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

Vedant Vohra Jacob Goldin

August 2, 2024

Abstract

The effect of a treatment may depend on the intensity with which it is administered. We study identification of ordered treatment effects with a binary instrument, focusing on the effect of moving from the treatment's minimum to maximum intensity. With arbitrary heterogeneity across units, standard IV 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 sharp, 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)

*Vohra: University of California San Diego. Goldin: University of Chicago and NBER. 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 the causal effect of a change in the intensity of a non-binary ordered treatment when the researcher has access to a valid binary instrument. We focus primarily on a parameter we refer to as the cumulative complier effect (CCE), which captures the average effect of moving from a treatment's minimum intensity to its maximum intensity among those who are induced by the instrument to move to a higher treatment intensity. 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). We highlight how extrapolating from the ACR to the CCE is subject to two forms of potential biases. 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 be induced by the instrument. Thus, depending on the specific instrument being analyzed, the ACR may yield a very different picture of a treatment's effects than the CCE. In fact, under the standard set of IV assumptions, 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.

In this paper, 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). Here, we develop the conditions under which this approach identifies bounds on the CCE. In some cases, the solution to the constrained optimization problem takes the form of sharp analytic bounds, which we characterize below.

We apply our approach to study two randomized evaluations, 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 relative to providing no coverage. 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 months of insurance taken up, and the relationship between months of insurance and medical outcomes may be non-linear.

Moreover, some treatments are designed to be taken up for their entire duration, as with the after-school instruction program that we study. In such settings, the CCE is a natural parameter of interest because it captures the "all-or-nothing" manner in which the treatment is intended to be provided.

Although our primary focus is on the CCE, we also consider treatment intensity changes within the interior of the dose-response function. We find that the assumptions that permit partial identification of the CCE may also yield informative bounds for these related parameters. We provide details of this analysis in the Online Appendix.

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 alternative policy interventions. A related literature focuses on extrapolating parameters beyond the LATE from IV research designs, but the methods studied in this literature typically 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., Balke and Pearl, 1997; Mogstad and Torgovitsky, 2018; Mogstad, Santos and Torgovitsky, 2018). One exception is Kamat, Norris and Pecenco (2023), which considers partial identification of a range of treatment effect parameters in the presence of multiple treatments and a discrete-valued instrument; identification is facilitated by restrictions on selection into treatment and unobserved heterogeneity. A second 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. Finally, a recent working paper by Chernozhukov et al. (2024) studies treatment effect heterogeneity in an IV framework that accommodates non-binary treatments by imposing restrictions on the relationship between the treatment assignment and potential outcomes. In contrast to these methods, identification under our approach relies on restricting heterogeneity across compliant subgroups in conjunction with economically motivated restrictions on the shape

of the treatment's dose-response function.[1]

In addition, our work adds to a large literature that studies partial identification of treatment effect models; recent examples include Fan, Sherman and Shum (2014), Arnold, Dobbie and Hull (2022), Rambachan and Roth (2023), and Tebaldi, Torgovitsky and Yang (2023). With respect to IV models in particular, many of these papers study what can be learned when the standard IV assumptions do not hold (e.g., Manski and Pepper, 2000; De Chaisemartin, 2017); see Swanson et al. (2018) for a review. However, as we show below, even when the standard IV assumptions do hold, the data do not constrain the CCE when there is a single binary instrument and a non-binary treatment. Hence, our approach is to study how imposing additional structure beyond Angrist and Imbens (1995) allows a researcher to partially identify the CCE.

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, 2024). In contrast, the parameter we study is based on the dose-response relationship for the original (uncollapsed) 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 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

[1] Lochner and Moretti (2015) impose a similar restriction on treatment effect heterogeneity to motivate a proposed test for exogeneity of a multi-valued treatment using a binary instrument.

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

Assumption 3: Monotonicity

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

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:

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

where

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

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:

$$CCE = E[Y_i(J) - Y_i(0) \mid D_i(1) > D_i(0)] \quad (2)$$

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. 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. There are therefore two potential forms of selection that may cause the ACR to diverge from the CCE. First, those individuals who are induced by the intervention to increase their treatment intensity may also have higher per-unit treatment effects. Second, 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).[2]

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, the answer is "not much." In fact, the CCE is entirely unconstrained under Assumptions 1-3, as we show formally in Appendix C. 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.[3]

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. We consider the following assumption:

Assumption 4: Homogeneous Incremental Effect Across Compliant Subgroups

$E[Y_i(j) - Y_i(j-1) \mid D_i(1) \geq k > D_i(0)] = \beta_j$ for all $k = 1, ..., J$

[2] In Appendix B, we formally relate the ACR and CCE in terms of these biases.

[3] When the unit causal effects are known to be bounded, the CCE is bounded as well; see Appendix C for details.

This assumption is restrictive, but as we illustrate in Section 5, it can be plausible in real-world applications of interest (potentially after conditioning on observable characteristics along the lines of Angrist and Fernandez-Val (2010)). The value of the 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. When Assumption 4 holds, the CCE corresponds to the effect of moving from minimum treatment intensity to maximum treatment intensity for any compliant subgroup. Lochner and Moretti (2015) impose a similar assumption to test exogeneity of a multi-valued treatment.

Under Assumption 4, 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:

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

Note that the ACR and set of weights, $\{w_j\}_{j=1}^{J}$, in **LP.1** are identified and estimable under Assumptions 1-3 (Angrist and Imbens, 1995). The ACR can be estimated using the sample moments corresponding to the expectations on the left-hand side of Equation (1). The weights can be estimated by comparing the share of units at each treatment level under $Z = 0$ versus $Z = 1$.

Denote the maximum feasible value of this objective function, $\sum_{j=1}^{J} \beta_j$, by $\overline{CCE}$ and the min-

imum by $\underline{CCE}$. Since β_j can be arbitrarily large or small, the CCE is generally unbounded under 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 the CCE.[4]

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.

Proposition 1: Under Assumptions 1-5, the following sharp bounds hold:

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

where $\underline{j} = \arg\min_{j \in \{1,...J\}} \{w_j\}$ and $\overline{j} = \arg\max_{j \in \{1,...J\}} \{w_j\}$.[5]

The proof of Proposition 1, and all subsequent results, is contained in the Online Appendix. The bounds provided in the Proposition are sharp, in the sense that all values of the CCE in the identified set are compatible with Assumptions 1-5 and the data (see Appendix G for details).

In some settings, it will be reasonable to impose additional assumptions beyond restrictions on

[4] In settings where additional exogenous variation is available in the form of multiple instruments, Angrist and Imbens (1995) show that each instrument can be used to identify an ACR and a set of compliance weights. When the number of instruments is equal to the number of treatment levels, the CCE may be point-identified. More generally, when the number of treatment levels exceeds the number of instruments, the researcher can incorporate the variation from the additional instruments to tighten the bounds on the CCE by adding them as constraints to LP.1. (see Appendix D for details).

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

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.[6] We refer to this assumption as concavity of the dose-response function:

Assumption 6: Concavity

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

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

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

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

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

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

As described above, the ACR and weights in **LP.2** are identified and estimable under Assumptions 1-3. 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 Assump-

[6] A different potential restriction that a researcher might impose is that the magnitude of each unit causal effect is bounded between known values. We extend Proposition 1 to this setting in Appendix H.1. A related restriction is that the outcome of interest has bounded support. In addition to implying bounds on the unit causal effects, this assumption constrains the values the unobserved potential outcomes can take on. We discuss this setting in Appendix H.2.

tion 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, 2024). In settings where these conditions hold, the bounds that emerge from LP.2 take the following simple analytic form:

Proposition 2: Under Assumptions 1-7, the following sharp bounds hold:

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

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. The bounds provided in Proposition 2 are sharp, in the sense that all values of the CCE in the identified set are compatible with Assumptions 1-7 and the data (see Appendix G for details).

Finally, in some settings the researcher may be interested in parameters relating to the dose-response function other than the CCE, such as when a potential intervention is intended to shift behavior across a subset of treatment values. In the appendix, we study what may be identified for a broader class of parameters, relating to the effect of moving from treatment intensity j_1 to j_2 where $j_2 > j_1$ and $j_1, j_2 \in \{0, 1, ..., J\}$. We show that the sharp analytic bounds provided in Propositions 1 and 2 extend naturally to this setting; see Online Appendix I for details.

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.[7] 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

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

approach is between 6.9 and 10.1 percentage points. Similarly, we can also derive bounds on the effect associated with providing uncovered individuals with a full year of coverage by replacing the objective function with $\sum_{j=1}^{12} \beta_j$.[8] Doing so implies an effect of a full year of Medicaid coverage of between 6.3 and 7.6 percentage points.[9]

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).[10] 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. Indeed, since the program was designed to be taken up for its entire duration, the CCE is a natural parameter of interest.

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. 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 equal-

[8] Note that under Assumption 4, this parameter reflects the effect of a full year of Medicaid coverage for those compliers who would have enrolled in partial-year coverage absent the intervention ($D_i(0) > 0$), as well as for those compliers who would not have otherwise enrolled in any coverage ($D_i(0) = 0$).

[9] In general, the width of the bounds on $\sum_{j=1}^{12} \beta_j$ is a function of the particular range of the dose-response function of interest (in this case, 0 to 12 months of coverage) and the distribution of treatment intensity changes induced by the instrument; see Online Appendix I for details.

[10] 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.

ity 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.

Arnold, David, Will Dobbie, and Peter Hull. 2022. "Measuring Racial Discrimination in Bail Decisions." *American Economic Review*, 112(9).

Balke, Alexander, and Judea Pearl. 1997. "Bounds on treatment effects from studies with imperfect compliance." *Journal of the American statistical Association*, 92(439): 1171–1176.

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, Iván Fernández-Val, Sukjin Han, and Kaspar Wüthrich. 2024. "Estimating Causal Effects of Discrete and Continuous Treatments with Binary Instruments." *arXiv preprint arXiv:2403.05850*.

De Chaisemartin, Clement. 2017. "Tolerating defiance? Local average treatment effects without monotonicity." *Quantitative Economics*, 8(2): 367–396.

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

Fan, Yanqin, Robert Sherman, and Matthew Shum. 2014. "Identifying treatment effects under data combination." *Econometrica*, 82(2): 811–822.

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

Kamat, Vishal, Samuel Norris, and Matthew Pecenco. 2023. "Identification in Multiple Treatment Models under Discrete Variation." *arXiv preprint arXiv:2307.06174*.

Kline, Brendan, and Elie Tamer. 2023. "Recent developments in partial identification." *Annual Review of Economics*, 15: 125–150.

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.

Lochner, Lance, and Enrico Moretti. 2015. "Estimating and testing models with many treatment levels and limited instruments." *Review of Economics and Statistics*, 97(2): 387–397.

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

Manski, Charles F., and John V. Pepper. 2000. "Monotone Instrumental Variables: With an Application to the Returns to Schooling." *Econometrica*, 68(4): 997–1010.

