to identify the weights corresponding to each unit-response in Equation 1, $\{w_j\}_{j=1}^{J}$. In Section 4, we propose using these estimated weights, along with the estimated ACR and setting-specific assumptions about the unit causal effects, to bound the CCE.

3 Challenges to Identifying the CCE

Although both the ACR and CCE are functions of a treatment's unit causal effects, the former cannot be directly extrapolated to identify the latter. Intuitively, whereas the ACR summarizes the average causal effect associated with a specific intervention, the CCE summarizes the average causal effect of the treatment across *all* compliers and *all* treatment intensities. More concretely, suppose that we decompose the unit causal effect as

$$\beta_{ij} = \overline{\gamma} + \gamma_i + \gamma_j + \tilde{\gamma}_{ij}$$

where $\overline{\gamma}$ denotes the average unit causal effect among compliers, $\overline{\gamma} = \frac{CCE}{J}$; γ_i captures unit-level heterogeneity, $\gamma_i = \frac{1}{J}\sum_j \beta_{ij} - \overline{\gamma}$; γ_j captures heterogeneity across treatment intensities, $\gamma_j = \frac{1}{N}\sum_i \beta_{ij} - \overline{\gamma}$; and $\tilde{\gamma}_{ij}$ captures the remaining heterogeneity among unit causal effects, $\tilde{\gamma}_{ij} = \beta_{ij} - \overline{\gamma} - \gamma_i - \gamma_j$. We can then express the relationship between the ACR and the CCE as:

$$ACR = \frac{1}{J}\,CCE + \frac{1}{\overline{\delta}}\left(\mathrm{Cov}(\gamma_i, \delta_i) + \mathrm{Cov}(\gamma_j, \delta_j) + \mathrm{Cov}(\tilde{\gamma}_{ij}, \delta_{ij})\right) \tag{4}$$

A derivation for Equation (4), and all other results, is provided in the Online Appendix. The equation highlights two potential forms of selection that may cause the ACR to diverge from the CCE. First, $\text{Cov}(\delta_i\,,\,\gamma_i)$ captures the fact that those individuals who are induced by the intervention to increase their treatment intensity may also have higher per-unit treatment effects. The second covariance term, $\text{Cov}(\delta_j\,,\,\gamma_j)$, captures the fact that the additional levels of the treatment induced by the intervention may have higher per-unit treatment effects than other levels of the treatment (that were not induced by the intervention). Finally, the third covariance term, $\text{Cov}(\tilde{\gamma}_{ij},\ \delta_{ij})$ reflects the potential interaction between these two types of selection.

Given the potential wedge between the ACR and the CCE, a natural question to ask is what can be learned about the CCE under the assumptions that identify the ACR? Unfortunately, as the next result shows, the answer is "not much." Intuitively, since not all compliers move along the full length of the dose-response function from treatment intensity 0 to J, but still contribute to the CCE, we need to impose additional structure on the unit-causal effects to extrapolate *across* complier types.

To see this, note that the CCE can be re-written as

$$CCE = \frac{1}{N_c}\left[\delta\, ACR + \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))(1-\delta_{ij})\beta_{ij}\right] \tag{5}$$

where N_c denotes the number of compliers and δ is equal to $\sum_i \sum_j \delta_{ij}$. Since the CCE is increasing in the unit causal effects, it can be made arbitrarily large or small unless the unit-causal effects are constrained in magnitude. However, for i, j such that $\delta_{ij} = 0$, β_{ij} can be made arbitrarily large or small without affecting the ACR. Thus, the standard IV assumptions, on their own, do not meaningfully constrain which values of the CCE are consistent with the data.

In the Online Appendix, we consider the additional restriction of bounds on the distribution of the outcome. In this case, it can be seen from Equation (2) that the unit-causal effects, and hence the CCE, are also bounded. Unfortunately, as discussed in the Online Appendix, these bounds are typically uninformative without additional assumptions on the magnitude of the unit causal effects.

4 Identifying the Cumulative Complier Effect Under Complier Effect Homogeneity

In this section, we abstract from potential heterogeneity in the unit causal effects among compliers to facilitate identification of the CCE. The following assumption is restrictive, but as we illustrate in Section 5, it can be plausible in real-world applications of interest.[1]

Assumption 4: Homogeneous Incremental Effect Among Compliers

$D_i(1) > D_i(0) \implies Y_i(j) - Y_i(j-1) = \beta_j$

The value of this assumption is that it focuses attention on the uncertainty in the identification of the CCE that arises due to the unknown shape of the treatment's dose-response function. Under this homogeneity assumption, we can express the ACR and CCE as

$$ACR = \sum_{j=1}^{J} w_j \beta_j$$

and

$$CCE = \sum_{j=1}^{J} \beta_j$$

To shed light on the range of cumulative effects of the treatment consistent with the data, we find bounds on the CCE by casting it as a linear optimization problem. Under Assumptions 1-4, the optimization problem can be written as:

$$\begin{aligned} \underset{\{\beta_j\}}{\text{Maximize/Minimize}} \quad & \sum_{j=1}^{J} \beta_j && \textbf{LP.1} \\ \text{subject to} \quad & \sum_{j=1}^{J} \beta_j w_j = ACR && \text{(C.1)} \end{aligned}$$

Denote the maximum feasible value of this objective function, $\sum_{j=1}^{J} \beta_j$, by $\overline{CCE}$ and the minimum by $\underline{CCE}$. Since β_j can be arbitrarily large or small, CCE is generally unbounded under

[1] In addition, as we argue in our application, Assumption 4 may be more likely to hold after conditioning on observable characteristics along the lines of Angrist and Fernandez-Val (2010).

Assumption 1-4. More formally, for any $a \in \mathbb{R}$, non-uniform set of weights $\{w\}$, and observed ACR value, there exists a feasible solution vector $(\beta_1, \beta_2, ..., \beta_J)$ such that $CCE = a$. Hence, additional assumptions regarding the unit causal effects are needed to obtain meaningful bounds on CCE.

