Testing instrument validity for LATE identification based on inequality moment constraints

Martin Huber and Giovanni Mellace

University of St. Gallen, Dept. of Economics

Abstract: This paper proposes bootstrap tests for the validity of instrumental variables (IV) in just identified treatment effect models with endogeneity. We demonstrate that the IV assumptions required for the identification of the local average treatment effect (LATE) allow us to both point identify and bound the mean potential outcomes (i) of the always takers (those treated irrespective of the instrument) under treatment and (ii) of the never takers (never treated irrespective of the instrument) under non-treatment. The point identified means must lie within their respective bounds, which provides four testable inequality moment constraints for IV validity. Furthermore, we show that a similar logic applies to testing the assumptions needed to identify distributional features (e.g., local quantile treatment effects). Finally, we discuss how testing power can be increased by imposing dominance/equality assumptions on the potential outcome distributions of different subpopulations.

Keywords: specification test, instrument, treatment effects, LATE, inequality moment constraints.

JEL classification: C12, C15, C21, C26.

We have benefited from comments by Alberto Abadie, Clément De Chaisemartin, Stefan Hoderlein, Bo Honoré, Guido Imbens, Toru Kitagawa, Michael Lechner, Jeff Smith, Anthony Strittmatter, Gautam Tripathi, and seminar/conference participants in Bludenz (internal research seminar, June 2011) and Harvard (IZA Conference on Labor Market Policy Evaluation, October 2011). Addresses for correspondence: Martin Huber (martin.huber@unisg.ch), Giovanni Mellace (giovanni.mellace@unisg.ch), SEW, University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland.

1 Introduction

In many economic evaluation problems causal inference is complicated by endogeneity, implying that the explanatory or treatment variable of interest is correlated with unobserved factors that also affect the outcome. E.g., when estimating the returns to education, the schooling choice is plausibly influenced by unobserved ability (see for instance Card, 1999) which itself most likely has an impact on the earnings outcome. Due to the endogenous treatment selection (also known as selection on unobservables) the earnings effect of education is confounded with the unobserved terms. In the presence of endogeneity, identification relies on the availability of an instrumental variable (IV) that generates exogenous variation in the treatment. In heterogenous treatment effect models with a binary treatment (which allow for effect heterogeneity across different subpopulations), an instrument is valid if (i) the potential outcomes are mean independent of the instrument, (ii) the potential treatment states are not confounded by instrument assignment, and (iii) the treatment is weakly monotonic in the instrument. In this case, the local average treatment effect (LATE) on those who switch their treatment state as a reaction to a change in the instrument (the so called compliers) is identified,[1] see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996).[2]

As endogenous treatment selection is an ubiquitous problem in economics, it is no surprise that IV estimation is a corner stone of empirical research. Taking the estimation of the returns to education as an example, a range of instruments have been suggested to control for the endogenous choice of schooling. Angrist and Krueger (1991) use quarter of birth which is related to years of education through regulations concerning the school starting age but arguably does not have a direct effect on income. Card (1995) exploits geographical proximity to college (which should affect the cost of college education) as instrument for going to college. Further influential studies in labor economics include Angrist (1990), who uses the US draft lottery as instrument for Vietnam veteran status in order to estimate its income effect, and Angrist

[1] For the identification of (local) quantile treatment effects the mean independence assumptions have to be strengthened to full independence of the instrument and the joint distribution of potential treatments and potential outcomes, see Frölich and Melly (2008).

[2] Note that under the strong restrictions of effect homogeneity and linearity of the outcome equation, an instrument is valid if it is correlated with the treatment and uncorrelated with the error term (monotonicity is imposed by construction in this kind of models), see for instance the text book discussions in Peracchi (2001), Wooldridge (2002), Cameron and Trivedi (2005), and Greene (2008). In this case, the IV estimand can be interpreted as the average treatment effect (ATE), given that the model is correctly specified. Clearly, the weaker IV restrictions (uncorrelation instead of the mean independence restrictions and no assumptions on the first stage) are bought by stronger structural assumptions. Then, IV validity cannot be tested in just identified models. In the subsequent discussion we will focus on heterogenous treatment effect models and show that the LATE assumptions have testable implications for IV validity.

and Evans (1998), who investigate the impact of fertility on female labor supply. They use the sex ratio of the first two children as an instrument. Having two children with the same sex should increase the likelihood of a third birth, given that some parents have a preference for a mixed sibling-sex composition.

Many, if not most instruments are far from being undisputed. E.g., the validity of quarter of birth instruments is contested in Bound, Jaeger, and Baker (1995). Based on their estimations and other studies, the authors present evidence on seasonal patterns of births that are related to family income, physical and mental health, and school attendance rates, factors which may be correlated with potential wages. As a further example, Rosenzweig and Wolpin (2000) criticize the sex ratio-instrument of Angrist and Evans (1998) by arguing that having mixed-sex siblings may directly affect both the marginal utility of leisure and child rearing costs and, thus, labor supply. Up to date, arguments in favor or against IV validity are predominantly discussed on theoretical and behavioral bases, which are frequently not unanimously accepted among researchers. In contrast, hypothesis tests have not played any role in applications with just identified models.[3]

Kitagawa (2008), henceforth K08, provides the first formal test for just identified heterogenous treatment effect models with a binary instrument based on somewhat more restrictive assumptions than the ones outlined above, i.e., full independence of the potential outcomes/treatment states and the instrument instead of mean independence. His method is based on the fact that the potential outcome distribution under treatment of the always takers (those treated irrespective of the instrument) as well as the joint distribution of the always takers and compliers are point identified if the instrument is valid. As shown in Imbens and Rubin (1997), the difference of both yields to the distribution under treatment of the compliers. An equivalent result holds for the identification of the compliers outcome distribution under non-treatment. Naturally, the density of the complier outcomes under treatment and non-treatment must not be negative, which is a testable implication. Therefore, K08 tests whether negative densities occur in subsets of the outcome distribution and uses a bootstrap method for inference.