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

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters." *Econometrica*, 86(5): 1589–1619.

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.

Rambachan, Ashesh, and Jonathan Roth. 2023. "A more credible approach to parallel trends." *Review of Economic Studies*, 90(5): 2555–2591.

Rose, Evan K, and Yotam Shem-Tov. 2021. "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. 2024. "On recoding ordered treatments as binary indicators." National Bureau of Economic Research.

Swanson, Sonja A, Miguel A Hernán, Matthew Miller, James M Robins, and Thomas S Richardson. 2018. "Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes." *Journal of the American Statistical Association*, 113(522): 933–947.

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.

Tebaldi, Pietro, Alexander Torgovitsky, and Hanbin Yang. 2023. "Nonparametric estimates of demand in the california health insurance exchange." *Econometrica*, 91(1): 107–146.

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

Table of Contents

A Figures **A.2**

B Relationship between the ACR and CCE **A.2**

C Identification of the CCE Without Homogeneity **A.3**

C.1 Proof that the CCE is Unconstrained Under Assumptions 1-3 . . . . . A.3

C.2 Bounded Unit Causal Effects A.5

D Identification of CCE with Multiple Instruments **A.5**

E Proof for Proposition 1 **A.6**

F Proof for Proposition 2 **A.6**

G Sharpness of Identification Results **A.7**

G.1 Sharpness of Proposition 1 . . . . . A.9

G.2 Sharpness of Proposition 2 . . . . . A.11

G.3 Illustration A.12

H Identification of CCE with Additional Boundedness Assumptions **A.13**

H.1 Bounded Unit Causal Effects with Homogeneity A.13

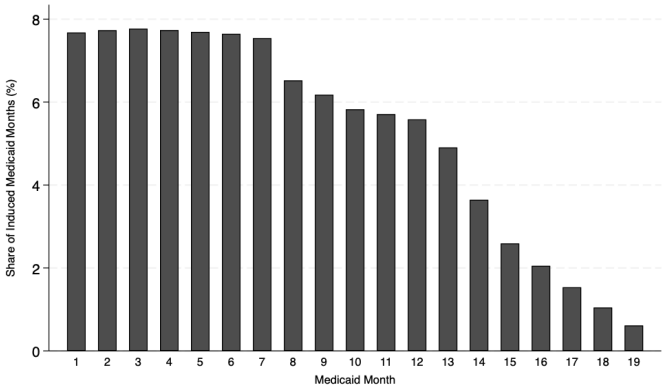
H.2 Bounded Outcomes with Homogeneity A.15

I Identification of Related Parameters **A.16**

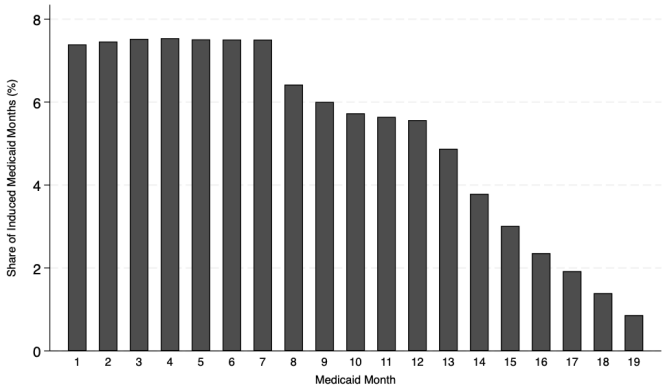
I.1 Proposition 1 . . . . . A.17

I.2 Proposition 2 . . . . . A.18

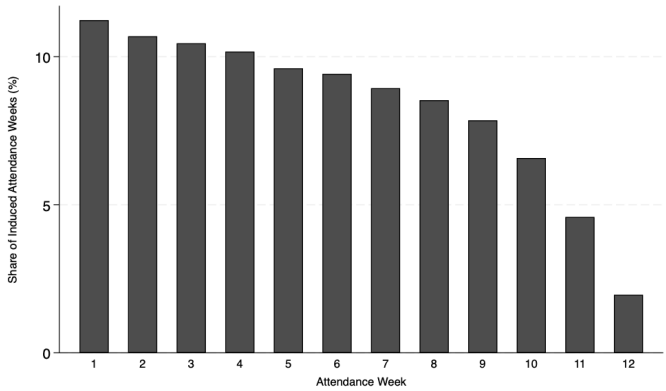
A Figures



(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

It will be convenient to introduce the following notation. Let β_{ij} denote the unit-causal effect for individual i at treatment intensity j: $\beta_{ij} = Y_i(j) - Y_i(j-1)$. Let δ_{ij} indicate whether the instrument induces unit i to cross treatment intensity j, $\delta_{ij} = \mathbb{1}\{D_i(1) \geq j > D_i(0)\}$.

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

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

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, $\gamma_i = \frac{1}{J}\sum_j \beta_{ij} - \overline{\gamma}$; γ_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, $\delta_i = \frac{1}{J}\sum\limits_{j=1}^{J} \delta_{ij}$; δ_j denote the share of compliers induced by the instrument to cross treatment intensity j, $\delta_j = \frac{1}{N_c}\sum\limits_{i=1}^{N_c} \delta_{ij}$; 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:

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

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

$$ACR - \overline{\gamma} = \frac{1}{\overline{\delta}}\left(\text{Cov}(\gamma_i, \delta_i) + \text{Cov}(\gamma_j, \delta_j) + \text{Cov}(\tilde{\gamma}_{ij}, \delta_{ij})\right) \tag{3}$$

To obtain Equation (3), we simply plug $\overline{\gamma} = CCE/J$ into the above.

C Identification of the CCE Without Homogeneity

This section explores the bounds on the CCE after relaxing Assumption 4, allowing for potential heterogeneity in the unit causal effects among compliers.

C.1 Proof that the CCE is Unconstrained Under Assumptions 1-3

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

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

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,

$$w_j = \sum_{i=1}^{N}\delta_{ij} \Big/ \sum_{j'=1}^{J}\sum_{i=1}^{N}\delta_{ij'}$$

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:

$$\delta\, ACR = \sum_{i=1}^{N}\sum_{j=1}^{J}\delta_{ij}\,\beta_{ij} \tag{4}$$

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

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

where the last equality uses (4).

To see that the CCE is unbounded under assumptions 1-3, note that we can always increase (or decrease) β_{ij} arbitrarily for any i and j pair with $\delta_{ij} = 0$ and $D_i(1) > D_i(0)$, since these unit causal effects are unconstrained by the ACR or other

observable moments. Therefore, as long as every single complier doesn't move along the full length of the dose response function, i.e., there exists some i for which $D_i(1) > D_i(0)$ and either $D_i(1) < J$ or $D_i(0) > 0$, the CCE is unbounded absent additional assumptions.[11]

C.2 Bounded Unit Causal Effects

One approach to restricting the free unit-causal effects is to impose an additional assumption on the support of the unit-causal effects, $\beta_{ij} \in \left[\underline{\beta}, \overline{\beta}\right]$. Suppose that $\beta_{ij} = \beta$ for all (i, j) pairs such that $\delta_{ij} = 0$. We can then express the CCE (scaled by J) as a convex combination of the ACR and β:

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

where the weight on the ACR, $\frac{\delta}{J\,N_c}$, corresponds to the fraction of the maximum potential effect of the instrument that is reflected in the ACR. We can therefore express the maximum CCE as

$$\begin{aligned}\overline{CCE} &= \left(\frac{\delta}{J\,N_c}\right) J\,ACR + \left(1 - \frac{\delta}{J\,N_c}\right) J\,\overline{\beta} \\ &= \frac{1}{N_c}\left[\delta\,ACR + (J\,N_c - \delta)\overline{\beta}\right]\end{aligned}$$

Analogously, the minimum CCE is

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

Note that the width of these bounds is increasing in N_c, which is not point-identified by the data. The maximum value of N_c, and hence the full range of CCE values consistent with the data, can be estimated based on the method described in Huang et al. (2016).

D Identification of CCE with Multiple Instruments

In this section, we consider identification of the CCE in the presence of additional identifying variation in the form of multiple instruments. Suppose there are $K \leq J$ mutually orthogonal binary instruments, $Z_k \in \{0, 1\}$. Assume that Assumption 1-3 are satisfied for each Z_k. Angrist and Imbens (1995) show that using K orthogonal indicators as instruments can equivalently be thought of as using a single $K + 1$ valued instrument, $Z_i \in \{0, 1, ..., K\}$. Let $\{w_j^k\}_{j=1}^J$ and ACR^k be the weighting scheme and ACR corresponding to instrument $d_k = \mathbb{1}(Z = k)$ respectively. Angrist and Imbens (1995) show that the K linearly independent dummy variables, $d_k = \mathbb{1}(Z_i = k)$, can be used as instruments to identify K linearly independent ACRs where

$$ACR^k = \frac{E[Y \mid Z = k] - E[Y \mid Z = 0]}{E[D \mid Z = k] - E[D \mid Z = 0]}$$

Recall that under Assumption 4 (homogeneity across complier types), we can re-write ACR^k as

$$ACR^k = \sum_{j=1}^{J} w_j^k \beta_j \tag{5}$$

for each $k = 1, ..., K$. When $K = J$, we have a system of J equations with J unknowns. Under an assumption that this system is full-rank, this allows for point-identification of each unit-causal effects, β_j. Therefore, when $K = J$, one can

[11] If all compliers do move along the full length of the dose response function, the setting can be analyzed as involving a binary treatment. In this case, the CCE is point-identified, and is equal to $ACR \times J$.

point-identify any function of the unit-causal effects including the CCE under Assumptions 1-4 and an additional full-rank assumption (requiring that the instruments induce compliers to move across distinct regions of the dose-response function).

Next, we consider identification in the case where $K < J$. In this case, we can use the additional identifying variation available through the additional instruments to tighten the bounds on the CCE. Specifically, we can incorporate (5) as a constraint in **LP.1** and **LP.2** for each $k = 1, .., K$:

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

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

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

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

The extent to which incorporating additional instruments into the analysis will tighten the bounds on the CCE will depend on the degree to which the additional instruments induce compliers to move across different portions of the treatment dose response function. For example, if $J = 2$ and a first instrument primarily moves compliers from $j = 0$ to $j = 1$, incorporating a second instrument will tend to tighten the estimate of the CCE if it primarily moves compliers from $j = 1$ to $j = 2$. In contrast, it will tend to provide less new information if it also primarily moves compliers from $j = 0$ to $j = 1$.

E 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. ■

F 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 CCE is higher than under 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. ■

G Sharpness of Identification Results

In this section, we prove that the bounds on the CCE in Proposition 1 and 2 are sharp, in the sense that all values of the CCE in the identified sets in Propositions 1 and 2 are compatible with the assumptions and the data (e.g., Kline and Tamer, 2023).