We now consider a range of additional assumptions that may be appropriate to impose depending on the application, in the spirit of Manski (2003). In some settings, the direction of the treatment effect will be known from theory or prior research. In such cases, without loss of generality, we can impose that the direction of each unit effects is positive:

Assumption 5: Uniform Sign of Unit Causal Effects

$$\beta_j \geq 0 \; \forall \; j$$

Under Assumptions 1-5, the CCE can be bounded based on the ACR and the empirically observable weights.[2]

Proposition 1: Under Assumptions 1-5,

$$CCE \in \left[\frac{ACR}{w_{\overline{j}}} \, , \, \frac{ACR}{w_{\underline{j}}} \right]$$

where $\underline{j} = \arg\min_{j \in \{1,...J\}} \{w_j\}$ and $\overline{j} = \arg\max_{j \in \{1,...J\}} \{w_j\}$.[3]

In some settings, it will be reasonable to impose additional assumptions beyond restrictions on the sign of the treatment effect. For example, the researcher may have good reason to believe that the magnitude of the unit causal effects is non-increasing in the intensity of the treatment. We refer to this assumption as concavity of the dose-response function:

Assumption 6: Concavity

$$\beta_j \geq \beta_{j+1} \; \forall \; j = 1, ..., J-1$$

[2] In the Online Appendix, we consider the bounds on the CCE under a modification of Assumption 5, where the unit causal effects are additionally bounded from above.

[3] Note that if one of the weights is zero, then $w_{\underline{j}} = 0$; in this case, $CCE \in \left[\frac{ACR}{w_{\overline{j}}} \, , \, \infty \right)$.

Under Assumptions 1-6, bounds on the CCE can be obtained from the linear optimization problem:

$$\underset{\{\beta_j\}}{\text{Maximize/Minimize}} \quad \sum_{j=1}^{J} \beta_j \qquad \textbf{LP.2}$$

$$\text{subject to} \quad \sum_{j=1}^{J} \beta_j w_j = ACR \qquad \text{(C.1)}$$

$$\beta_j \geq 0 \qquad \forall\ j = 1, ..., J \qquad \text{(C.2)}$$

$$\beta_j \geq \beta_{j+1} \qquad \forall\ j = 1, ..., J-1 \qquad \text{(C.3)}$$

Goldin, Lurie and McCubbin (2021) solved this linear problem to estimate bounds on the cumulative effect of health insurance coverage in their setting.

Finally, in some settings the share of compliers that are induced by an instrument to cross a particular treatment intensity threshold will be non-decreasing in the threshold level:

Assumption 7: Monotonic Complier-Share Weights

$w_j \geq w_{j+1}\ \forall\ j = 1, ..., J-1$, and the inequality is strict for at least one such j.

Unlike Assumptions 5 and 6, Assumption 7 is empirically verifiable because the complier-share weights are identified under our maintained Assumptions 1-3. A sufficient condition for Assumption 7 to hold is that all individuals who increase treatment intensity in response to the instrument do so on the extensive margin (Rose and Shem-Tov, 2021*b*).

In settings where these conditions hold, we can obtain the following sharp bounds on the CCE:

Proposition 2: Under Assumptions 1-7,

$$CCE \in \left[\frac{ACR}{w_1},\ ACR \times J\right]$$

When Proposition 2 applies, the CCE is maximized when the dose-response function is linear and minimized when intensive margin changes in the intensity of the treatment have no effect on the outcome.

5 Applications

We illustrate the method by applying it in two empirical settings: the Oregon Health Insurance Experiment (Taubman et al., 2014) and a randomized intervention studying the effect of a computer-aided learning program for middle-school students on test scores in India (Muralidharan, Singh and Ganimian, 2019). Table 1 shows the ACR and the bounds for the CCE for both settings. Following Andrews (2000), we provide confidence intervals on the CCE bounds using a modified bootstrapping procedure that provides accurate coverage in linear optimization settings like the one we study.

5.1 Health Insurance and Emergency Department Usage

In 2008, certain low-income adults in Oregon were selected through a lottery to enroll in Medicaid. Taubman et al. (2014) use this random variation to study the effect of Medicaid on Emergency Department (ED) usage. To shed light on the cumulative effect of obtaining a full period of Medicaid coverage, we define the treatment intensity as the number of months an individual was enrolled in Medicaid during the 19-month study period. The data imply an ACR of Medicaid on ED use of 0.53 percentage points. While this captures the average effect of the additional months of Medicaid coverage induced by the particular intervention being studied, hospitals and policymakers may also be interested in understanding the effect of providing annual coverage to previously uninsured individuals.

It is likely that there is heterogeneity in the effect of Medicaid on ED usage across compliers. However, Kowalski (2021) shows that previous ER utilization can explain the vast majority of this heterogeneity; the marginal treatment effect curve is (approximately) flat after conditioning on prior ED use. Motivated by this finding, we apply Assumption 4 conditional on prior ED use and bound the CCE separately for those with and without a history of ED use before the Medicaid lottery. We assume that the effect of Medicaid on ED use is non-negative (Assumption 5), and that the per-month effect is non-increasing in the number of months of enrollment (Assumption 6). However, it is likely that the Medicaid lottery increased months of enrollment for some individuals who would have enrolled in coverage even absent the treatment, suggesting that Assumption 7 may

not hold in this context.[4] Hence, we obtain bounds on the CCE by solving the linear program in **LP.2**.

In this context, the CCE captures the causal effect on ED usage of enrolling in Medicaid for the full 19-month study period. Our results suggest this effect was an increase in the share of individuals using the ED at least once of between 6.7 and 9.8 percentage points for those *without* a prior history of ED use, and between 7.4 and 10.6 percentage points for those *with* a prior history of ED use. We can combine these into a CCE for the overall sample by taking a weighted average of the two, where the weights correspond to the observed distribution of prior ED use. The implied CCE using this approach is between 6.9 and 10.1 percentage points. Similarly, we can also derive bounds on the CCE associated with providing uncovered individuals with a full year of coverage by replacing the objective function with $\sum_{j=1}^{12} \beta_j$. Doing so implies an effect of a full year of Medicaid coverage of between 6.3 and 7.6 percentage points.