The first contribution of this paper is the proposition of alternative tests that are based on mean potential outcomes instead of densities. In the case of a binary instrument, the underlying intuition is as follows: Under IV validity, the mean potential outcome of the always takers under treatment is point identified. It simply corresponds to the observed mean outcome in the treated subpopulation that does

[3] In contrast, tests for IV validity are available for overidentified models where the number of instruments exceeds the number of endogenous regressors. Sargan (1958) was the first to propose such a test for the linear IV model with homogenous effects.

not receive the instrument. For the same potential outcome, one can derive upper and lower bounds in the treated subpopulation receiving the instrument, where the width of the bounds depends of the relative shares of compliers and always takers. Clearly, the point identified mean outcome in the absence of the instrument must lie within the parameter bounds in the presence of the instrument.

If this condition is violated, the instrument either has a direct effect on the mean potential outcome of the always takers, or the treatment is not monotonic in the instrument, or both. An equivalent result holds for the never takers (those never treated irrespective of the instrument) by considering the outcomes of the non-treated receiving the instrument and the non-treated not receiving the instrument. Therefore, the LATE framework provides us with four testable inequality moment constraints based on point identifying and bounding the mean potential outcomes of the always takers under treatment and the never takers under non-treatment. For the practical implementation we consider three different bootstrap methods of which the minimum p-value-based test with partial recentering proposed by Bennett (2009) appears to have the best finite sample properties. As the K08 test, our approach tests for necessary, albeit not sufficient conditions for IV validity. The latter requires the mean potential outcomes of the always/never takers to be equal across different instrument states. However, only the inequality moment constraints are testable, rather than equality of means. For this reason, our test becomes more powerful as the bounds shrink or, put differently, as the compliers' share becomes relatively smaller to the fractions of always takers and never takers, respectively.

As a second contribution, we therefore show how the width of the bounds can be tightened to increase testing power by imposing dominance of the mean potential outcome of one population over another (see also Huber and Mellace, 2010, and Zhang and Rubin, 2003). Testing power is maximized if equality in mean potential outcomes is assumed. Then, the bounds collapse to a point and the inequality constraints turn into equality constraints. E.g., given that the mean potential outcomes of the the always takers and compliers are equal, IV validity implies that the mean outcome of the treated receiving the instrument is equal to that of the treated not receiving the instrument. This can be easily tested by difference of means tests. An analogous result holds for the never takers and the compliers under non-treatment.

Our third contribution is the extension of our testing approach to potential outcome distributions rather than potential means, which requires joint independence of the instrument and the potential treatments/outcomes. Starting with the upper bounds on the potential outcome distributions of the always takers under treatment and the never takers under non-treatment, we derive constrains that are equivalent to K08, namely that complier densities must not be negative in the mixed populations (where both compliers and always or never takers occur). In addition, we show that also the lower bounds provide two testable implications which have not been considered yet. The latter reflect the intuitive fact that under the null, the joint probability of being a complier and having an outcome that lies within a subinterval of the support must never be larger than the (unconditional) complier share in the population. Similar to the tests based on mean independence, we also demonstrate how power can be further increased by imposing stochastic dominance or equality assumptions on the potential outcome distributions of different subpopulations.

The remainder of the paper is organized as follows. Section 2 discusses the IV assumptions in the LATE framework and the testable implications. Section 3 proposes bootstrap tests based on moment inequality constraints. Section 4 shows how mean dominance and equality restrictions can be used (on top of the standard assumptions) to increase testing power. A generalization to non-binary instruments is provided in Section 5. Testing under the stronger joint independence assumption is discussed in Section 6. Simulation results are presented in Section 7. In Section 8, we apply our methods to the labor market data of Card (1995). Section 9 concludes.

2 IV assumptions and testable implications

Suppose that we are interested in the average effect of a binary and endogenous treatment $D \in \{1, 0\}$ (e.g., participation in a training) on an outcome Y (e.g., earnings) with bounded support evaluated at some point in time after the treatment. Under endogeneity, the effect of D is confounded with some unobserved term U that is correlated with both the treatment and the outcome. Therefore, identification of treatment effects requires an instrument (Z) that shifts the treatment but does not have a direct effect on the mean outcome (i.e., any mean impact other than through the treatment). Denote by $D(z)$ the potential treatment state for $Z = z$, and by $Y(d, z)$ the potential outcome for treatment $D = d$ and $Z = z$ (see for instance Rubin, 1974, for a discussion of the potential outcome notation). In heterogenous treatment effect models, the observed outcome of some individual i can be written as $Y_i = \varphi(D_i, Z_i, U_i)$, where φ denotes a general function that might be unknown to the researcher. Likewise, the potential outcome is the value individual i would receive if the treatment and the instrument were set to particular states, $Y_i(d, z) = \varphi(d, z, U_i)$.

As we observe only one potential outcome for each individual, any identification strategy relies on identifying assumptions, which may or may not be (partially) testable. Here, we will focus on those assumptions required for LATE identification. The first restriction maintained throughout the discussion is the so-called Stable Unit Treatment Value Assumption (SUTVA, e.g., Rubin, 1990), which rules out interference between units and general equilibrium effects of the treatment. The SUTVA is formalized in Assumption 1 (see also Angrist, Imbens, and Rubin, 1996) and states that the potential treatments and outcomes of any subject i are unrelated to the actual treatment and instrument states of any other individual:

Assumption 1:

$Y_i(d,z) \perp (D_j, Z_j)$ and $D_i(z) \perp Z_j$, $\forall j \neq i$, $d \in \{0,1\}$, and z in the support of Z (SUTVA).