The data we observe is the joint distribution of (Y_i, D_i, Z_i). We will show that for identification of the CCE, the relevant moments of the joint distribution are $E[Y \mid D = j, Z = z]$ and $P[D = j \mid Z = z]$ for $j \in \{0, 1..., J\}$ and $z \in \{0, 1\}$. Under the assumptions of instrument validity (Assumptions 1-3), we establish the relationship between the unobserved joint distribution of $D_i(0)$ and $D_i(1)$ and the observed moments. Similarly, we show how the observed moments constrain the relationship between the outcome and treatment dosage under Assumptions 1-3. Next, we posit a dose-response relationship between the treatment and the outcome that attains the bounds in Propositions 1 and 2. We then show that this relationship is consistent with the assumptions and the observed moments, thereby establishing sharpness of our proposed bounds. We illustrate our argument in a simple example with three treatment levels ($J = 2$).

We begin by enumerating the relevant observed moments of the joint distribution of (Y_i, D_i, Z_i). Let $P[D_i(0) = m, D_i(1) = n] \equiv \pi_{mn}$ and $E[Y_i(j)|D_i(0) = m, D_i(1) = n] \equiv Y_{mn}(j)$ for $j, m, n \in \{0, 1, ..., J\}$. Using this notation, monotonicity (Assumption 3) implies $\pi_{mn} = 0$ for all m, n with $m > n$. Imposing independence (Assumption 2), we obtain the following constraints on the joint distribution of $D_i(0)$ and $D_i(1)$:

$$P[D = k \mid Z = 0] = P[D(0) = k \mid Z = 0] = P[D(0) = k] = \sum_{j=k}^{J} \pi_{kj} \tag{6}$$

$$P[D = k \mid Z = 1] = P[D(1) = k \mid Z = 1] = P[D(1) = k] = \sum_{j=0}^{k} \pi_{jk} \tag{7}$$

for each $k \in \{0, 1, ..., J\}$.

Relevance of the instrument (Assumption 1) implies that $E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0] = \sum_{j=1}^{J} P[D_i(1) \geq j >$

$D_i(0)] = \sum_{j=1}^{J} \sum_{p=j}^{J} \sum_{q=0}^{j-1} \pi_{qp} > 0$.

Turning to the relationship between the treatment dosage and the outcome, $Y_{mn}(m)$ and $Y_{mn}(n)$ are similarly constrained by the observed moments under independence and monotonicity (Assumptions 2-3):

$$E[Y \mid Z = 0, D = k]P[D = k \mid Z = 0] = \sum_{j=k}^{J} \pi_{kj} Y_{kj}(k) \tag{8}$$

$$E[Y \mid Z = 1, D = k]P[D = k \mid Z = 1] = \sum_{j=0}^{k} \pi_{jk} Y_{jk}(k) \tag{9}$$

for each $k \in \{0, 1, ..., J\}$.

Next, we show that while we observe the entire joint distribution of (Y_i, D_i, Z_i), the only features of the distribution that are relevant for identification of the CCE are $E[Y|D = j, Z = z]$ and $P(D = j \mid Z = z)$ for $j \in \{0, 1, ..., J\}$ and $z \in \{0, 1\}$. We show this by expressing the CCE under homogeneity (Assumption 4) as a function of the features of the joint distribution of (Y_i, D_i, Z_i) that are constrained by (6)-(9). Specifically, we show that the CCE can be expressed as follows:

$$\begin{aligned} CCE = \frac{1}{P[D = J \mid Z = 1] - P[D = J \mid Z = 0]} \times \\ \Big[E[Y \mid Z = 1, D = J]P[D = J \mid Z = 1] - E[Y \mid Z = 0, D = J]P[D = J \mid Z = 0] \\ + E[Y \mid Z = 1, D = 0]P[D = 0 \mid Z = 1] - E[Y \mid Z = 0, D = 0]P[D = 0 \mid Z = 0] \\ + \sum_{j=1}^{J-1} \Big(\pi_{0j} Y_{0j}(0) - \pi_{jJ} Y_{jJ}(0) \Big) \Big] \end{aligned}$$

To see this, we start with the observed moment conditions (8) and (9) for $D = 0$ and $D = J$ respectively:

$$\begin{aligned} & E[Y \mid Z = 1, D = J]P[D = J \mid Z = 1] - E[Y \mid Z = 0, D = 0]P[D = 0 \mid Z = 0] \\ &= \sum_{j=0}^{J} \pi_{jJ} Y_{jJ}(J) - \sum_{j=0}^{J} \pi_{0j} Y_{0j}(0) \\ &= \pi_{0J} \Big(Y_{0J}(J) - Y_{0J}(0) \Big) + \sum_{j=1}^{J} \pi_{jJ} Y_{jJ}(J) - \sum_{j=0}^{J-1} \pi_{0j} Y_{0j}(0) \\ &= \pi_{0J} CCE + \pi_{JJ} Y_{JJ}(J) - \pi_{00} Y_{00}(0) + \sum_{j=1}^{J-1} \Big(\pi_{jJ} Y_{jJ}(J) - \pi_{0j} Y_{0j}(0) \Big) \\ &= \pi_{0J} CCE + \pi_{JJ} Y_{JJ}(J) - \pi_{00} Y_{00}(0) + \sum_{j=1}^{J-1} \Big(\pi_{jJ} CCE + \pi_{jJ} Y_{jJ}(0) - \pi_{0j} Y_{0j}(0) \Big) \\ &= \Big(\sum_{j=0}^{J} \pi_{jJ} - \pi_{JJ} \Big) CCE + \pi_{JJ} Y_{JJ}(J) - \pi_{00} Y_{00}(0) + \sum_{j=1}^{J-1} \Big(\pi_{jJ} Y_{jJ}(0) - \pi_{0j} Y_{0j}(0) \Big) \end{aligned}$$

Plugging in moment conditions (6) and (7) for $D = J$, and moment conditions (8) and (9) for $D = J$ and $D = 0$

respectively, allows us to rewrite the above as

$$\begin{aligned}
&E[Y \mid Z=1, D=J]P[D=J \mid Z=1] - E[Y \mid Z=0, D=0]P[D=0 \mid Z=0] \\
&= \Big(P[D=J \mid Z=1] - P[D=J \mid Z=0]\Big)CCE \\
&\quad + E[Y \mid Z=0, D=J]P[D=J \mid Z=0] - E[Y \mid Z=1, D=0]P[D=0 \mid Z=1] \\
&\quad + \sum_{j=1}^{J-1} \Big(\pi_{jJ} Y_{jJ}(0) - \pi_{0j} Y_{0j}(0)\Big)
\end{aligned}$$

Rearranging terms, we can re-write the CCE as follows,

$$\begin{aligned}
CCE = &\frac{1}{P[D=J \mid Z=1] - P[D=J \mid Z=0]} \times \\
&\Bigg[E[Y \mid Z=1, D=J]P[D=J \mid Z=1] - E[Y \mid Z=0, D=J]P[D=J \mid Z=0] \\
&+ E[Y \mid Z=1, D=0]P[D=0 \mid Z=1] - E[Y \mid Z=0, D=0]P[D=0 \mid Z=0] \\
&+ \sum_{j=1}^{J-1} \Big(\pi_{0j} Y_{0j}(0) - \pi_{jJ} Y_{jJ}(0)\Big)\Bigg]
\end{aligned}$$

This shows that the CCE is only a function of the features of the joint distribution of (Y_i, D_i, Z_i) that are constrained by (6)-(9).

G.1 Sharpness of Proposition 1

Recall that Proposition 1 shows that, under Assumptions 1-5, $CCE \in \left[\frac{ACR}{w_{\bar{j}}}, \frac{ACR}{w_{\underline{j}}}\right]$ where $\underline{j} = \arg\min_{j \in \{1,...J\}}\{w_j\}$ and $\bar{j} = \arg\max_{j \in \{1,...J\}}\{w_j\}$. To simplify exposition, we additionally assume that the weights are monotonically declining, so $\bar{j} = 1$ and $\underline{j} = J$; the same logic extends to the case where this condition does not hold .[12]

Consider the following candidate relationship between the outcome and treatment dosage for all complier types: $Y_{mn}(0) = Y_{mn}(1) = ... = Y_{mn}(J-1)$ and $Y_{mn}(J) - Y_{mn}(J-1) = ACR/w_J$ for all $n > m$ and $m, n \in \{0, 1, ..., J\}$.

We begin by showing that this candidate dose-response relationship is consistent with Assumptions 1-5. Since the choice of the dose-response relationship does not affect the validity of the instrument, it is consistent with Assumptions 1-3. Assumption 4 imposes that the unit causal effects, $Y_{mn}(j) - Y_{mn}(j-1)$ are constant across all $n > m$ and $m, n \in \{0, 1, ..., J\}$. Since we propose the same candidate dose-response relationship for all complier types, Assumption 4 is satisfied by construction. It is also easy to see that the unit causal effects for all compliers, $\beta_j = Y_{mn}(j) - Y_{mn}(j-1)$ are non-negative, satisfying Assumption 5.

Next, we show that this candidate dose-response relationship attains the maximum value of the CCE in Proposition 1. Under Assumption 4, the CCE can be written as $Y_{mn}(J) - Y_{mn}(0)$ for any $n > m$ and $m, n \in \{0, 1, ..., J\}$ (i.e., for any complier). Adding and subtracting $Y_{mn}(J-1)$, we have $CCE = Y_{mn}(J) - Y_{mn}(J-1) + Y_{mn}(J-1) - Y_{mn}(0) = ACR/w_J$, where the final equality follows from the fact that $Y_{mn}(J-1) = Y_{mn}(0)$ under the candidate dose response function. This shows that the candidate dose-response relationship attains the maximum value in Proposition 1.