5.2 After-School Instruction and Educational Outcomes

Muralidharan, Singh and Ganimian (2019) study the effect of a technology-aided after-school instruction program on test scores in urban India, using a lottery that provided winners with free access to the program. Using the outcome of the lottery to instrument for weeks of program attendance, the data from the paper imply an ACR for the program of a 0.045 standard deviation increase in math test scores and a 0.030 standard deviation increase for Hindi (see Table 1).[5] While the ACR identifies a specific weighted average effect of attending the Mindspark centers for a week on test scores, a policymaker considering whether to scale up the program might be particularly interested in the overall effect of providing a full course of after-school instruction to a new set of students.

The authors suggest that it is likely that the effects of the program are homogeneous across students using three pieces of evidence. First, the observed treatment effects are similar across the distribution of student achievement prior to the program – a likely source of heterogeneity.

[4] That being said, Appendix Figure 1(a) and 1(b) suggests only slight deviations from monotonicity.

[5] In their original analysis, the authors focused on days of attendance rather than weeks. Here, we pool the treatment intensity to the week-level to reduce statistical variability.

Second, the authors cannot reject equality of the IV and OLS value-added estimates, suggesting the ATE and LATE might be similar and heterogeneity in the compliant subgroups might be small. Finally, the authors note that the outcomes for the control group and never-takers are similar, suggesting equality of potential outcomes across different compliance groups. Taken together, these observations provide suggestive evidence in support of Assumption 4.

In addition, Assumptions 5 and 6 both seem likely to hold in this context: one would not expect the program to *reduce* learning and it seems plausible (though not guaranteed) that there would be non-increasing returns to scale to program attendance. Because only students who won the lottery could attend the program, we know that all compliers were affected on the extensive margin. Hence, as discussed in Section 4, the weights are guaranteed to be monotonically declining so that Assumption 7 holds. Appendix Figure 1(c) empirically verifies that this is the case.

When these conditions hold, Proposition 2 provides sharp bounds for the CCE of weeks of attendance. In particular, translating the ACR point estimate into bounds for the CCE implies that the full program course (12 weeks) would increase math scores by between 0.40 and 0.54 standard deviations and Hindi scores by between 0.27 and 0.36 standard deviations (Table 1, Column 2). Like the main IV estimate, these bounds grow substantially less precise, but continue to exclude zero, once statistical uncertainty is taken into account (Table 1, Column 3).

6 Conclusion

Researchers are often interested in evaluating the effect of non-binary treatments. In such settings, the cumulative effect of the treatment, i.e., the effect of moving from minimum to maximum treatment intensity, is a natural parameter of interest. In this paper, we have highlighted conditions that allow for partial identification of this parameter among compliers and show that, in their absence, meaningful identification is generally not feasible. While the required assumptions are strong, we illustrate with two applications that they are plausible in settings of interest to applied researchers.

References

Andresen, Martin E, and Martin Huber. 2021. "Instrument-based estimation with binarised treatments: issues and tests for the exclusion restriction." *The Econometrics Journal*, 24(3): 536–558.

Andrews, Donald W. K. 2000. "Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space." *Econometrica*, 68(2): 399–405.

Angrist, Joshua, and Ivan Fernandez-Val. 2010. "Extrapolate-ing: External validity and overidentification in the late framework." National Bureau of Economic Research.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association*, 90(430): 431–442.

Bickel, Peter J., and Anat Sakov. 2008. "On the choice of m in the m-out-of-n bootstrap and confidence bounds for extrema." *Statistica Sinica*, 18(3): 967–985.

Chernozhukov, Victor, and Christian Hansen. 2005. "An IV model of quantile treatment effects." *Econometrica*, 73(1): 245–261.

Demuynck, Thomas. 2015. "Bounding average treatment effects: A linear programming approach." *Economics Letters*, 137: 75 – 77.

Goldin, Jacob, Ithai Z Lurie, and Janet McCubbin. 2021. "Health insurance and mortality: Experimental evidence from taxpayer outreach." *The Quarterly Journal of Economics*, 136(1): 1–49.

Heckman, James J, and Edward J Vytlacil. 2007. "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments." *Handbook of econometrics*, 6: 4875–5143.

Heckman, James J, Sergio Urzua, and Edward Vytlacil. 2006. "Understanding instrumental variables in models with essential heterogeneity." *The review of economics and statistics*, 88(3): 389–432.

Huang, Emily J., Ethan X. Fang, Daniel F. Hanley, and Michael Rosenblum. 2016. "Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few?" *Biostatistics*, 18(2): 308–324.

Imbens, Guido W. 2007. "Nonadditive models with endogenous regressors." *Econometric Society Monographs*, 43: 17.

Imbens, Guido W, and Whitney K Newey. 2009. "Identification and estimation of triangular simultaneous equations models without additivity." *Econometrica*, 77(5): 1481–1512.

Kowalski, Amanda E. 2021. "Reconciling seemingly contradictory results from the Oregon health insurance experiment and the Massachusetts health reform." *Review of Economics and Statistics*, 1–45.

Manski, Charles F. 2003. *Partial identification of probability distributions.* Vol. 5, Springer.

Mogstad, Magne, and Alexander Torgovitsky. 2018. "Identification and extrapolation of causal effects with instrumental variables." *Annual Review of Economics*, 10: 577–613.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review*, 109(4): 1426–60.

Rose, Evan K, and Yotam Shem-Tov. 2021*a*. "How does incarceration affect reoffending? estimating the dose-response function." *Journal of Political Economy*, 129(12): 3302–3356.