For the sake of expositional ease, we will henceforth assume the instrument to be binary ($Z \in \{0,1\}$), while Section 5 will generalize the results to bounded non-binary instruments. As discussed in Angrist, Imbens, and Rubin (1996), the population can then be categorized into four types (denoted by T), according to the treatment behavior as a function of the binary instrument. The compliers react on the instrument in the intended way by taking the treatment when $Z = 1$ and abstaining from it when $Z = 0$. For the remaining three types $D(z) \neq z$ for either $Z = 1$, or $Z = 0$, or both: The always takers are always treated irrespective of the instrument state, the never takers are never treated, and the defiers only take the treatment when $Z = 0$, see Table 1.

Table 1: Types

Type T	D(1)	D(0)	Notion
at	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
nt	0	0	Never takers

We cannot directly infer on the type of any individual as either $D(1)$ or $D(0)$ is observed, but never both. Without further assumptions, neither the share of the different types nor their mean potential outcomes are identified. We therefore impose the following unconfounded type assumption, which implies that the instrument is assigned independently of the potential treatment states:

Assumption 2:

$\Pr(T = t|Z = 1) = \Pr(T = t|Z = 0)$ for $t \in \{at, c, d, nt\}$ (unconfounded type).

Under Assumption 2, the share of any type conditional on the instrument is equal to its unconditional proportion in the entire population. It is worth noting that it is not required in the linear model with effect homogeneity, because the latter imposes considerably stronger functional form assumptions than the nonparametric framework considered here. Let $\pi_t \equiv \Pr(T = t)$, $t \in \{at, c, nt\}$, represent the (unobserved) probability to belong to type T in the population and denote by $P_{d|z} \equiv \Pr(D = d|Z = z)$ the (observed) conditional treatment probability given the instrument. Assumption 2 implies that any of the four conditional treatment probabilities is a combination of two unobserved type proportions, see Table 2.

Table 2: Observed probabilities and type proportions

Cond. treatment prob.	type proportions
$P_{1\|1} \equiv \Pr(D = 1\|Z = 1)$	$\pi_{at} + \pi_c$
$P_{0\|1} \equiv \Pr(D = 0\|Z = 1)$	$\pi_d + \pi_{nt}$
$P_{1\|0} \equiv \Pr(D = 1\|Z = 0)$	$\pi_{at} + \pi_d$
$P_{0\|0} \equiv \Pr(D = 0\|Z = 0)$	$\pi_c + \pi_{nt}$

Similarly, each of the four observed conditional means $E(Y|D = d, Z = z)$ is a mixture or weighted average of the mean potential outcomes of two types (denoted by $E(Y(d, z)|T = t)$), where the weights depend on the relative proportions:

$$E(Y|D = 1, Z = 1) = \frac{\pi_{at}}{\pi_{at} + \pi_c} \cdot E(Y(1, 1)|T = at) + \frac{\pi_c}{\pi_{at} + \pi_c} \cdot E(Y(1, 1)|T = c), \quad (1)$$

$$E(Y|D = 1, Z = 0) = \frac{\pi_{at}}{\pi_{at} + \pi_d} \cdot E(Y(1, 0)|T = at) + \frac{\pi_d}{\pi_{at} + \pi_d} \cdot E(Y(1, 0)|T = d), \quad (2)$$

$$E(Y|D = 0, Z = 0) = \frac{\pi_c}{\pi_{nt} + \pi_c} \cdot E(Y(0, 0)|T = c) + \frac{\pi_{nt}}{\pi_{nt} + \pi_c} \cdot E(Y(0, 0)|T = nt), \quad (3)$$

$$E(Y|D = 0, Z = 1) = \frac{\pi_d}{\pi_{nt} + \pi_d} \cdot E(Y(0, 1)|T = d) + \frac{\pi_{nt}}{\pi_{nt} + \pi_d} \cdot E(Y(0, 1)|T = nt). \quad (4)$$

From Table 2 and expressions (1) to (4) it becomes obvious that further assumptions are necessary to identify the LATE, namely a mean exclusion restriction, monotonicity of the treatment in the instrument, and the existence of compliers. Starting with the mean exclusion restriction, it is required that the instrument does not exhibit an effect on the mean potential outcomes within any subpopulation (however, it may affect higher moments):

Assumption 3:

$E(Y(d, 1)|T = t) = E(Y(d, 0)|T = t) = E(Y(d)|T = t)$ for $d \in \{0, 1\}$ and $t \in \{at, c, d, nt\}$ (mean exclusion restriction),

where the last equality makes explicit that the mean potential outcomes are not a function of the instrument. Notice that the mean exclusion restriction is a stronger assumption than uncorrelation of the instrument and the unobserved term, which is invoked in standard IV models with a linear outcome equation and homogeneous treatment effects. This is, together with Assumption 2, the price to pay for considering more flexible models (in terms of effect heterogeneity), which also gives rise to our testable implications.

By the mean exclusion restriction,

$$E(Y(1,1)|T=at) = E(Y(1,0)|T=at) = E(Y(1)|T=at)$$

and

$$E(Y(0,1)|T=nt) = E(Y(0,0)|T=nt) = E(Y(0)|T=nt),$$

which provides the base for the testable implications outlined further below. Alternatively to Assumptions 2 and 3, one may assume that they only hold conditional on a vector of observed variables X as considered in Frölich (2007), who shows nonparametric identification of the LATE in the presence of a conditionally valid instrument (given X). In the subsequent discussion, conditioning on X will be kept implicit, such that all results either refer to an supposedly unconditionally valid instrument or to an analysis within cells of X.

The final two assumptions required for LATE identification put restrictions on the (non-)existence of particular types.

Assumption 4:

$\Pr(D(1) \geq D(0)) = 1$ (monotonicity).