To show that the upper bound in Proposition 1 is sharp, it remains to show that the candidate dose-response relationship is consistent with the observable moments. We prove this by showing that there exists a vector of unobserved potential outcomes that satisfies Equations (6)-(9) under the candidate dose-response relationship. We begin by plugging in the

[12] If set of weights $\{w_j\}$ are not monotonically declining, the dose-response relationship that would attain the bounds in Proposition 1 would maximize the unit effect at the range of the dose response function where the weight is smallest. The remainder of the proof would follow similarly.

dose-response relationship $Y_{mn}(0) = Y_{mn}(k)$ for $k \in \{1, ..., J-1\}$ in (8) and (9):

$$E[Y \mid Z=0, D=k]P[D=k \mid Z=0] = \pi_{kk}Y_{kk}(k) + \sum_{j=k+1}^{J} \pi_{kj}Y_{kj}(0)$$

$$E[Y \mid Z=1, D=k]P[D=k \mid Z=1] = \pi_{kk}Y_{kk}(k) + \sum_{j=0}^{k-1} \pi_{jk}Y_{jk}(0)$$

For $k = 0$ in (9),

$$E[Y \mid Z=1, D=0]P[D=0 \mid Z=1] = \pi_{00}Y_{00}(0)$$

Plugging the above into (8) for $k = 0$,

$$E[Y \mid Z=0, D=0]P[D=0 \mid Z=0] - E[Y \mid Z=1, D=0]P[D=0 \mid Z=1] = \sum_{j=1}^{J} \pi_{0j}Y_{0j}(0)$$

For $k = J$ in (8),

$$E[Y \mid Z=0, D=J]P[D=J \mid Z=0] = \pi_{JJ}Y_{JJ}(J)$$

Plugging the above into (9) for $k = J$,

$$E[Y \mid Z=1, D=J]P[D=J \mid Z=1] - E[Y \mid Z=0, D=J]P[D=J \mid Z=0] = \sum_{j=0}^{J-1} \pi_{jJ}Y_{jJ}(J)$$

Next, plugging in the dose-response relationship, $Y_{jJ}(J) = ACR/w_J + Y_{jJ}(0)$, into the above equation yields the following:

$$\begin{aligned} E[Y \mid Z=1, D=J]P[D=J \mid Z=1] - E[Y \mid Z=0, D=J]P[D=J \mid Z=0] \\ = \frac{ACR}{w_J}\left(\sum_{j=0}^{J-1} \pi_{jJ}\right) + \sum_{j=0}^{J-1} \pi_{jJ}Y_{jJ}(0) \end{aligned}$$

Together, this yields the following system of $2J$ equations:

$$E[Y \mid Z=0, D=k]P[D=k \mid Z=0] = \pi_{kk}Y_{kk}(k) + \sum_{j=k+1}^{J} \pi_{kj}Y_{kj}(0) \quad \text{for} \quad k = 1, ..., J-1$$

$$E[Y \mid Z=1, D=k]P[D=k \mid Z=1] = \pi_{kk}Y_{kk}(k) + \sum_{j=0}^{k-1} \pi_{jk}Y_{jk}(0) \quad \text{for} \quad k = 1, ..., J-1$$

$$E[Y \mid Z=0, D=0]P[D=0 \mid Z=0] - E[Y \mid Z=1, D=0]P[D=0 \mid Z=1] = \sum_{j=1}^{J} \pi_{0j}Y_{0j}(0)$$

$$\begin{aligned} E[Y \mid Z=1, D=J]P[D=J \mid Z=1] - E[Y \mid Z=0, D=J]P[D=J \mid Z=0] \\ = \frac{ACR}{w_J}\left(\sum_{j=0}^{J-1} \pi_{jJ}\right) + \sum_{j=0}^{J-1} \pi_{jJ}Y_{jJ}(0) \end{aligned}$$

In this system of equations, $Y_{mn}(0)$ is unknown for all $n > m$ and $m, n \in \{0, 1, ..., J\}$. Additionally, $Y_{kk}(k)$ is unknown for all $k \in \{1, ..., J-1\}$. Together, this implies that there are $\frac{J(J+1)}{2} + J - 1$ unknowns. For all $J \geq 2$, this implies that there are at least as many unknown potential outcomes as there are equations, and so there is at least 1 vector of unknown potential outcomes that satisfies this set of equations.

Together, this shows that the upper bound for the CCE in Proposition 1 can be attained under a dose-response relationship between the treatment and the outcome that is consistent with the assumptions and the data. This establishes that the bounds in Proposition 1 are sharp. The proof for the lower bound is symmetric.

G.2 Sharpness of Proposition 2

Recall that Proposition 2 shows that, under Assumptions 1-7, the CCE is bounded above by $ACR \times J$ and bounded below by ACR/w_1. Below, we provide a proof for the sharpness of the upper bound. The proof that the lower bound is sharp is symmetric to the proof in Proposition 1.

Consider the following relationship between the outcome and treatment dosage for all complier types: $Y_{mn}(j) - Y_{mn}(j-1) = ACR$ for all $j \in \{1, ..., J\}$ and all $n > m$.

We begin by showing that this candidate dose-response relationship is consistent with Assumptions 1-6. Since the choice of the dose-response relationship does not affect the validity of the instrument, it is consistent with Assumptions 1-3. Assumption 4 imposes that the unit causal effects, $Y_{mn}(j) - Y_{mn}(j-1)$ are constant across all $n > m$ and $m, n \in \{0, 1, ..., J\}$. Since we propose the same candidate dose-response relationship for all complier types, Assumption 4 is satisfied by construction. It is also easy to see that the unit causal effects for all compliers, $Y_{mn}(j) - Y_{mn}(j-1) = ACR$ are non-negative, satisfying Assumption 5. Assumption 6 imposes that the dose-response relationship between the treatment and outcome is (weakly) concave, i.e., $Y_{mn}(j) - Y_{mn}(j-1) \geq Y_{mn}(j+1) - Y_{mn}(j)$ for all $j \in \{1, 2, ..., J-1\}$. The candidate dose-response relationship we specify is such that $Y_{mn}(j) - Y_{mn}(j-1) = Y_{mn}(j+1) - Y_{mn}(j)$ for all $j \in \{1, 2, ..., J-1\}$, and is therefore compatible with Assumption 6.

Next, we show that the candidate dose-response relationship attains the maximum value of the CCE in Proposition 2. Under Assumption 4, the CCE can be written as $\sum_{j=1}^{J} Y_{mn}(j) - Y_{mn}(j-1)$ for any $n > m$ and $m, n \in \{0, 1, ..., J\}$. Plugging in $Y_{mn}(j) - Y_{mn}(j-1) = ACR$, we obtain that $CCE = ACR \times J$. This shows that our candidate dose-response relationship attains the maximum value in Proposition 2.

To show that the upper bound in Proposition 2 is sharp, it remains to show that the candidate dose-response relationship is consistent with the observable moments. We prove this by showing that there exists a vector of unobserved potential outcomes that satisfies Equations (6)-(9) under the specified dose-response relationship.

We begin by plugging in the dose-response relationship $Y_{mn}(k) = Y_{mn}(0) + ACR \times k$ for $k \in \{1, ..., J-1\}$ in (8) and (9):

$$E[Y \mid Z = 0, D = k]P[D = k \mid Z = 0] = \pi_{kk} Y_{kk}(k) + \sum_{j=k+1}^{J} \pi_{kj} \Big(Y_{kj}(0) + ACR \times k \Big)$$

$$E[Y \mid Z = 1, D = k]P[D = k \mid Z = 1] = \pi_{kk} Y_{kk}(k) + \sum_{j=0}^{k-1} \pi_{jk} \Big(Y_{jk}(0) + ACR \times k \Big)$$

For $k = 0$ in (9),

$$E[Y \mid Z = 1, D = 0]P[D = 0 \mid Z = 1] = \pi_{00} Y_{00}(0)$$

Plugging in the above into (8) for $k = 0$,

$$E[Y \mid Z = 0, D = 0]P[D = 0 \mid Z = 0] - E[Y \mid Z = 1, D = 0]P[D = 0 \mid Z = 1] = \sum_{j=1}^{J} \pi_{0j} Y_{0j}(0)$$

For $k = J$ in (8),

$$E[Y \mid Z = 0, D = J]P[D = J \mid Z = 0] = \pi_{JJ} Y_{JJ}(J)$$

Plugging in the above into (9) for $k = J$,

$$E[Y \mid Z = 1, D = J]P[D = J \mid Z = 1] - E[Y \mid Z = 0, D = J]P[D = J \mid Z = 0] = \sum_{j=0}^{J-1} \pi_{jJ} Y_{jJ}(J)$$

Next, plugging in the dose-response relationship, $Y_{jJ}(J) = ACR \times J + Y_{jJ}(0)$, into the above equation yields the following:

$$E[Y \mid Z = 1, D = J]P[D = J \mid Z = 1] - E[Y \mid Z = 0, D = J]P[D = J \mid Z = 0] = \sum_{j=0}^{J-1} \pi_{jJ} \Big(Y_{jJ}(0) + ACR \times J \Big)$$

Together, this yields the following system of $2J$ equations:

$$E[Y \mid Z = 0, D = k]P[D = k \mid Z = 0] = \pi_{kk} Y_{kk}(k) + \sum_{j=k+1}^{J} \pi_{kj} \Big(Y_{kj}(0) + ACR \times k \Big) \quad \text{for} \quad k = 1, ..., J-1$$

$$E[Y \mid Z = 1, D = k]P[D = k \mid Z = 1] = \pi_{kk} Y_{kk}(k) + \sum_{j=0}^{k-1} \pi_{jk} \Big(Y_{jk}(0) + ACR \times k \Big) \quad \text{for} \quad k = 1, ..., J-1$$

$$E[Y \mid Z = 0, D = 0]P[D = 0 \mid Z = 0] - E[Y \mid Z = 1, D = 0]P[D = 0 \mid Z = 1] = \sum_{j=1}^{J} \pi_{0j} Y_{0j}(0)$$

$$E[Y \mid Z = 1, D = J]P[D = J \mid Z = 1] - E[Y \mid Z = 0, D = J]P[D = J \mid Z = 0] = \sum_{j=0}^{J-1} \pi_{jJ} \Big(Y_{jJ}(0) + ACR \times J \Big)$$

In this system of equations, $Y_{mn}(0)$ is unknown for all $n > m$ and $m, n \in \{0, 1, ..., J\}$. Additionally, $Y_{kk}(k)$ is unknown for all $k \in \{1, ..., J-1\}$. Together, this implies that there are $\frac{J(J+1)}{2} + J - 1$ unknowns. For all $J \geq 2$, this implies that there are at least as many unknown potential outcomes as there are equations, and so there is at least 1 vector of unknown potential outcomes that satisfies this set of equations.

Together, this shows that the upper bound for the CCE in Proposition 2 can be attained under a dose-response relationship between the treatment and the outcome that is consistent with the assumptions and the data. The proof for the lower bound follows the same logic as Section G.1.[13]

G.3 Illustration

In this section, we illustrate the sharpness of the Proposition 1 bounds in the simple case where $J = 2$.