Rose, Evan K, and Yotam Shem-Tov. 2021*b*. "On Recoding Ordered Treatments as Binary Indicators." *arXiv preprint arXiv:2111.12258*.

Taubman, Sarah L, Heidi L Allen, Bill J Wright, Katherine Baicker, and Amy N Finkelstein. 2014. "Medicaid increases emergency-department use: evidence from Oregon's Health Insurance Experiment." *Science*, 343(6168): 263–268.

Torgovitsky, Alexander. 2015. "Identification of nonseparable models using instruments with small support." *Econometrica*, 83(3): 1185–1197.

Table 1: Bounds on the Cumulative Complier Effect (CCE)

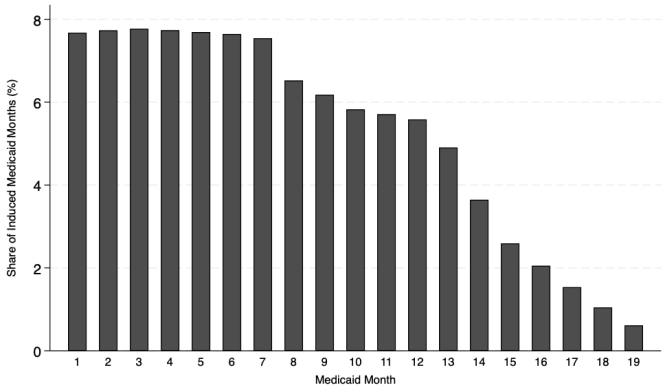
	(1)	(2)	(3)
	Average Causal Response	CCE Bounds	CCE Bounds (with 95% CI)
Taubman et al. (2014)			
Emergency Department Usage	0.529	[6.918, 10.053]	[2.160, 18.064]
	(0.181)		
Muralidharan, Singh and Ganimian (2019)			
Math Test Score	0.045	[0.400, 0.538]	[0.175, 0.722]
	(0.007)		
Hindi Test Score	0.030	[0.265, 0.357]	[0.099, 0.630]
	(0.007)		

Notes: The table illustrates the application of the proposed method to estimating the effect of (1) health insurance coverage on emergency department usage and (2) after-school instruction on test scores. The Average Causal Response (Column 1) is obtained from a two-stage least-squares regression using the data reported in the specified study. The point estimates for the CCE bounds (Column 2) are calculated as described in Section 4. The 95% confidence intervals (Column 3) are obtained from an m-out-of-n bootstrapping procedure (Demuynck, 2015; Huang et al., 2016): instead of drawing samples (with replacement) of size n equal to the sample size, we draw samples of size $m \ll n$. To select the appropriate m for this procedure, we follow the method proposed in Bickel and Sakov (2008).

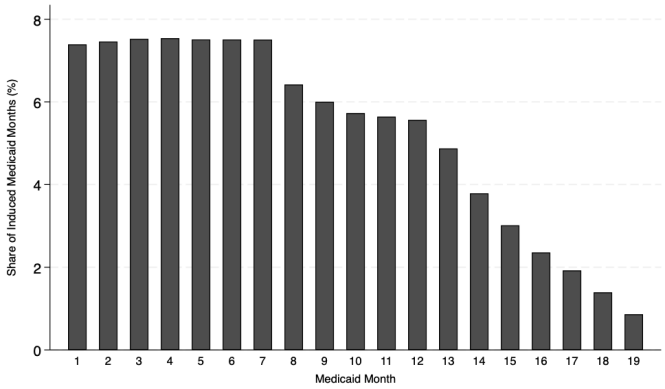
Appendix (For Online Publication)

Table of Contents

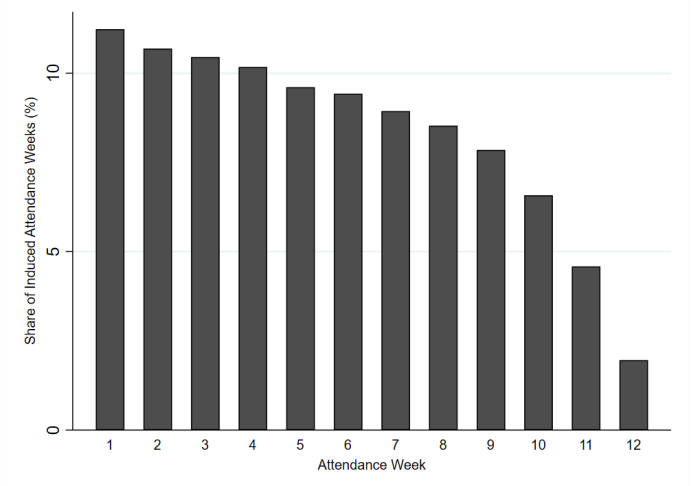
A Figures



(a) Taubman et al. (2014) No Prior ED Use Sample: Medicaid Months



(b) Taubman et al. (2014) Prior ED Use Sample: Medicaid Months



(c) Muralidharan, Singh and Ganimian (2019): Attendance Weeks

Figure A.1: Distribution of Estimation Unit Weights

B Relationship between the ACR and CCE

Without loss of generality, suppose that i is indexed such that $D_i(1) > D_i(0) \iff i \leq N_c$. In this case, the ACR can be written as:

$$ACR = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\delta_{ij}\beta_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{J}\delta_{ij}}$$

and the CCE can be written as

$$\begin{aligned} CCE &= \frac{1}{N_c}\sum_{i=1}^{N}\sum_{j=1}^{J}\mathbb{1}(D_i(1) > D_i(0))\beta_{ij} \\ &= \frac{1}{N_c}\sum_{i=1}^{N_c}\sum_{j=1}^{J}\beta_{ij} \end{aligned}$$

Define $\overline{\gamma}$ as the average unit causal effect among compliers:

$$\overline{\gamma} = CCE/J$$

Contrasting $\overline{\gamma}$ with the ACR, the ACR is a weighted average of unit causal effects for compliers and weights corresponding to a specific treatment level. On the other hand, the $\overline{\gamma}$ considers the average effect of a unit increase in treatment for *all* types of compliers if they had all moved along the full length of the causal response. In contrast to the weighting scheme of the ACR, $\overline{\gamma}$ weighs all compliers equally.