Assumption 4 states that the potential treatment state of any individual does not decrease in the instrument. This rules out the existence of defiers (type d). Note that monotonicity is also implicitly assumed in the linear IV model, where the effect of the instrument on the treatment is represented by a homogenous first stage coefficient.

Assumption 5:

$\Pr(D(1) > D(0)) > 0$ (existence of compliers).

By Assumption 5, a subpopulation of individuals reacts on the instrument such that compliers do exist. Assumptions 4 and 5 together state that $E(D|Z=1) - E(D|Z=0) > 0$, i.e., that the instrument has an effect on the treatment. In the IV linear model this implies that the first stage coefficient must not be zero, which is also referred to as IV relevance.

As defiers do not exist, the proportions of the remaining types are identified by $P_{0|1} = \pi_{nt}$, $P_{1|0} = \pi_{at}$, $P_{1|1} - P_{1|0} = P_{0|0} - P_{0|1} = \pi_c$. Furthermore, the mean potential outcomes of the always takers under treatment and the never takers under non-treatment are point identified. Expression (2) simplifies to $E(Y|D = 1, Z = 0) = E(Y(1, 0)|T = at) = E(Y(1)|T = at)$ under Assumptions 1 to 4, and (4) becomes $E(Y|D = 0, Z = 1) = E(Y(0, 1)|T = nt) = E(Y(0)|T = nt)$. This allows identifying the mean potential outcomes of the compliers under treatment and non-treatment by

$$\begin{aligned} & E(Y(1)|T = c) = E(Y(1, 1)|T = c) \\ = & \frac{\Pr(D = 1|Z = 1) \cdot E(Y|D = 1, Z = 1) - \Pr(D = 1|Z = 0) \cdot E(Y|D = 1, Z = 0)}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)} \end{aligned} \quad (5)$$

and

$$\begin{aligned} & E(Y(0)|T = c) = E(Y(0, 0)|T = c) \\ = & \frac{\Pr(D = 0|Z = 0) \cdot E(Y|D = 0, Z = 0) - \Pr(D = 0|Z = 1) \cdot E(Y|D = 0, Z = 1)}{\Pr(D = 1|Z = 1) - \Pr(D = 1|Z = 0)}. \end{aligned} \quad (6)$$

Therefore, the LATE on the compliers is obtained by the difference of (5) and (6) , which simplifies to

$$\begin{aligned} & \frac{\Pr(D = 1|Z = 1) \cdot E(Y|D = 1, Z = 1) + \Pr(D = 0|Z = 1) \cdot E(Y|D = 0, Z = 1)}{E(D|Z = 1) - E(D|Z = 0)} \\ - & \frac{\Pr(D = 1|Z = 0) \cdot E(Y|D = 1, Z = 0) + \Pr(D = 0|Z = 0) \cdot E(Y|D = 0, Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} \\ = & \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}. \end{aligned} \quad (7)$$

The last line gives the well known result that the LATE is identified by the ratio of two differences of conditional expectations, namely the intention to treat effect divided by the share of compliers, see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996).

Our discussion has demonstrated that Assumptions 1-4 allow pinning down the type proportions as well as the mean potential outcomes of the always takers and never takers under treatment and non-treatment. Together with Assumption 5, this identifies the LATE and the mean potential outcomes of the compliers.[4] However, this framework also provides testable implications for IV validity based on deriving bounds on

[4] An equivalent result for the potential outcome distributions of the compliers under slightly stronger assumptions has been derived by Imbens and Rubin (1997).

the mean potential outcomes of the always takers and never takers in equations (1) and (3), respectively. In fact, the mean potential outcome of the always takers in equation (1) is bounded by the mean over the upper and lower proportion of outcomes that corresponds to the share of the always takers in this mixed population. It is obvious that $E(Y|D=1,Z=0)=E(Y(1)|T=at)$ must lie within these bounds, otherwise either Z has a direct effect on the mean of Y, or the potential treatment state is confounded with the instrument, or defiers exist in (2), or any combination of these violations occurs. An equivalent result applies to the never takers under non-treatment.

To formalize the discussion, we introduce some further notation and assume for the moment that the outcome is continuous, while Appendix A.4 shows how the following intuition and the test procedure discussed in the next section can be adapted to discrete outcomes. Define the qth conditional quantile of the outcome $y_q \equiv G^{-1}(q)$, with G being the cdf of Y given $Z=1$ and $D=1$. Furthermore, let q correspond to the proportion of always takers in (1): $q=\frac{\pi_{at}}{\pi_{at}+\pi_c}=\frac{P_{1|0}}{P_{1|1}}$. By the results of Horowitz and Manski (1995) (see also the discussion in Huber and Mellace, 2010), $E(Y|D=1,Z=1,Y\leq y_q)$ is the sharp lower bound of the mean potential outcome of the always takers, implying that all the always takers are concentrated in the lower tail of the distribution that corresponds to their proportion. Similarly, $E(Y|D=1,Z=1,Y\geq y_{1-q})$ is the upper bound by assuming that any always taker occupies a higher rank in the outcome distribution than any complier. Therefore, the IV assumptions imply that

$$E(Y|D=1,Z=1,Y\leq y_q)\leq E(Y|D=1,Z=0)\leq E(Y|D=1,Z=1,Y\geq y_{1-q}). \tag{8}$$