Assume the same candidate dose-response relationship as considered in Section G.1, $Y_{mn}(0) = Y_{mn}(1)$ and $Y_{mn}(2) = ACR/w_2$ for all $n > m$. We showed that this dose-response relationship attains the upper bound of the CCE provided in Proposition 1 and is consistent with Assumptions 1-5. Here, we provide a vector of the unknown potential outcomes that is consistent with (6)-(9) to illustrate the sharpness of the provided bounds. Re-writing Equations (8) and (9) using the

[13] The dose-response relationship under which the lower bound is attained is given by $Y_{mn}(1) - Y_{mn}(0) = ACR/w_1$ and $Y_{mn}(1) = ... = Y_{mn}(J)$ for all $n > m$. In addition to the proof in Section G.1, one only needs to show that this dose-response relationship is also consistent with Assumption 6. This is easy to see, since $\beta_1 = ACR/w_1 > 0 = \beta_j$ for all $j > 1$.

specified dose-response relationship, as in Section G.1:

$$E[Y \mid Z=0, D=1]P[D=1 \mid Z=0] = \pi_{11}Y_{11}(1) + \pi_{12}Y_{12}(0)$$
$$E[Y \mid Z=1, D=1]P[D=1 \mid Z=1] = \pi_{11}Y_{11}(1) + \pi_{01}Y_{01}(0)$$
$$E[Y \mid Z=1, D=2]P[D=2 \mid Z=1] - E[Y \mid Z=0, D=2]P[D=2 \mid Z=0] = \frac{ACR}{w_J}\Big(\pi_{02}+\pi_{12}\Big) + \pi_{02}Y_{02}(0) + \pi_{12}Y_{12}(0)$$
$$E[Y \mid Z=0, D=0]P[D=0 \mid Z=0] - E[Y \mid Z=1, D=0]P[D=0 \mid Z=1] = \pi_{01}Y_{01}(0) + \pi_{02}Y_{02}(0)$$

The vector of unknowns here is $\pi_{01}Y_{01}(0)$, $\pi_{02}Y_{02}(0)$, $\pi_{11}Y_{11}(1)$, and $\pi_{12}Y_{12}(0)$. Assume the following values of the unknown potential outcomes:

$$\pi_{01}Y_{01}(0) = E[Y \mid Z=1, D=1]P[D=1 \mid Z=1] - E[Y \mid Z=0, D=1]P[D=1 \mid Z=0]$$
$$\pi_{02}Y_{02}(0) = E[Y \mid Z=1, D=2]P[D=2 \mid Z=1] - E[Y \mid Z=0, D=2]P[D=2 \mid Z=0] - \frac{ACR}{w_J}\Big(\pi_{02}+\pi_{12}\Big)$$
$$\pi_{11}Y_{11}(1) = E[Y \mid Z=0, D=1]P[D=1 \mid Z=0]$$
$$\pi_{12}Y_{12}(0) = 0$$

It is easy to see simply by plugging in the assumed values of the unknown potential outcomes that this vector of unknowns satisfies the first three of the four moment conditions shown above. It remains to show that

$$E[Y \mid Z=0, D=0]P[D=0 \mid Z=0] - E[Y \mid Z=1, D=0]P[D=0 \mid Z=1] = \pi_{01}Y_{01}(0) + \pi_{02}Y_{02}(0)$$
$$\Leftrightarrow \frac{ACR}{w_J} = \frac{1}{\pi_{02}+\pi_{12}} \times \left(\sum_{j=0}^{2} E[Y \mid Z=1, D=j]P[D=j \mid Z=1] - E[Y \mid Z=0, D=j]P[D=j \mid Z=0] \right)$$

Using the law of total probability, the numerator on the RHS may be written as $E[Y \mid Z=1] - E[Y \mid Z=0]$. Next, it can be seen that $\pi_{02}+\pi_{12} = P[D_i(1) \geq 2 > D_i(0)]$. Dividing the numerator and denominator by the first stage, we obtain that the RHS is equal to ACR/w_J. This shows that the assumed values of the unknown potential outcomes are consistent with the observed moments and assumptions, and they attain the upper bound on the CCE provided in Proposition 1.

H Identification of CCE with Additional Boundedness Assumptions

This section considers bounds for the CCE under Assumptions 1-5, when either the unit causal effects or the support of the outcome are known to be bounded.

H.1 Bounded Unit Causal Effects with Homogeneity

We first consider the additional assumption that the unit causal effects are bounded from above, i.e., $\beta_j \in [0, \overline{\beta}]$. We are interested in solving the following linear program:

$$\underset{\{\beta_j\}}{\text{Maximize/Minimize}} \quad \sum_{j=1}^{J} \beta_j \qquad \textbf{LP.4}$$

$$\text{subject to} \quad \sum_{j=1}^{J} \beta_j w_j = ACR \qquad \text{(C.1)}$$
$$\beta_j \geq 0 \qquad \text{(C.2)}$$
$$\beta_j \leq \overline{\beta} \qquad \text{(C.3)}$$

For ease of notation, we assume that the weights are monotonically declining (Assumption 7).

Beginning with the upper bound, define $j_u = \max\{1, ..., J\}$ such that $\sum_{j=j_u}^{J} w_j \geq \frac{ACR}{\overline{\beta}}$. Note that j_u is well-defined under our assumptions; if not, it would imply $ACR > \overline{\beta} \sum_{j=1}^{J} w_j = \overline{\beta}$. Because the ACR is a weighted average of unit causal effects, all of which are bounded above by $\overline{\beta}$, this is impossible. Therefore, j_u must be well-defined. Next, consider the case in which $j_u = J$, which is obtained when $w_J \geq \frac{ACR}{\overline{\beta}}$, or equivalently, $\overline{\beta} \geq \frac{ACR}{w_J}$. In this case, the vector of unit-causal effects that attain the bounds in the absence of C.3 (as shown in Proposition 1) are still feasible, since $\overline{\beta} \geq \frac{ACR}{w_J} = \beta_J$ and $\overline{\beta} > \frac{ACR}{w_1} = \beta_1$. Therefore, in the case that $j_u = J$, C.3 does not bind and **LP.4** collapses into the case studied in Proposition 1.

The other possibility is that $j_u \in \{1, ..., J-1\}$. To find an upper bound for the CCE in this case, we will first show that the CCE is maximized when $\beta_j = \overline{\beta}$ for all $j > j_u$ and $\beta_j = 0$ for all $j < j_u$.

To prove $\beta_j = \overline{\beta}$ for all $j > j_u$, assume towards a contradiction that the CCE is maximized subject to C.1-C.3 by a vector of unit causal effects $\beta = (\beta_1, ..., \beta_J)$ where $\beta_{k_1} < \overline{\beta}$ for some $k_1 > j_u$. Observe that, by construction of j_u, we have:

$$\begin{aligned} \overline{\beta} &< \frac{ACR}{\sum_{j=j_u+1}^{J} w_j} \\ \implies \sum_{j=j_u+1}^{J} \overline{\beta} w_j &< ACR = \sum_{j=1}^{J} \beta_j w_j \end{aligned}$$

where the last equality follows from C.1. Because $\beta_j \leq \overline{\beta}$ for all j, this implies that there must exist a $k \leq j_u$ such that $w_k > 0$ and $\beta_k > 0$. We assume without loss of generality that there is only one such non-zero unit-causal effect for $k \leq j_u$.[14] Consider a new vector of unit causal effects $\beta^* = (\beta_1^*, ..., \beta_J^*)$ equal to β, except that $\beta_k^* = \beta_k - (\overline{\beta} - \beta_{k_1}) \frac{w_{k_1}}{w_k}$ and $\beta_{k_1}^* = \overline{\beta}$. Note that β^* satisfies C.1-C.3. Comparing the CCE under β^* and β:

$$\sum_{j=1}^{J} \beta_j^* - \sum_{j=1}^{J} \beta_j = \left(\beta_k^* - \beta_k\right) + \left(\beta_{k_1}^* - \beta_{k_1}\right) = -\left(\overline{\beta} - \beta_{k_1}\right) \frac{w_{k_1}}{w_k} + \left(\overline{\beta} - \beta_{k_1}\right) = \left(\overline{\beta} - \beta_{k_1}\right)\left(1 - \frac{w_{k_1}}{w_k}\right) > 0$$

where the final inequality follows since $\frac{w_{k_1}}{w_k} < 1$. Thus, the initial vector of unit treatment effects must not be the solution to the constrained maximization problem, proving that the CCE is maximized only when $\beta_j = \overline{\beta}$ for all $j > j_u$.

To prove $\beta_j = 0$ for all $j < j_u$, assume towards a contradiction that the CCE is maximized by a vector of unit causal effects $\beta = (\beta_1, ..., \beta_J)$ where $\beta_{k_2} > 0$ for some $k_2 < j_u$. Because β maximizes the CCE, we know from above that $\beta_j = \overline{\beta}$ for all $j > j_u$. Consider a new vector of unit treatment effects, $\beta^* = (\beta_1^*, ..., \beta_J^*)$, identical to β except that the specified non-zero effect has been set to zero, $\beta_{k_2}^* = 0$, and β_{j_u} has been increased to maintain feasibility with respect to the ACR constraint: $\beta_{j_u}^* = \beta_{j_u} + \beta_{k_2} w_{k_2} / w_{j'}$. For it to be feasible to increase β_{j_u} in this way, we need that $\beta_{j_u}^* \leq \overline{\beta}$. This follows from the construction of j_u:

$$\frac{ACR}{\sum_{j=j_u}^{J} w_j} \leq \overline{\beta} \iff \frac{ACR - \sum_{j=j_u+1}^{J} w_j \overline{\beta}}{w_{j_u}} \leq \overline{\beta}$$

[14] This is without loss of generality because we can always re-define the problem with β_k being a weighted average of all non-zero unit causal effects. For instance, if $\beta_p, \beta_q > 0$ for $p, q \leq j_u$, we could re-define $\beta_k = \frac{\beta_p w_p + \beta_q w_q}{w_p + w_q}$ and $w_k = w_p + w_q$.

To see this, note that

$$\frac{ACR - \sum_{j=j_u+1}^{J} w_j\overline{\beta}}{w_{j_u}} = \frac{\sum_{j=1}^{j_u} w_j\beta_j}{w_{j_u}} = \frac{w_{j_u}\beta_{j_u} + w_{k_2}\beta_{k_2}}{w_{j_u}} = \beta_{j_u} + \frac{w_{k_2}\beta_{k_2}}{w_{j_u}} = \beta^*_{j_u} \leq \overline{\beta}$$

where the first equality follows from the ACR constraint, the second equality follows from our assumption that β_k is the only non-zero unit causal effect for $k < j_u$, the fourth equality follows from our definition of $\beta^*_{j_u}$. But with these changes, the value of the objective function is higher than that of the original vector of unit effects:

$$\sum \beta^*_j - \sum \beta_j = \beta_k \frac{w_k}{w_{j_u}} - \beta_k > 0,$$

where the last inequality follows from the fact that $\frac{w_k}{w_{j_u}} > 1$. Thus, the initial vector of unit treatment effects must not be the solution to the constrained maximization problem. Together, this implies that the CCE is maximized only when $\beta_j = \overline{\beta}$ for all $j > j_u$ and $\beta_j = 0$ for all $j < j_u$. The constraint C.1 pins down the value for β_{j_u} and implies that the upper bound on the CCE is

$$\frac{ACR}{w_{j_u}} - \overline{\beta}\ \frac{\sum_{j=j_u+1}^{J} w_j}{w_{j_u}} + (J - j_u)\overline{\beta}$$

Turning to the lower bound, define $j_l = \min\{1, ..., J\}$ such that $\sum_{j=1}^{j_l} w_j \geq \frac{ACR}{\overline{\beta}}$. Note that j_l is well-defined under our assumptions; if not, it would imply $ACR > \overline{\beta} \sum_{j=1}^{J} w_j = \overline{\beta}$. Because the ACR is a weighted average of unit causal effects, all of which are bounded above by $\overline{\beta}$, this is impossible. Therefore, j_l must be well-defined. Next, consider the case in which $j_l = 1$, which is obtained when $w_1 \geq \frac{ACR}{\overline{\beta}}$, or equivalently, $\overline{\beta} \geq \frac{ACR}{w_1}$. In this case, the vector of unit-causal effects that attain the lower bounds in the absence of C.3 (as shown in Proposition 1) is still feasible, since $\overline{\beta} \geq \frac{ACR}{w_1} = \beta_1$. Therefore, in the case that $j_l = 1$, C.3 does not bind and **LP.4** collapses into the case studied in Proposition 1. The other possibility is that $j_l \in \{2, ..., J\}$. The CCE is minimized when $\beta_j = \overline{\beta}$ for all $j < j_l$ and $\beta_j = 0$ for all $j > j_l$. The constraint C.1 pins down the value for β_{j_l} and implies that the lower bound on the CCE is

$$\frac{ACR}{w_{j_l}} - \overline{\beta}\ \frac{\sum_{j=1}^{j_l-1} w_j}{w_{j_l}} + (j_l - 1)\overline{\beta}$$

The derivation for the lower bound is analogous to the proof for the upper bound.