Next, note that each unit causal effect, β_{ij}, can be decomposed as follows:

$$\beta_{ij} = \overline{\gamma} + \gamma_i + \gamma_j + \tilde{\gamma}_{ij}$$

where γ_i captures unit-level heterogeneity, $\gamma_i = \frac{1}{J}\sum_j \beta_{ij} - \overline{\gamma}$; β_j captures heterogeneity across treatment intensities, $\gamma_j = \frac{1}{N_c}\sum_i^{N_c} \beta_{ij} - \overline{\gamma}$; and $\tilde{\gamma}_{ij}$ captures the remaining heterogeneity among unit causal effects, $\tilde{\gamma}_{ij} = \beta_{ij} - \overline{\gamma} - \gamma_i - \gamma_j$. Note that $E[\gamma_i] = E[\gamma_j] = E[\tilde{\gamma}_{ij}] = 0$.

Finally, let δ_i denote i's (mean) change in treatment intensity induced by the instrument, $\delta_i =$

$\frac{1}{J}\sum_{j=1}^{J}\delta_{ij}$; δ_j denote the share of compliers induced by the instrument to cross treatment intensity j, $\delta_j = \frac{1}{N_c}\sum_{i=1}^{N_c}\delta_{ij}$; and $\overline{\delta}$ denote the mean increase in treatment units induced by the instrument, $\overline{\delta} = \frac{1}{N_c}\frac{1}{J}\sum_{i=1}^{N_c}\sum_{j=1}^{J}\delta_{ij}$. Substituting this decomposition and notation into the ACR formula, we have:

$$\begin{aligned}
N_c\, J\, \overline{\delta}\, ACR &= \sum_{i=1}^{N_c}\sum_{j=1}^{J}\delta_{ij}\beta_{ij} \\
&= \sum_{i=1}^{N_c}\sum_{j=1}^{J}\delta_{ij}\left(\overline{\gamma}+\gamma_i+\gamma_j+\tilde{\gamma}_{ij}\right) \\
&= \overline{\gamma}\, N_c\, J\, \overline{\delta} + \sum_i \gamma_i \sum_j \delta_{ij} + \sum_j \gamma_j \sum_i \delta_{ij} + \sum_i\sum_j \tilde{\gamma}_{ij}\delta_{ij} \\
&= \overline{\gamma}\, N_c\, J\, \overline{\delta} + \sum_i \gamma_i\, J\, \delta_i + \sum_j \gamma_j\, N_c\, \delta_j + \sum_i\sum_j \tilde{\gamma}_{ij}\delta_{ij} \\
&= \overline{\gamma}\, N_c\, J\, \overline{\delta} + \frac{N_c}{N_c}\, J \sum_i \gamma_i\, \delta_i + \frac{J}{J}\, N_c \sum_j \gamma_j\, \delta_j + \frac{N_c\, J}{N_c\, J}\sum_i\sum_j \tilde{\gamma}_{ij}\delta_{ij} \\
&= \overline{\gamma}\, N_c\, J\, \overline{\delta} + N_c\, J\, \left(\mathrm{Cov}(\gamma_i,\delta_i) + \mathrm{Cov}(\gamma_j,\delta_j) + \mathrm{Cov}(\tilde{\gamma}_{ij},\delta_{ij})\right)
\end{aligned}$$

Dividing through by $N_c\, J\, \overline{\delta}$, we have:

$$ACR - \overline{\gamma} = \frac{1}{\overline{\delta}}\left(\mathrm{Cov}(\gamma_i,\delta_i) + \mathrm{Cov}(\gamma_j,\delta_j) + \mathrm{Cov}(\tilde{\gamma}_{ij},\delta_{ij})\right)$$

To obtain Equation (4), we simply plug $\overline{\gamma} = CCE/J$ into the above.

C Heterogeneity Within Compliers

This section explores the bounds on the CCE after relaxing Assumption 4, allowing for potential heterogeneity in the unit causal effects among compliers. Recall that the ACR is defined as

$$ACR = \sum_{j=1}^{J} w_j E[Y_i(j) - Y_i(j-1) \mid D_i(1) \geq j > D_i(0)]$$

where

$$w_j = \frac{P[D_i(1) \geq j > D_i(0)]}{\sum_{j=1}^{J} P[D_i(1) \geq j > D_i(0)]}$$

Let $\delta_{ij} = \mathbb{1}\{D_i(1) \geq j > D_i(0)\}$. We begin by noting that the ACR can be written as

$$ACR = \frac{\sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij}\, \beta_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij}} = \frac{1}{\delta} \sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij}\, \beta_{ij}$$

where $\delta = \sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij}$. To see this, first note that $P[D_i(1) \geq j > D_i(0)] = \sum_{i=1}^{N} \delta_{ij}/N$. Therefore,

$$w_j = \frac{\sum_{i=1}^{N} \delta_{ij}}{\sum_{j=1}^{J} \sum_{i=1}^{N} \delta_{ij}}$$

Additionally, observe that

$$E[Y_i(j) - Y_i(j-1) \mid D_i(1) \geq j > D_i(0)] = \frac{\sum_{i=1}^{N} \beta_{ij}\delta_{ij}}{\sum_{i=1}^{N} \delta_{ij}}$$

Putting the above together,

$$\begin{aligned} ACR &= \sum_{j=1}^{J} w_j E[Y_i(j) - Y_i(j-1) \mid D_i(1) \geq j > D_i(0)] \\ &= \sum_{j=1}^{J} \frac{\sum_{i=1}^{N} \delta_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{N} \delta_{ij}} \frac{\sum_{i=1}^{N} \beta_{ij}\delta_{ij}}{\sum_{i=1}^{N} \delta_{ij}} \\ &= \frac{\sum_{j=1}^{J}\sum_{i=1}^{N} \beta_{ij}\delta_{ij}}{\sum_{j=1}^{J}\sum_{i=1}^{N} \delta_{ij}} = \frac{1}{\delta} \sum_{i=1}^{N}\sum_{j=1}^{J} \delta_{ij}\, \beta_{ij} \end{aligned}$$