Equivalent arguments hold for the mixed outcome equation of never takers and compliers. Let $y_r \equiv F^{-1}(r)$, with F being the cdf of Y given $D=0,Z=0$ and $r=\frac{\pi_{nt}}{\pi_{nt}+\pi_c}=\frac{P_{0|1}}{P_{0|0}}$, i.e., the proportion of never takers in equation (3). Taking the mean over the lower and upper share of the outcome distribution corresponding to r we obtain the lower and upper bounds $E(Y|D=0,Z=0,Y\leq y_r)$, $E(Y|D=0,Z=0,Y\geq y_{1-r})$ on the mean potential outcome of the never takers. The latter is also point identified by $E(Y|D=0,Z=1)=E(Y(0)|T=nt)$, such that the IV assumptions require that

$$E(Y|D=0,Z=0,Y\leq y_r)\leq E(Y|D=0,Z=1)\leq E(Y|D=0,Z=0,Y\geq y_{1-r}). \tag{9}$$

Note that under one-sided non-compliance, only one of (8) and (9) can be tested. Furthermore, monotonicity holds by construction in this case such that a violation of the remaining testable constraint points to a non-satisfaction of the exclusion restriction. E.g., when there are no observations with $Z = 0$ and $D = 1$, always takers do not exist ($\pi_{at} = 0$) and $E(Y|D = 1, Z = 0)$ is not defined. In addition, the latter case also rules out the existence of defiers. Therefore, monotonicity is satisfied, but (9) is still useful to test the exclusion restriction on the never takers.

3 Testing

Expressions (8) and (9) provide us with four testable inequality moment constraints.[5] Under the null hypothesis that the instrument is valid it must hold that

$$H_0 : \begin{pmatrix} E(Y|D=1,Z=1,Y\leq y_q) - E(Y|D=1,Z=0) \\ E(Y|D=1,Z=0) - E(Y|D=1,Z=1,Y\geq y_{1-q}) \\ E(Y|D=0,Z=0,Y\leq y_r) - E(Y|D=0,Z=1) \\ E(Y|D=0,Z=1) - E(Y|D=0,Z=0,Y\geq y_{1-r}) \end{pmatrix} \equiv \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \qquad (10)$$

Under the alternative hypothesis that IV validity does not hold at least one and at most two constraints *might* be binding. This is the case because violations of the first and second as well as of the third and fourth constraints are mutually exclusive, respectively. Furthermore, note that even if no inequality constraint is violated, IV validity may not be satisfied. I.e., we detect violations only if they are large enough such that the point identified mean outcomes of the always takers and/or never takers lie outside their respective bounds in the mixed populations. Ideally, we would like to test for the equality of the mean outcomes of the respective population across instrument states. However, this is not feasible as it remains unknown which individuals in the mixed populations belong to the group of always/never takers or compliers. Therefore, without further assumptions, testing based on inequality moment constraints is the best one can get. It is obvious that such tests gain power as the proportion of compliers decreases, implying that the bounds on the mean outcomes of the always and never takers become tighter.

Several methods have been proposed for testing inequality constraints. The first approach generalizes the standard Wald, LM, and LR statistics to test for inequality constraints and goes back to Wolak (1987, 1989b), who considers a linear regression framework, as well as Kodde and Palm (1986) and Wolak (1989a,

[5] Note that expressions (8) and (9) hold under Assumptions 1-4 alone, i.e., the existence of compliers (Assumption 5) is not required. In principle, one could therefore test for IV validity even if the LATE is not identified, which might, however, not be an interesting exercise in applied work.

1991), who propose tests for nonlinear models. The idea is to compare the parameter estimates of the constrained model with those of the unconstrained model under the least favorable configuration (LFC, i.e., the parameter configuration for which the null is rejected with the lowest probability) to obtain a test with asymptotically exact size. The limitation of this approach is that when the covariance matrix of the parameters depends on unknown parameters, finding the LFC might become complicated. Indeed, the limiting distribution of the test statistic is a non-trivial mixture of χ^2 distributions (see Perlman, 1969 and Kudo, 1963) with weights that depend on the covariance matrix of the parameters which in turn depends on the unknown parameters. The appendix outlines the estimation of the parameter vector $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)^T$ in a GMM framework and the construction of the test proposed by Wolak (1991).

An alternative is to use tests that are based on the bootstrap (see Efron, 1979). As shown in Appendix A.1, we can easily obtain asymptotically normally distributed estimators of all components of θ such that θ itself has a continuous asymptotic distribution, which justifies the use of the bootstrap as valid inference method. Apart from circumventing the problem of deriving the non-trivial limiting distribution of the test statistic, bootstrap procedures are often more accurate in finite samples than methods relying asymptotic theory (which may be a poor approximation for the sample at hand). Therefore, we consider three different bootstrap tests and evaluate their finite sample performance in Section 7.

Let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)^T$ denote the vector of estimates of the respective population parameters θ based on an i.i.d. sample containing n observations. Furthermore, denote by $\hat{\theta}_b = (\hat{\theta}_{1,b}\hat{\theta}_{2,b}, \hat{\theta}_{3,b}, \hat{\theta}_{4,b})^T$ ($b \in \{1, 2, ..., B\}$, where B is the number of bootstrap replications) the estimates in a particular bootstrap sample b containing n observations that are randomly drawn from the original data with replacement. The first method is based on the classical nonparametric bootstrap, with the exception that it includes a Bonferroni adjustment to account for the fact that we test four hypotheses jointly.

The basic steps are the following. In each bootstrap sample, we compute the recentered parameter vector $\tilde{\theta}_b = \hat{\theta}_b - \hat{\theta}$. Then, the vector of p-values $P_{\hat{\theta}}$ is estimated by the share of bootstrap replications in which the recentered parameters are larger than the estimates in the original sample:

$$P_{\hat{\theta}} = B^{-1} \sum_{b=1}^{B} I\{\tilde{\theta}_b > \hat{\theta}\}, \tag{11}$$

where $I\{\cdot\}$ denotes the indicator function.