H.2 Bounded Outcomes with Homogeneity

In this subsection, we consider bounds on the CCE under Assumptions 1-5 in the case that the outcome has bounded support between known values, $Y_i \in \left[\underline{Y}, \overline{Y}\right]$.

Let $P[D_i(0) = m, D_i(1) = n] \equiv \pi_{mn}$ and $E[Y_i(j)|D_i(0) = m, D_i(1) = n] \equiv Y_{mn}(j)$ for $j, m, n \in \{0, 1, ..., J\}$. In Appendix Section G, we establish how the observed moments constrain the relationship between the outcome and treatment dosage as well as the relationship between the unobserved joint distribution of $D_i(0)$ and $D_i(1)$ and the observed moments under Assumptions 1-3. Here, we incorporate the moment conditions as constraints in the following optimization problem.

For any $m, n \in \{0, 1, ..., J\}$ and $n > m$:

$$\underset{\{Y_{jk}(d), \pi_{jk}\}_{j,k,d=0}^{J}}{\text{Maximize/Minimize}} \quad Y_{mn}(J) - Y_{mn}(0) \qquad \textbf{LP.5}$$

$$\text{subject to} \quad Y_{mn}(j) - Y_{mn}(j-1) \geq 0 \qquad \text{(C.1)}$$

$$\sum_{j=k}^{J} \pi_{kj} = P[D = k \mid Z = 0] \text{ for all } k \in \{0, 1, ..., J\} \qquad \text{(C.2)}$$

$$\sum_{j=0}^{k} \pi_{jk} = P[D = k \mid Z = 1] \text{ for all } k \in \{0, 1, ..., J\} \qquad \text{(C.3)}$$

$$\sum_{j=k}^{J} \pi_{kj} Y_{kj}(k) = E[Y \mid Z = 0, D = k] P[D = k \mid Z = 0] \text{ for all } k \in \{0, 1, ..., J\} \qquad \text{(C.4)}$$

$$\sum_{j=0}^{k} \pi_{jk} Y_{jk}(k) = E[Y \mid Z = 1, D = k] P[D = k \mid Z = 1] \text{ for all } k \in \{0, 1, ..., J\} \qquad \text{(C.5)}$$

$$Y_{jk}(d) \in \left[\underline{Y}, \overline{Y}\right] \text{ for all } j, k, d \in \{0, 1, ..., J\} \qquad \text{(C.6)}$$

Absent C.6, **LP.5** simplifies to the case studied in Proposition 1. As shown in Appendix G, the bounds in Proposition 1 are sharp under C.1 through C.5 and the data available to the researcher. Adding C.6 allows us to tighten the bounds described in that Proposition. In particular, one implication of C.6 is that the magnitude of each unit causal effect does not exceed $\overline{Y} - \underline{Y}$. Hence, the bounds described in Appendix H.1 are valid for **LP.5**, with $\overline{\beta} = \overline{Y} - \underline{Y}$. However, this restriction does not necessarily exploit all of the information contained in C.6; C.6 also provides information concerning the level of the potential outcomes, which is not reflected in constraints relating to the magnitude of the treatment effect. As such, directly solving **LP.5** may yield tighter bounds compared to the bounds described in Appendix H.1.

I Identification of Related Parameters

Thus far, we have focused on identification of the effect of moving from the treatment's minimum intensity to the treatment's maximum intensity. In this section, we explore identification of a broader class of parameters. Specifically, we consider identification of the effect among compliers of moving from treatment intensity j_1 to j_2 where $j_1, j_2 \in \{0, 1, ..J\}$ and $j_1 < j_2$. We refer to this as the Complier Effect (CE) for j_1 to j_2: $CE(j_1, j_2) = E[Y_i(j_2) - Y_i(j_1) | D_i(1) > D_i(0)]$. Under Assumption 4, the unit-causal effects are homogeneous across different types of compliance subgroups. Specifically, Assumption 4 requires that $E[Y_i(j_2) - Y_i(j_1) | D_i(1) > D_i(0)] = E[Y_i(j_2) - Y_i(j_1) | D_i(1) = j_2, D_i(0) = j_1]$. Therefore, we may re-write $CE(j_1, j_2)$ as $CE(j_1, j_2) = E[Y_i(j_2) - Y_i(j_1) | D_i(1) = j_2, D_i(0) = j_1] = \sum_{j=j_1+1}^{j_2} \beta_j$. In the special case where $j_1 = 0$ and $j_2 = J$,

$$CE(0, J) = CCE = E[Y_i(J) - Y_i(0) | D_i(1) = J, D_i(0) = 0] = \sum_{j=1}^{J} \beta_j$$

I.1 Proposition 1

We begin by asking what may be identified for $CE(j_1, j_2)$ under Assumptions 1-5. As in Section 4, we cast the problem of identifying $CE(j_1, j_2)$ as the following linear optimization problem:

$$\underset{\{\beta_j\}}{\text{Maximize/Minimize}} \quad \sum_{j=j_1+1}^{j_2} \beta_j \qquad \textbf{LP.6}$$

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

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

In this section, we extend Proposition 1 and show the following:[15]

Proposition 1*: Let $\underline{j}(j_1, j_2) = \arg\min_{j \in \{j_1+1,...,j_2\}}\{w_j\}$ and $\overline{j}(j_1, j_2) = \arg\max_{j \in \{j_1+1,...,j_2\}}\{w_j\}$. Under Assumptions 1-5, the following sharp bounds hold:

$$CE(0, J) = CCE \in \left[\frac{ACR}{w_{\overline{j}(0,J)}}, \frac{ACR}{w_{\underline{j}(0,J)}}\right]$$

and for $(j_1, j_2) \neq (0, J)$

$$CE(j_1, j_2) \in \left[0, \frac{ACR}{w_{\underline{j}(j_1,j_2)}}\right]$$

.

Proof: We begin by showing that $CE(j_1, j_2)$ is minimized at 0, for any $(j_1, j_2) \neq (0, J)$. Recall that, under Assumption 5, $\beta_j \geq 0$. This implies that $CE(j_1, j_2)$ is always bounded below at 0 for all j_1, j_2: $CE(j_1, j_2) = \sum_{j=j_1+1}^{j_2} \beta_j \geq 0$. Next, we show that for any $(j_1, j_2) \neq (0, J)$, we cannot rule out that $CE(j_1, j_2)$ takes on a value 0. Consider the case where $j_1 = 0$. This implies that $j_2 \leq J - 1$. Define k such that $j_2 < k \leq J$. Then, constraint (C.1) may be satisfied with $\beta_k = \frac{ACR}{w_k}$ and $\beta_j = 0$ for all $j \neq k$. This implies that $\sum_{j=j_1+1}^{j_2} \beta_j = 0$ and shows that $CE(j_1, j_2)$ is minimized at 0. Similarly, consider the case where $j_1 \neq 0$. Define k such that $0 \leq k < j_1$. Then, constraint (C.1) may be satisfied with $\beta_k = \frac{ACR}{w_k}$ and $\beta_j = 0$ for all $j \neq k$. This shows that the $CE(j_1, j_2)$ is minimized at 0, for any $(j_1, j_2) \neq (0, J)$. This shows that $CE(j_1, j_2)$ is minimized at 0 for any $(j_1, j_2) \neq (0, J)$.

Next, we show that $CE(j_1, j_2)$ is maximized at $\frac{ACR}{w_{\underline{j}(j_1,j_2)}}$ where $\underline{j}(j_1, j_2) = \arg\min_{j \in \{j_1+1,...,j_2\}}\{w_j\}$. To establish that $CE(j_1, j_2)$ is bounded above by $\frac{ACR}{w_{\underline{j}(j_1,j_2)}}$, we will first show that the $CE(j_1, j_2)$ is maximized if and only if $\beta_j = 0$ for all $j < j_1$ and $j > j_2$ and then show that $CE(j_1, j_2)$ is maximized when $\beta_{\underline{j}(j_1,j_2)}$ is the only non-zero unit causal effect, i.e., $\beta_j = 0$ for all $j \neq \underline{j}(j_1, j_2)$.

Proceeding by contradiction to show the first step, suppose instead that $CE(j_1, j_2)$ was maximized by a vector of unit treatment effects $(\beta_1, ..., \beta_J)$ with $\beta_k > 0$ for some $k < j_1$ or $k > j_2$. Consider a new vector of unit treatment effects, $(\beta_1^*, ..., \beta_J^*)$ identical to $(\beta_1, ..., \beta_J)$, except that the specified non-zero effect has been set to zero, $\beta_k^* = 0$, and the unit effect corresponding to the smallest weight, $\beta_{\underline{j}(j_1,j_2)}$, has been increased to maintain feasibility with respect to the ACR constraint: $\beta^*_{\underline{j}(j_1,j_2)} = \beta_{\underline{j}(j_1,j_2)} + \beta_k\, w_k / w_{\underline{j}(j_1,j_2)}$. But, with these changes, the value of the objective function is larger

[15] In this proposition, we assume $w_k > 0$ for at least one $k \leq j_1$ or $k > j_2$. If $w_k = 0$ for all $k \leq j_1$ and $k > j_2$, then the lower bound on $CE(j_1, j_2)$ would be the same as the lower bound on the CCE.

than that of the original vector of unit effects:

$$\sum_{j=j_1+1}^{j_2} \beta_j^* - \sum_{j=j_1+1}^{j_2} \beta_j = \beta_k \frac{w_k}{w_{\underline{j}(j_1,j_2)}} > 0$$

where the final inequality follows by the assumption that $\beta_k > 0$.