Rearranging, we get:

$$\delta\, ACR = \sum_{i=1}^{N}\sum_{j=1}^{J} \delta_{ij}\, \beta_{ij} \tag{6}$$

Recall that the parameter of interest, the CCE, is defined as

$$\begin{aligned} CCE &= E[Y_i(J) - Y_i(0) | D_i(1) > D_i(0)] \\ &= E\left[\sum_{j=1}^{J} \beta_{ij} \;\middle|\; D_i(1) > D_i(0)\right] \\ &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))\beta_{ij}}{\sum_{i=1}^{N} \mathbb{1}(D_i(1) > D_i(0))} \end{aligned}$$

Let N_c be the number of compliers, $N_c := \sum_{i=1}^{N} \mathbb{1}(D_i(1) > D_i(0))$.

$$
\begin{aligned}
CCE &= \frac{1}{N_c}\left[\sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))\beta_{ij}\right] \\
&= \frac{1}{N_c}\left[\sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))\Big(\beta_{ij} + \beta_{ij}\delta_{ij} - \beta_{ij}\delta_{ij}\Big)\right] \\
&= \frac{1}{N_c}\left[\sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))\Big(\beta_{ij}\delta_{ij} + (1-\delta_{ij})\beta_{ij}\Big)\right] \\
&= \frac{1}{N_c}\left[\sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))\beta_{ij}\delta_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))(1-\delta_{ij})\beta_{ij}\right] \\
&= \frac{1}{N_c}\left[\sum_{i=1}^{N}\sum_{j=1}^{J} \beta_{ij}\delta_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))(1-\delta_{ij})\beta_{ij}\right] \\
&= \frac{1}{N_c}\left[\delta\, ACR + \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))(1-\delta_{ij})\beta_{ij}\right]
\end{aligned}
$$

where the last equality uses (6). We begin by maximizing the CCE. Note that the CCE is increasing in $\sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))(1-\delta_{ij})\beta_{ij}$, and so it is maximized when β_{ij} is large. If β_{ij} is unbounded for some i with $D_i(1) > D_i(0)$, the CCE can be made arbitrarily large by increasing β_{ij}. Note that one can always increase β_{ij} for all i and j such that $\delta_{ij} = 0$, since these unit causal effects are unconstrained by the ACR, or other observable quantities. Therefore, we assume that $\beta_{ij} \in \left[\underline{\beta}, \overline{\beta}\right]$. Then, the maximum CCE is

$$
\begin{aligned}
\overline{CCE} &= \frac{1}{N_c}\left[\delta\, ACR + \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))(1-\delta_{ij})\overline{\beta}\right] \\
&= \frac{1}{N_c}\left[\delta\, ACR + (J\, N_c - \delta)\overline{\beta}\right]
\end{aligned}
$$

Analogously, the minimum CCE is

$$\underline{CCE} = \frac{1}{N_c}\left[\delta\, ACR + (J\, N_c - \delta)\underline{\beta}\right]$$

Note that the number of compliers, N_c, can be bounded by the approach suggested by Huang et al. (2016). We first consider the case where the outcome is unbounded. Since $\beta_{ij} = Y_i(j) - Y_i(j-1)$, the unit causal effects are also unbounded. Because δ is, by definition, weakly less than $J\, N_c$, the bounds on the CCE can be made arbitrarily large or small depending on the user-specified bounds on the unit causal effects, $\overline{\beta}$ and $\underline{\beta}$.

Next, we specialize to the case where the outcome is bounded. Specifically, suppose that $Y_i \in \left[\underline{Y}, \overline{Y}\right]$. Let $\Delta_Y = \overline{Y} - \underline{Y}$. The CCE is maximized when $Y_i(0) = \underline{Y}$ and $Y_i(J) = \overline{Y}$ for *all* compliers, implying that $\overline{CCE} = \Delta_Y$. While the bounds on the outcome provide a natural bound on the CCE, we show next that these bounds are sharp.

We assume that for all i such that $D_i(1) > D_i(0)$, there exists a $k \in \{1, ..., J\}$ s.t. $\delta_{ik} = 0$. This assumption states that it is consistent with the observed marginal distributions of $D_i(1)$ and $D_i(0)$, that *none* of the individuals moved by instrument move along the full length of the dose-response function.[6] For example, a sufficient condition for this assumption to hold is that the first stage is less than the total sample of treated units, $\sum_i \sum_j \delta_{ij} \leq N$, because in that case the data is consistent with each complier increasing their treatment intensity by only a single unit, i.e., $\forall\, i$, $\exists\, j_i$ s.t. $\delta_{i,j_i} = 1$ and $\delta_{ij} = 0\, \forall\, j \neq j_i$. A similar assumption is considered by Imbens (2007).

For each individual i, let $k(i) \in \{1, 2, ..., J\}$ be defined such that $\delta_{i,k(i)} = 0$. Let $\beta_{i,k(i)} =$

[6] This assumption can be verified in the data using the approach suggested by Huang et al. (2016).