Even though the p-values are consistent for assessing each constraint separately, they are not appropriate for making judgements about the joint satisfaction of the constraints. To the latter end, we use a simple Bonferroni adjustment. As for instance discussed in MacKinnon (2007), the Bonferroni inequality implies that the p-value for joint hypotheses can be computed by multiplying the minimum p-value by the number of constraints, in our case four. Therefore, the p-value of the bootstrap test, denoted by $\hat{p}_{\text{bs}}$, is

$$\hat{p}_{\text{bs}} = 4 \cdot \min(P_{\hat{\theta}}). \tag{12}$$

While this procedure is easy to implement, the Bonferroni adjustment has the disadvantage that it yields too conservative p-values when the test statistics are positively correlated, see for example the discussion in Romano, Shaikh, and Wolf (2008). A further and probably more important limitation is that the power of the test decreases as the number of non-binding constraints increases, which is particularly relevant for the non-binary instrument framework of Section 5. Indeed, $\min(P_{\hat{\theta}})$ is not affected by adding irrelevant constraints, but it will be multiplied by a larger number. This problem and the importance of allowing for a sample dependent null distribution (for which the number of binding constraints are estimated from the data) has been acknowledged in a number of papers such as Andrews and Jia (2008), Andrews and Soares (2010), Bennett (2009), Chen and Szroeter (2009), and Hansen (2005).[6] One might use any approach proposed in these papers or in Donald and Hsu (2010) (which is in the spirit of Hansen, 2005, with the exception that a simulation approach is used instead of bootstrapping) to overcome the limitations of the Bonferroni adjustment.

Here, we consider the novel minimum p-value-type test proposed by Bennett (2009) for joint inequality moment constraints. The test relies on the following, quite general assumptions (see his Assumption 1) which are satisfied in a standard GMM framework that may also be used to characterize our testing problem: (i) i.i.d. sampling, (ii) bounded second moments, (iv) Lipschitz continuity of the moment functions with the Lipschitz function having bounded second moments, (v) linear representation of the testing problem. The Bennett (2009) test not only has an asymptotically exact size, but is - in contrast to Andrews and Jia (2008), Andrews and Soares (2010), Hansen (2005), and Donald and Hsu (2010) - also invariant to studentization. Compared to Chen and Szroeter (2009), it has the advantage that it does not require the choice of any smoothing function. Furthermore, the test does not rely on the double

[6] It is worth mentioning that testing inequality constraints is closely related to the fast evolving literature on inference in models with moment inequalities, see for instance: Andrews and Guggenberger (2007), Andrews and Soares (2010), Chernozhukov, Hong, and Tamer (2007), Fan and Park (2007), Guggenberger, Hahn, and Kim (2008), Linton, Song, and Whang (2008), and Rosen (2008).

(i.e., nested) bootstrap (see Beran, 1988) to estimate the distribution of the minimum p-value $\min(P_{\hat{\theta}})$ as suggested in Godfrey (2005), which may be computationally intensive. Instead, it only demands two individual bootstraps, where the second resamples from the distribution of the first bootstrap. Bennett (2009) considers both full (i.e., standard) recentering of inequality constraints and partial recentering of only those constraints which are either violated in the sample or not violated but within a small neighborhood of the boundary of the null hypothesis. He shows that partial recentering (henceforth mP.p) has weakly superior finite sample properties than full recentering (henceforth mP.f). The algorithm of both methods can be sketched as follows:

1. Estimate the vector of parameters $\hat{\theta}$ in the original sample.

2. Draw B_1 bootstrap samples of size n from the original sample.

3. In each bootstrap sample, compute the recentered vector $\tilde{\theta}_b^f \equiv \hat{\theta}_b - \hat{\theta}$ for the mP.f test and the partially recentered vector $\tilde{\theta}_b^p \equiv \hat{\theta}_b - \max(\hat{\theta}, -\delta_n)$ for the mP.p test, where δ_n is a sequence such that $\delta_n \to 0$ and $\sqrt{n} \cdot \delta_n \to \infty$ as $n \to \infty$.[7]

4. Estimate the vector of p-values for mP.f, denoted by $P_{\tilde{\theta}^f}$:

$$P_{\tilde{\theta}^f} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{\sqrt{n} \cdot \tilde{\theta}_b^f > \sqrt{n} \cdot \hat{\theta}\}. \tag{13}$$

5. Compute the minimum P-values for mP.f:

$$\hat{p}_f = \min(P_{\tilde{\theta}^f}). \tag{14}$$

6. Draw B_2 values from the distributions of $\tilde{\theta}_b^f$ and $\tilde{\theta}_b^p$. We denote by $\tilde{\theta}_{b_2}^f$ and $\tilde{\theta}_{b_2}^p$ the resampled observations in the second bootstrap.

7. In each bootstrap sample, compute the minimum P-values of mP.f and mP.p, denoted by $\hat{p}_{f,b_2}$ and $\hat{p}_{p,b_2}$:

$$\hat{p}_{f,b_2} = \min(P_{\tilde{\theta}^f,b_2}), \qquad \hat{p}_{p,b_2} = \min(P_{\tilde{\theta}^p,b_2}), \tag{15}$$

[7] In the simulations and applications further below, we choose $\delta_n = \sqrt{\frac{2 \cdot \ln(\ln(n))}{n}} \cdot \hat{\sigma}_{\theta_i}, \quad i \in \{1,2,3,4\}$, where $\hat{\sigma}_{\theta_i}$ is the estimated (in the B_1 first stage bootstrap samples) standard deviation of the i-th inequality constraint, as suggested by Bennett (2009). It is, however, not guaranteed that this choice is optimal, see for instance the discussion in Donald and Hsu (2010).

where

$$P_{\tilde{\theta}^f,b_2} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{\sqrt{n} \cdot \tilde{\theta}_b^f > \sqrt{n} \cdot \tilde{\theta}_{b_2}^f\}, \quad P_{\tilde{\theta}^p,b_2} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{\sqrt{n} \cdot \tilde{\theta}_b^f > \sqrt{n} \cdot \tilde{\theta}_{b_2}^p\}. \tag{16}$$