Proceeding similarly by contradiction to show the second step, suppose instead that $CE(j_1, j_2)$ was maximized by a vector of unit treatment effects $(\beta_1, ..., \beta_J)$ with $\beta_k > 0$ for some $k \neq \underline{j}(j_1, j_2)$ and $j_1 \leq k \leq j_2$. Consider a new vector of unit treatment effects, $(\beta_1^*, ..., \beta_J^*)$ identical to $(\beta_1, ..., \beta_J)$, except that the specified non-zero effect has been set to zero, $\beta_k^* = 0$, and the unit effect corresponding to the smallest weight, $\beta_{\underline{j}(j_1,j_2)}$, has been increased to maintain feasibility with respect to the ACR constraint: $\beta^*_{\underline{j}(j_1,j_2)} = \beta_{\underline{j}(j_1,j_2)} + \beta_k \, w_k / w_{\underline{j}(j_1,j_2)}$. But, with these changes, the value of the objective function is larger than that of the original vector of unit effects:

$$\sum_{j=j_1+1}^{j_2} \beta_j^* - \sum_{j=j_1+1}^{j_2} \beta_j = \beta_k \frac{w_k}{w_{\underline{j}(j_1,j_2)}} - \beta_k > 0,$$

where the last inequality follows from the fact that $\beta_k > 0$ and the definition of $\underline{j}(j_1, j_2)$. Thus, the initial vector of unit treatment effects must not be the solution to the constrained maximization problem, proving the contradiction. This shows that $CE(j_1, j_2)$ is maximized at $\frac{ACR}{w_{\underline{j}(j_1,j_2)}}$ for any $(j_1, j_2) \neq (0, J)$. Finally, we note that one can use the same argument as in Appendix Section G to establish that these bounds are sharp.

I.2 Proposition 2

We consider identification for $CE(j_1, j_2)$ under Assumptions 1-7. As in Section 4, we may cast the problem of identifying $CE(j_1, j_2)$ as the following linear optimization problem:

$$\underset{\{\beta_j\}}{\text{Maximize/Minimize}} \quad \sum_{j=j_1+1}^{j_2} \beta_j \qquad \textbf{LP.7}$$

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

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

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

In this section, we extend Proposition 2 to show the following.

Proposition 2*: Under Assumptions 1-7, the following sharp bounds hold: $CCE \in \left[\frac{ACR}{w_1}, \, ACR \times J\right]$.

(a) When $j_1, j_2 \in \{0, 1, ..., J\}$ where $j_2 > j_1$, $CE(j_1, j_2) \leq \frac{ACR}{\sum_{j=1}^{j_2} w_j} \times (j_2 - j_1)$

(b) When $j_1, j_2 \in \{1, ..., J\}$ where $j_2 > j_1$, $CE(j_1, j_2) \geq 0$

(c) When $j_1 = 0$ and $j_2 \in \{1, ..., J\}$ where $j_2 \geq \frac{1}{w_1}$, $CE(0, j_2) \geq \frac{ACR}{w_1}$

(d) When $j_1 = 0$ and $j_2 \in \{1, ..., J\}$ where $j_2 < \frac{1}{w_1}$, $CE(0, j_2) \geq ACR \times j_2$

Proof: We begin with the proof of Proposition 2*(a) and show that the $CE(j_1, j_2)$ is maximized at $\frac{ACR}{\sum_{j=1}^{j_2} w_j} \times (j_2 - j_1)$.

We proceed in three steps. First, we show that $CE(j_1, j_2)$ is maximized when $\beta_j = 0$ for all $j > j_2$. In the second step, we show that the $CE(j_1, j_2)$ is maximized when $\beta_{j_1+1} = ... = \beta_{j_2}$. Finally, we show that it is only maximized when $\beta_1 = ... = \beta_{j_1+1}$.

Proceeding by contradiction for the first step, suppose that $CE(j_1, j_2)$ is maximized subject to constraints (C.1)-(C.3) by a vector of unit effects $\beta = (\beta_1, \beta_2, ..., \beta_J)$ with $\beta_k > 0$ for some $k > j_2$. Consider a new vector of unit treatment effects, $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, ..., \tilde{\beta}_J)$, where $\tilde{\beta}_j = \beta_j + \frac{w_k \beta_k}{w_j j_2}$ for all $j \leq j_2$ and $\tilde{\beta}_k = 0$. To see that $\tilde{\beta}$ satisfies (C.1),

$$\sum_{j=1}^{J} \tilde{\beta}_j w_j = \sum_{j=1}^{j_2} \tilde{\beta}_j w_j = \sum_{j=1}^{j_2} \Big(\beta_j + \frac{w_k \beta_k}{w_j j_2}\Big) w_j = \sum_{j=1}^{j_2} \beta_j w_j + \sum_{j=1}^{j_2} \frac{w_k \beta_k}{j_2} = \sum_{j=1}^{j_2} \beta_j w_j + w_k \beta_k = ACR$$

where the first equality follows from the fact that $\tilde{\beta}_j = 0$ for all $j > j_2$ and the final equality follows by the assumption that the vector of unit effects β satisfies C.1. Additionally, $\tilde{\beta}$ satisfies C.2 since $\beta_j \geq 0$ for all j implies that $\tilde{\beta}_j = \beta_j + \frac{w_k \beta_k}{w_j j_2} \geq 0$. Finally, C.3 is satisfied since $\beta_j \geq \beta_{j+1}$ implies that $\beta_j + \sum_{j=1}^{j_2} \frac{w_k \beta_k}{w_j j_2} \geq \beta_{j+1} + \sum_{j=1}^{j_2} \frac{w_k \beta_k}{w_j j_2} \implies \tilde{\beta}_j \geq \tilde{\beta}_{j+1}$. Finally, comparing the objective functions under the two vectors of unit effects,

$$\sum_{j=j_1+1}^{j_2} \tilde{\beta}_j - \sum_{j=j_1+1}^{j_2} \beta_j = \sum_{j=j_1+1}^{j_2} \frac{w_k \beta_k}{w_j j_2} \geq 0$$

where the last inequality follows by the assumption that $\beta_k > 0$. This shows that $CE(j_1, j_2)$ is maximized when $\beta_j = 0$ for all $j > j_2$.

Second, we show that the $CE(j_1, j_2)$ is maximized when $\beta_{j_1+1} = ... = \beta_{j_2}$. Proceeding by contradiction, suppose that the $CE(j_1, j_2)$ is maximized by a vector of unit treatment effects $(\beta_1, \beta_2, ..., \beta_J)$ with $\beta_k > \beta_{k+1}$ for some $k \in \{j_1 + 1, ..., j_2 - 1\}$ (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:

$$\sum_{j=j_1+1}^{j_2} \tilde{\beta}_j - \sum_{j=j_1+1}^{j_2} \beta_j = \frac{(\beta_{k+1} - \beta_k)(w_{k+1} - w_k)}{w_{k+1} + w_k} > 0,$$

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. This shows that the $CE(j_1, j_2)$ is maximized when $\beta_{j_1+1} = ... = \beta_{j_2}$.

Finally, we show that $CE(j_1, j_2)$ is maximized when $\beta_1 = ... = \beta_{j_1+1}$. Proceeding by contradiction, suppose that $CE(j_1, j_2)$ is maximized by a vector of unit effects $\beta = (\beta_1, \beta_2, ...\beta_J)$ where $\beta_j = 0$ for all $j > j_2$ and $\beta_j = \beta_{j+1}$ for all $j \leq j_2$ except $j = k$ for some $k \in \{1, 2, ..., j_1\}$. This implies that $\beta_k > \beta_{k+1}$ (concavity rules out the reverse ordering). We can re-write (C.1) under the vector β as:

$$\sum_{j=1}^{J} \beta_j w_j = \sum_{j=1}^{j_2} \beta_j w_j = \sum_{j=1}^{k} \beta_j w_j + \sum_{j=k+1}^{j_2} \beta_j w_j = \sum_{j=1}^{k} \beta_k w_j + \sum_{j=k+1}^{j_2} \beta_{j_1+1} w_j = ACR$$

where the first equality follows the assumption that $\beta_j = 0$ for all $j > j_2$ and the third equality follows from the assumption that $\beta_j = \beta_k$ for all $j \leq k$ and $\beta_j = \beta_{j_1+1}$ for all $j \in \{k+1, ..., j_2\}$.

Define $\tilde{w}_k = \sum_{j=1}^{k} w_j$ and $\tilde{w}_{j_1+1} = \sum_{j=k+1}^{j_2} w_j$. Then, $\beta_k \tilde{w}_k + \beta_{j_1+1} \tilde{w}_{j_1+1} = ACR$.

Consider a new vector of unit treatment effects, $(\tilde{\beta}_1, ..., \tilde{\beta}_J)$, with $\tilde{\beta}_k = \beta_k - \frac{(\beta_k - \beta_{j_1+1}) \tilde{w}_{j_1+1}}{\tilde{w}_k + \tilde{w}_{j_1+1}}$ and $\tilde{\beta}_{j_1+1} = \beta_{j_1+1} + \frac{\tilde{w}_k (\beta_k - \beta_{j_1+1})}{\tilde{w}_k + \tilde{w}_{j_1+1}}$, so that $\tilde{\beta}_k = \tilde{\beta}_{j_1+1}$. Note that $(\tilde{\beta}_1, ..., \tilde{\beta}_J)$ is a feasible unit 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:

$$\sum_{j=j_1+1}^{j_2} \tilde{\beta}_j - \sum_{j=j_1+1}^{j_2} \beta_j = \frac{\tilde{w}_k(\beta_k - \beta_{j_1+1})}{\tilde{w}_k + \tilde{w}_{j_1+1}} > 0,$$

where the last inequality follows from the assumption that $\beta_k > \beta_{k+1} = \beta_{j_1+1}$. This shows that $CE(j_1, j_2)$ is maximized when $\beta_1 = ... = \beta_{j_2}$.

Finally, using the ACR constraint, this implies that

$$\sum_{j=1}^{J} \beta_j w_j = \sum_{j=1}^{j_2} \beta_j w_j = ACR \implies \beta_j = \frac{ACR}{\sum_{j=1}^{j_2} w_j}$$

Using the above, we can re-write $CE(j_1, j_2)$ as

$$CE(j_1, j_2) = \sum_{j=j_1+1}^{j_2} \beta_j = \sum_{j=j_1+1}^{j_2} \frac{ACR}{\sum_{j=1}^{j_2} w_j} = \frac{ACR(j_2 - j_1)}{\sum_{j=1}^{j_2} w_j}$$

Next, we prove Proposition 2*(b) and show that $CE(j_1, j_2)$ is minimized at 0 for any $j_1 > 0$. Recall that under Assumption 5, $\beta_j \geq 0$ for all j. This implies that $CE(j_1, j_2)$ is bounded below at 0 for any j_1, j_2, $CE(j_1, j_2) = \sum_{j=j_1+1}^{j_2} \beta_j \geq 0$. Next, we show that for any $j_1 > 0$, we cannot rule out that $CE(j_1, j_2)$ takes on a value of 0. The constraint (C.1) may be satisfied with $\beta_1 = \frac{ACR}{w_1}$ and $\beta_j = 0$ for all $j > 1$. Note that this vector of unit effects $(\beta_1, \beta_2, ..., \beta_J) = \left(\frac{ACR}{w_1}, 0, ..., 0\right)$ also satisfies Constraint (C.2) and (C.3) since $\beta_1 = \frac{ACR}{w_1} \geq \beta_j = 0$ for any $j > 1$. Under this vector of unit effects, $CE(j_1, j_2) = \sum_{j=j_1+1}^{j_2} \beta_j = 0$. This shows that $CE(j_1, j_2)$ is minimized at 0 for any $j_1 > 0$.