$\Delta_Y - \sum_{j=1}^{J} \delta_{ij}\beta_{ij}$. Additionally, for all $\delta_{ij} = 0$ *and* $j \neq k(i)$, let $\beta_{ij} = 0$. Then,

$$\begin{aligned}
CCE &= \frac{1}{N_c}\left[\sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))\Big(\beta_{ij}\delta_{ij} + (1-\delta_{ij})\beta_{ij}\Big)\right] \\
&= \frac{1}{N_c}\left[\sum_{i=1}^{N} \mathbb{1}(D_i(1) > D_i(0))\left(\sum_{j=1}^{J}\beta_{ij}\delta_{ij} + \sum_{j=1}^{J}(1-\delta_{ij})\beta_{ij}\right)\right] \\
&= \frac{1}{N_c}\left[\sum_{i=1}^{N} \mathbb{1}(D_i(1) > D_i(0))\left(\sum_{j=1}^{J}\beta_{ij}\delta_{ij} + \Delta_Y - \sum_{j=1}^{J}\beta_{ij}\delta_{ij}\right)\right] \\
&= \frac{1}{N_c}\left[\sum_{i=1}^{N} \mathbb{1}(D_i(1) > D_i(0))\Delta_Y\right] = \Delta_Y
\end{aligned}$$

This shows that the $\overline{CCE}$ can attain the value Δ_Y. Similarly, it can be shown that $\underline{CCE}$ can attain a minimum value of $-\Delta_Y$.

We now explore how imposing additional restrictions on the unit causal effects might help tighten the bounds. Consider a modification of Assumption 5, $\beta_{ij} \in \Big[\underline{\beta}, \overline{\beta}\Big]$ for all i and all j, where $\Big[\underline{\beta}, \overline{\beta}\Big] \subseteq \Big[-\Delta_Y, \Delta_Y\Big]$. Then, as above, the maximum of the CCE is given by

$$\overline{CCE} = \frac{1}{N_c}\left[\delta\, ACR + \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))(1-\delta_{ij})\overline{\beta}\right]$$

We show that this additional assumption sharpens the bounds on the CCE, i.e., $\overline{CCE} \leq \Delta_Y$

$$\begin{aligned}
\overline{CCE} &= \frac{1}{N_c}\left[\delta\, ACR + \sum_{i=1}^{N}\sum_{j=1}^{J} \mathbb{1}(D_i(1) > D_i(0))(1-\delta_{ij})\overline{\beta}\right] \\
&= \frac{1}{N_c}\left[\delta\, ACR + \sum_{i=1}^{N} \mathbb{1}(D_i(1) > D_i(0))\sum_{j=1}^{J}\overline{\beta} - \sum_{i=1}^{N}\sum_{j=1}^{J}\delta_{ij}\overline{\beta}\right] \\
&\leq \Delta_Y + \frac{\delta}{N_c}\left[\, ACR - \overline{\beta}\right] \\
&\leq \Delta_Y
\end{aligned}$$

Similarly, $\underline{CCE} \geq -\Delta_Y$. This shows that the bounds can be tightened by imposing additional restrictions on the unit causal effects, similar to Assumption 5.

Next, we consider the implication of additionally assuming that the dose response function for each individual is (weakly) concave, i.e., $\beta_{i,j} \geq \beta_{i,j+1}$ for all i and $j < J$. Recall that in the absence of the concavity assumption, one can always choose $\beta_{ij} \in [\underline{\beta}, \overline{\beta}]$ for all i and j such that $\delta_{ij} = 0$, since these unit causal effects are unconstrained by the ACR, or other observable quantities. However, the concavity assumption constrains these unit causal effects. To see this, suppose that i is moved by the instrument from 0 to k units of treatment intensity. This implies that, β_{ik} which is constrained by the ACR, provides an upper bound for all β_{ij} where $j > k$ even though $\delta_{ij} = 0$. Since we cannot identify where the compliers are moved along the dose response function, this assumption alone cannot be used to further tighten the bounds. Indeed, this motivates an assumption of homogeneous unit causal effects within compliers, considered in Section 4 of the main text.

D Proof for Proposition 1

To establish that $CCE \geq ACR/w_{\bar{j}}$, we will first show that CCE is minimized when $\beta_{\bar{j}}$ is the only non-zero unit causal effect, i.e., $\beta_j = 0$ for all $j \neq \bar{j}$. Proceeding by contradiction, suppose instead that CCE was minimized by a vector of unit treatment effects $(\beta_1, ..., \beta_J)$ with $\beta_k > 0$ for some $k \neq \bar{j}$. Consider a new vector of unit treatment effects, $(\beta_1^*, ..., \beta_J^*)$ identical to the first, except that the specified non-zero effect has been set to zero, $\beta_k^* = 0$, and the unit effect corresponding to the largest weight, $\beta_{\bar{j}}$, has been increased to maintain feasibility with respect to the ACR constraint: $\beta_{\bar{j}}^* = \beta_{\bar{j}} + \beta_k \, w_k / w_{\bar{j}}$. But, with these changes, the value of the objective function is lower than that of the original vector of unit effects:

$$\sum \beta_j^* - \sum \beta_j = \beta_k \, \frac{w_k}{w_{\bar{j}}} - \beta_k < 0,$$

where the last inequality follows from Assumption 5 and the definition of $\bar{j}$. Thus, the initial vector of unit treatment effects must not be the solution to the constrained minimization problem, proving the contradiction. The proof that $CCE \leq ACR/w_{\underline{j}}$ is analogous. ■

E Proof for Proposition 2

To show that $CCE \geq ACR/w_1$, we will first show that the CCE is minimized when $\beta_j = 0$ $\forall\ j > 1$. Proceeding by contradiction, suppose that the CCE is minimized by a vector of unit treatment effects $(\beta_1, ..., \beta_J)$ with $\beta_j > 0$ for some $j > 1$. Let k denote the highest intensity non-zero unit effect, more precisely $k = \max_j\{2, ..., J \mid \beta_j > 0\}$. Along with Assumption 6, this implies $\beta_1 \geq \beta_2 \geq ... \geq \beta_k > 0$. Consider a new vector of unit treatment effects, $(\tilde{\beta}_1, ..., \tilde{\beta}_J)$, which is identical to the first vector, except that β_k has been set to zero, $\tilde{\beta}_k = 0$, and the first unit effect has been increased to maintain feasibility with respect to the ACR constraint: $\tilde{\beta}_1 = \beta_1 + \beta_k w_k/w_1$. But, with these changes, the value of the objective function is lower than that of the original vector of unit effects:

$$\sum \tilde{\beta}_j - \sum \beta_j = \beta_k \frac{w_k}{w_1} - \beta_k < 0,$$

where the last inequality follows from Assumptions 5 and 7. Thus, the initial vector of unit treatment effects must not be the solution to the constrained minimization problem, proving the contradiction. Finally, to prove the result, note that when $\beta_j = 0\ \forall\ j > 1$, it follows that $CCE = \sum_j \beta_j = \beta_1 = ACR/w_1$, where the last equality follows from the definition of the ACR.