8. Compute the p-values of the mP.f and mP.p tests by the share of bootstrapped minimum p-values that are smaller than the respective minimum p-value of the original sample:

$$\hat{p}_{\text{mP.f}} = B_2^{-1} \cdot \sum_{b_2=1}^{B_2} I\{\hat{p}_{f,b_2} \leq \hat{p}_f\}, \qquad \hat{p}_{\text{mP.p}} = B_2^{-1} \cdot \sum_{b_2=1}^{B_2} I\{\hat{p}_{p,b_2} \leq \hat{p}_f\}. \tag{17}$$

As already mentioned, mP.f and mP.p only differ in terms of recentering. The former test recenters all four constraints, while the latter recenters only the restrictions that either violate the null or are in the null but close (i.e., within δ_n) to equality in the original sample. Partial recentering allows estimating the number of binding constraints from the data and therefore provides a better approximation of the asymptotic distribution of the test under the null. It dominates mP.f in terms of power while yielding asymptotically exact size, see Bennett (2009). This finding is corroborated by the simulation results reported in Section 7.

4 Mean dominance and equality constraints

This section discusses restrictions on the order of the mean potential outcomes of different populations, which were for instance also considered by Huber and Mellace (2010) in an IV context and Zhang and Rubin (2003) in models with censored outcomes. If these mean dominance assumptions appear plausible to the researcher, they may be invoked to increase testing power.

The first assumption considered is mean dominance of the complier outcomes over those of the always takers under treatment:

Assumption 6:
$E(Y(1)|T = c) \geq E(Y(1)|T = at)$ (mean dominance of compliers).

Assumption 6 implies that the mean potential outcome of the compliers under treatment is at least as high as that of the always takers. Therefore, the upper bound of the mean potential outcome of the always takers in Equation (1) tightens to the conditional mean $E(Y|D = 1, Z = 1)$. Under Assumptions 1-6, (8) becomes

$$E(Y|D=1,Z=1,Y\leq y_q)\leq E(Y|D=1,Z=0)\leq E(Y|D=1,Z=1), \tag{18}$$

which generally increases testing power due to the tighter upper bound. Whether this assumption is plausible depends on the empirical application at hand and has to be justified by theory and/or empirical evidence. In fact, one could also assume the converse, i.e., that the mean potential outcome of the compliers cannot be higher than that of the always takers. This is formally stated in Assumption 7:

Assumption 7:
$E(Y(1)|T=c)\leq E(Y(1)|T=at)$ (mean dominance of always takers).

In this case, $E(Y|D=1,Z=1)$ constitutes the lower bound of the mean potential outcome of the always takers, and the testable implication becomes

$$E(Y|D=1,Z=1)\leq E(Y|D=1,Z=0)\leq E(Y|D=1,Z=1,Y\geq y_{1-q}). \tag{19}$$

Finally, the combination of Assumptions 6 and 7 results in the restriction that the mean potential outcomes under treatment of the always takers and compliers are the same, yielding the following equality constraint:

Assumption 8:
$E(Y(1)|T=c)=E(Y(1)|T=at)$ (equality of means).

Clearly, Assumption 8 entails the highest testing power and implies that

$$E(Y|D=1,Z=1)=E(Y|D=1,Z=0), \tag{20}$$

such that the inequality restrictions turn into an equality constraint. Then, the validity of the instrument can be tested by a simple two sample t-test for differences between means. To be precise, the latter tests the IV assumptions and Assumption 8 jointly: A non-rejection points to both a valid instrument and homogeneity of the mean potential outcomes of always takers and compliers under treatment. Note that equivalent results under mean dominance/equality apply to the compliers and never takers under non-treatment. E.g., assuming $E(Y(0)|T=c)=E(Y(0)|T=nt)$

amounts to testing whether

$$E(Y|D=0,Z=1)=E(Y|D=0,Z=0). \tag{21}$$

5 Generalization to non-binary instruments

This section generalizes the testable implications derived under mean independence to bounded non-binary instruments, which is straightforward in our framework based on moment inequalities. As discussed in Frölich (2007), in cases where the support of Z is bounded such that $Z \in [z_{\min}, z_{\max}]$, it is possible to define and identify LATEs with respect to any two distinct subsets of the support of Z. To this end, we need to invoke Assumption 1 along with 2NB to 5NB, which are generalizations of Assumptions 2 to 5:

Assumption 2NB:

$\Pr(T=t|Z=z)=\Pr(T=t)\ \forall\ z$ in the support of Z (unconfounded type).

Assumption 3NB:

$E(Y(d,z)|T=t)=E(Y(d)|T=t)\ \forall\ z$ in the support of Z, $d \in \{0,1\}$, and $t \in \{at,c,d,nt\}$ (mean exclusion restriction).

Assumption 4NB:

$\Pr(D(z) \geq D(z'))=1\ \forall\ z,z'$ satisfying $z_{\min} \leq z' < z \leq z_{\max}$ (monotonicity).

I.e., z and z' are two distinct subsets of the support of Z such that any element in z is larger than any element in z'.

Assumption 5NB:

$\Pr(D(z_{\max}) > D(z_{\min}))>0$ (existence of compliers).