Next, we prove Proposition 2*(c) and show that the lower bound on $CE(0, j_2)$ is given by ACR/w_1 if $j_2 \geq \frac{1}{w_1}$. Suppose that $j_2 \geq \frac{1}{w_1}$. To show that $CE(0, j_2) \geq ACR/w_1$, we will first show that the $CE(0, j_2)$ is minimized when $\beta_j = 0 \; \forall \; j > 1$. Proceeding by contradiction, suppose that the $CE(0, j_2)$ 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:

$$\sum_{j=1}^{j_2} \tilde{\beta}_j - \sum_{j=1}^{j_2} \beta_j = \beta_k \frac{w_k}{w_1} - \beta_k < 0,$$

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 $CE(0, j_2) = \sum_j \beta_j = \beta_1 = ACR/w_1$, where the last equality follows from the definition of the ACR. Next, we show that $CE(0, j_2) = ACR/w_1$ would not be a lower bound if $j_2 < \frac{1}{w_1}$. Proceeding by contradiction, suppose that the $CE(0, j_2)$ is minimized by a vector of unit treatment effects $\beta = (\beta_1, ..., \beta_J)$ with $\beta_1 = ACR/w_1$ and $\beta_j = 0$ for all $j > 1$. Consider a new vector of unit treatment effects, $\tilde{\beta} = (\tilde{\beta}_1, ..., \tilde{\beta}_J)$, where $\tilde{\beta}_j = \tilde{\beta}_{j+1}$ for all $j \in \{1, ..., J-1\}$. The ACR constraint implies that $\tilde{\beta}_j = ACR$ for all j. Comparing the objective

function under β and $\tilde{\beta}$,

$$\sum_{j=1}^{j_2} \tilde{\beta}_j - \sum_{j=1}^{j_2} \beta_j = ACR \times j_2 - \frac{ACR}{w_1} < 0$$

where the last inequality follows from the assumption that $j_2 < \frac{1}{w_1}$. This shows that $CE(0, j_2) = ACR/w_1$ would not be a lower bound if $j_2 < \frac{1}{w_1}$.

Finally, we prove Proposition 2*(d) and show that $CE(0, j_2)$ is minimized at $ACR \times j_2$ if $j_2 < \frac{1}{w_1}$. To show that $CE(0, j_2) \geq ACR \times j_2$, we will first show that $CE(0, j_2)$ is minimized when $\beta_1 = \beta_2 = ... = \beta_J$. Proceeding by contradiction, suppose that $CE(0, j_2)$ is minimized by a vector of unit effects $\beta = (\beta_1, \beta_2, ...\beta_J)$ where $\beta_j = \beta_{j+1}$ for all j except $j = k$ for some $k \in \{1, 2, ..., J-1\}$. This implies that $\beta_k > \beta_{k+1}$ (concavity rules out the reverse ordering). We can re-write (C.1) under the vector β as:

$$\sum_{j=1}^{J} \beta_j w_j = \sum_{j=1}^{k} \beta_j w_j + \sum_{j=k+1}^{J} \beta_j w_j = \sum_{j=1}^{k} \beta_k w_j + \sum_{j=k+1}^{J} \beta_{k+1} w_j = ACR$$

where the second equality follows from the assumption that $\beta_j = \beta_k$ for all $j \leq k$ and $\beta_j = \beta_{k+1}$ for all $j \in \{k+1, ..., J\}$. Define $\tilde{w}_k = \sum_{j=1}^{k} w_j$ and $\tilde{w}_{k+1} = \sum_{j=k+1}^{J} w_j$. Then, $\beta_k \tilde{w}_k + \beta_{k+1} \tilde{w}_{k+1} = ACR$.

Consider a new vector of unit treatment effects, $(\tilde{\beta}_1, ..., \tilde{\beta}_J)$, with $\tilde{\beta}_k = \beta_k - \frac{(\beta_k - \beta_{k+1})\tilde{w}_{k+1}}{\tilde{w}_k + \tilde{w}_{k+1}}$ and $\tilde{\beta}_{k+1} = \beta_{k+1} + \frac{\tilde{w}_k(\beta_k - \beta_{k+1})}{\tilde{w}_k + \tilde{w}_{k+1}}$, so that $\tilde{\beta}_k = \tilde{\beta}_{k+1}$. Note that $(\tilde{\beta}_1, ..., \tilde{\beta}_J)$ is a feasible unit vector since it satisfies C.1-C.3. Next, we show that with these changes, the value of the objective function is smaller than that of the original vector of the unit effects. If $k \geq j_2$,

$$\sum_{j=1}^{j_2} \tilde{\beta}_j - \sum_{j=1}^{j_2} \beta_j = -j_2 \times \frac{(\beta_k - \beta_{k+1})\tilde{w}_{k+1}}{\tilde{w}_k + \tilde{w}_{k+1}} < 0$$

where the last inequality follows from the assumption that $\beta_k > \beta_{k+1}$. If $k < j_2$,

$$\begin{aligned}
\sum_{j=1}^{j_2} \tilde{\beta}_j - \sum_{j=1}^{j_2} \beta_j &= \sum_{j=1}^{k} \left(\tilde{\beta}_j - \beta_j\right) + \sum_{j=k+1}^{j_2} \left(\tilde{\beta}_j - \beta_j\right) \\
&= -k \times \frac{(\beta_k - \beta_{k+1})\tilde{w}_{k+1}}{\tilde{w}_k + \tilde{w}_{k+1}} + (j_2 - k)\frac{\tilde{w}_k(\beta_k - \beta_{k+1})}{\tilde{w}_k + \tilde{w}_{k+1}} \\
&= \left(\frac{\beta_k - \beta_{k+1}}{\tilde{w}_k + \tilde{w}_{k+1}}\right)\left(j_2 \tilde{w}_k - k(\tilde{w}_k + \tilde{w}_{k+1})\right) \\
&\leq \left(\frac{\beta_k - \beta_{k+1}}{\tilde{w}_k + \tilde{w}_{k+1}}\right)\left(j_2 w_1 - k\right) < \left(\frac{\beta_k - \beta_{k+1}}{\tilde{w}_k + \tilde{w}_{k+1}}\right)\left(1 - k\right) \leq 0
\end{aligned}$$

where the fourth line follows from the fact that $w_1 > w_j$ for all $j > 1$ and $\tilde{w}_k + \tilde{w}_{k+1} = 1$, the fifth line follows from the fact that $j_2 < \frac{1}{w_1} \implies j_2 w_1 < 1$, and the final inequality follows from the fact that $\beta_k > \beta_{k+1}$ by assumption and $k \geq 1$.

This shows that when $j_2 < \frac{1}{w_1}$, $CE(0, j_2)$ is minimized when $\beta_1 = ... = \beta_J$. Using the ACR constraint, this implies that $\beta_j = ACR$ for all j. Using this to re-write $CE(0, j_2)$ as $CE(0, j_2) = \sum_{j=0}^{j_2} \beta_j = ACR \times j_2$. Finally, the same argument as in Appendix Section G can be used to establish that these bounds are sharp. ■

Proposition 2* also sheds light on the *width* of the bounds on $CE(j_1, j_2)$ under Assumptions 1-7. Here, we provide some discussion on the value of j_2 for which the bounds on the $CE(0, j_2)$ are the tightest. First, we note that the upper bound $CE(0, j_2)$ is increasing in $\frac{j_2}{\sum_{j=1}^{j_2} w_j}$. Therefore, increasing j_2 by one unit will increase the upper bound more if there

is a relatively greater drop-off in the share of compliers induced by the instrument to move beyond treatment intensity j_2 i.e., if w_j for $j > j_2$ is smaller. On the other hand, the lower bound on $CE(0, j_2)$ in Proposition 2* is maximized at treatment level j_2 where $j_2 > 1/w_1$. To summarize, under Assumptions 1-7, the bounds on the effect of moving from 0 to j_2 will be tightest when j_2 is such that w_j for $j > j_2$ is small *and* when $j_2 > 1/w_1$.

While Assumption 7 does not directly apply in our analysis of Taubman et al. (2014), this discussion clarifies how the width of the bounds might change depending on the distribution of compliers and the parameter of interest. In particular, we observe relatively tight bounds on the effect of one full-year effect of Medicaid coverage. To understand why this occurs, we can see in Figure A.1(a)-(b) that only about 17% of the compliers cross a treatment intensity of greater than 12 months, and $1/w_1 \approx 13$. This implies that, at $j_2 = 12$, the lower bound may be close to its maximum value and the upper bound may be close to its minimum value. Indeed, this is the case. Figure I.1 plots the lower bound and upper bound on the effect of moving from no Medicaid coverage to j_2 months of Medicaid coverage (shown on the x-axis). As can be seen from the figure, at $j_2 = 12$, the lower bound is only approximately 8% smaller than its maximum value and the upper bound is only approximately 9% larger than its minimum value; hence the bounds (the difference between the red and blue lines) is quite narrow around 12 months.

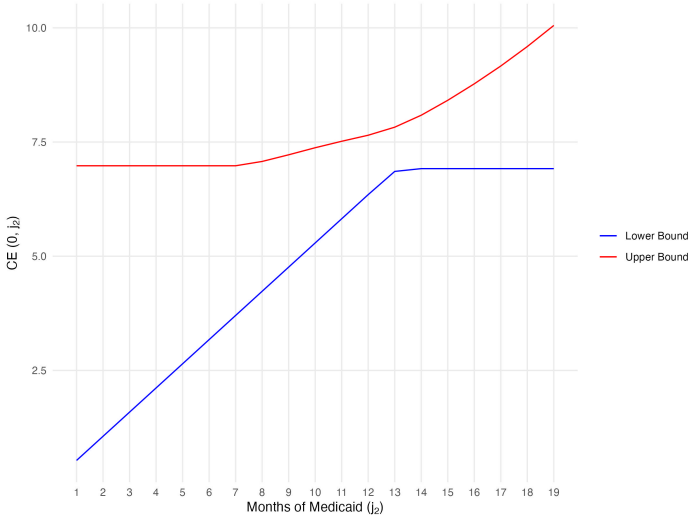


Figure I.1: Bounds on $CE(0, j_2)$ in Taubman et al. (2014)

Notes: The figure displays bounds on the effect of moving from 0 months of Medicaid coverage to j_2 months of Medicaid coverage on Emergency Department usage. Bounds on $CE(0, j_2)$ are displayed in percentage points on the y-axis. The bounds are computed by solving LP.7 for $j_2 = 1, ..., 19$.