To show that $CCE \leq ACR \times J$, we will first show that the CCE is maximized when the unit causal effects are equalized, i.e., when $\beta_1 = \beta_2 = ... = \beta_J$. Proceeding by contradiction, suppose that the CCE is maximized by a vector of unit treatment effects $(\beta_1, \beta_2, ..., \beta_J)$ with $\beta_k > \beta_{k+1}$ for some $k \in \{1, ..., J-1\}$ (concavity rules out the reverse ordering). Consider a new vector of unit treatment effects, $(\tilde{\beta}_1, ..., \tilde{\beta}_J)$, which is identical to the first vector, except that $\tilde{\beta}_k = \beta_k - \frac{(\beta_k - \beta_{k+1})w_{k+1}}{w_k + w_{k+1}}$ and $\tilde{\beta}_{k+1} = \beta_{k+1} + \frac{w_k(\beta_k - \beta_{k+1})}{w_k + w_{k+1}}$, so that $\tilde{\beta}_k = \tilde{\beta}_{k+1}$. Note that $(\tilde{\beta}_1, ..., \tilde{\beta}_J)$ is a feasible solution vector since it satisfies C.1-C.3. But, with these changes, the value of the objective function is higher than that of the original vector of the unit effects:

$$\sum \tilde{\beta}_j - \sum \beta_j = \frac{(\beta_{k+1} - \beta_k)(w_{k+1} - w_k)}{w_{k+1} + w_k} > 0,$$

where the last inequality follows from the assumption that $\beta_k > \beta_{k+1}$ (Assumption 6) and that $w_k > w_{k+1}$ (Assumption 7). Thus, the initial vector of unit treatment effects must not be the solution to the constrained maximization problem, proving the contradiction. ■

F Bounded Unit-Causal Effects with Homogeneity

This section considers the bounds for the CCE when Assumption 5 is modified such that the unit causal effects are additionally bounded from above, i.e., $\beta_j \in [\underline{\beta}, \overline{\beta}]$. If we are interested in solving **LP.1** under Assumptions 1-4 and the modified Assumption 5, we may re-write the problem as:

$$\underset{\{\beta_j\}}{\text{Maximize/Minimize}} \quad \sum_{j=1}^{J} \beta_j \tag{LP.3}$$

$$\text{subject to} \quad \sum_{j=1}^{J} \beta_j w_j = ACR \tag{C.1}$$

$$\beta_j \geq \underline{\beta} \tag{C.2}$$

$$\beta_j \leq \overline{\beta} \tag{C.3}$$

We begin by minimizing the above optimization problem. Following the same logic used in the proof for Proposition 1, note that the CCE is minimized when $\forall\ j \neq \overline{j},\ \beta_j = \underline{\beta}$. Proceeding by contradiction, suppose instead that CCE was minimized by a vector of unit treatment effects (β_1,...β_J) with $\beta_k > \underline{\beta}$ for some $k \neq \overline{j}$. Consider a new vector of unit treatment effects, $(\beta_1^*, ..., \beta_J^*)$ identical to the first except that β_k has been set to $\underline{\beta}$, $\beta_k^* = \underline{\beta}$, and the unit effect corresponding to the largest weight, $\beta_{\overline{j}}$, has been increased to maintain feasibility with respect to the ACR constraint: $\beta_{\overline{j}}^* = \beta_{\overline{j}} + (\beta_k - \underline{\beta})\frac{w_k}{w_{\overline{j}}}$. But, with these changes, the value of the objective function is lower than that of the original vector of unit effects:

$$\sum \beta_j^* - \sum \beta_j = (\underline{\beta} - \beta_k)\left(1 - \frac{w_k}{w_1}\right) < 0.$$

Now, we can re-write the ACR constraint as

$$\begin{aligned}\sum_{j=1}^{J}\beta_j w_j &= \beta_{\bar{j}} w_{\bar{j}} + \sum_{j\neq\bar{j}}\beta_j w_j \\ &= \beta_{\bar{j}} w_{\bar{j}} + \underline{\beta}\sum_{j\neq\bar{j}} w_j \\ &= \beta_{\bar{j}} w_{\bar{j}} + \underline{\beta}(1-w_{\bar{j}}) = ACR \\ \implies \beta_{\bar{j}} &= \frac{ACR}{w_{\bar{j}}} - \frac{\underline{\beta}(1-w_{\bar{j}})}{w_{\bar{j}}}\end{aligned}$$

This implies that,

$$\min(CCE) = \frac{ACR}{w_{\bar{j}}} - \frac{\underline{\beta}(1-w_{\bar{j}})}{w_{\bar{j}}} + (J-1)\underline{\beta}$$

Using a similar logic to the above, we can show that the CCE is maximized when $\forall j \neq \underline{j}$, $\beta_j = \underline{\beta}$. Then ,

$$\begin{aligned}\beta_{\underline{j}} &= \frac{ACR}{w_{\underline{j}}} - \frac{\underline{\beta}(1-w_{\underline{j}})}{w_{\underline{j}}} \\ \implies \max(CCE) &= \frac{ACR}{w_{\underline{j}}} - \frac{\underline{\beta}(1-w_{\underline{j}})}{w_{\underline{j}}} + (J-1)\underline{\beta}\end{aligned}$$

As should be expected, these bounds are weakly tighter than the bounds provided in Proposition 1.