Then, the LATE for compliers defined upon any z,z' satisfying $z_{\min} \leq z' \leq z \leq z_{\max}$ and $\Pr(D(z) > D(z'))>0$ is identified by

$$E(Y(1)-Y(0)|D(Z\in z) > D(Z \in z'))=\frac{E(Y|Z\in z)-E(Y|Z\in z')}{E(D|Z\in z)-E(D|Z\in z')}. \tag{22}$$

Theorem 8 in Frölich (2007) shows that the LATE on the largest complier population possible is identified by choosing $z=z_{\max}$ and $z'=z_{\min}$. In this light, the logic of Assumption 5NB becomes more apparent: If it is not satisfied for $z_{\max}, z_{\min}$, it does not hold for any pair of z, z'. However, Assumption 5NB merely states that compliers exist for at least one combination of distinct values of Z, but not necessarily for all pairs of subsets z, z'. As monotonicity of the binary treatment implies that each individual switches its treatment status as a reaction to the instrument at most once under the null, the complier share may be small or even zero for some pairs z, z'.

While small or zero complier shares are undesirable for LATE estimation, the contrary holds for testing, as $\pi_c = 0$ maximizes asymptotic power. A further dimension relevant to testing power is the number of subsets considered. I.e., it is useful to look at all possible pairs of neighboring[8] z and z' for which the moment inequalities in (24) must be satisfied under instrument validity. In large samples small subsets therefore appear preferable, firstly to minimize the complier share and secondly to maximize the number of neighboring pairs of z and z'. However, in small samples a trade-off between finite sample power and asymptotic power may well occur when doing so.

To generalize the testable implications to the non-binary case, define $\tilde{Z}$ as

$$\tilde{Z} = \begin{cases} 1 & \text{if } Z \in z \\ 0 & \text{if } Z \in z' \end{cases}. \tag{23}$$

Under Assumptions 1 and 2NB to 5NB, the results of Sections 2 and 3 must also hold when replacing Z by $\tilde{Z}$. This implies that for any $\tilde{Z}$, we obtain four inequality constraints:

$$\begin{pmatrix} E(Y|\tilde{Z}=1, D=1, Y \leq y_q) - E(Y|\tilde{Z}=0, D=1), \\ E(Y|\tilde{Z}=0, D=1) - E(Y|\tilde{Z}=1, D=1, Y \geq y_{1-q}), \\ E(Y|\tilde{Z}=0, D=0, Y \leq y_r) - E(Y|\tilde{Z}=1, D=0), \\ E(Y|\tilde{Z}=1, D=0) - E(Y|\tilde{Z}=0, D=0, Y \geq y_{1-r}) \end{pmatrix} \leq 0.$$

Let $n_{\tilde{Z}}$ be the number of possible choices of $\tilde{Z}$ with neighboring subsets. Testing IV validity amounts to applying the test procedures outlined in Section 3, where the number of inequality constraints is now $4 \cdot n_{\tilde{Z}}$ instead of 4. To give an example, consider the case that Z may take the values 0, 1, or 2. The

[8] Note that under Assumptions 4ND, 5ND and for any fixed z, neighboring z and z' give weakly lower complier shares than non-neighboring pairs and thus, entail a higher asymptotic power.

number of possible definitions of $\tilde{Z}$ with neighboring z, z' is 4:

$$\begin{aligned} z' &= 0 & z &= 1, \\ z' &= 1 & z &= 2, \\ z' &= 0 & z &= \{1,2\}, \\ z' &= \{0,1\} & z &= 2. \end{aligned}$$

This implies that we have $4 \times 4 = 16$ testable inequality constraints based on neighboring pairs.

Notice that also considering the non-neighboring pair $z' = 0, z = 2$ does neither increase finite sample power nor asymptotic power: A test base on the non-neighboring pair is weakly dominated by using $z' = \{0,1\}, z = 2$ and $z' = 0, z = \{1,2\}$ in terms of the sample size (which influences finite sample power) and entails a weakly higher complier share than any other neighboring pair. As a final remark, note that $n_{\tilde{Z}}$ becomes infinite when the instrument is continuous. In practice, the researcher will have to define a finite number of subsets that depends on the richness of the data in the application considered and will, thus, again face a trade-off between asymptotic and finite sample power.

6 Testing under joint independence

Even though stronger than necessary for LATE identification, the literature commonly imposes the following joint independence assumption instead of Assumptions 2 and 3, see for instance Imbens and Angrist (1994):

Assumption 2J:
$Y(d,z) = Y(d)$ and $Z \perp (Y(d), D(z))\ \forall\ d \in \{0,1\}$ and z in the support of Z (joint independence).

Assumption 2J states that the potential outcome is a function of treatment, but not of the the instrument (such that the exclusion restriction holds for any moment) and that the instrument is independent of the joint distribution of the potential treatment states and the potential outcomes. It is sufficient for the identification of local quantile treatment effects, see Frölich and Melly (2008), or other distributional features.

One plausible reason for the popularity of this assumption in LATE estimation is that in many empirical setups, it does not seem too unlikely that when mean independence holds, also the stronger joint independence is satisfied. E.g., if one is willing to assume that an instrument is mean independent of the outcome variable hourly wage, it might appear reasonable to assume that it is mean independent of the log of hourly wage, too. As the latter is a (one to one) nonlinear transformation of the original outcome variable, this also implies independence w.r.t. higher moments. From this perspective, strengthening mean independence to joint independence may often only come with little costs in terms of credibility.[9] The subsequent review of the K08 test and the adaptation of our method to joint independence makes it obvious that Assumption 2J allows constructing asymptotically more powerful tests based on probability measures (such as density functions) rather than means only. However, it remains to be shown how to optimally define these probability measures in finite samples. From a practical point of view, the mean-based tests may therefore appear useful even under joint independence due to their ease of implementation (and potentially better finite sample properties when compared to tests based on ill-chosen probability measures, see Section 7).

Henceforth assuming a binary instrument, the testing approach proposed in K08 exploits the fact that under IV validity (now relying on Assumption 2J instead of 2 and 3) and for any subset V of the support of Y, $\Pr(Y \in V, D = d|Z = d) - \Pr(Y \in V, D = d|Z = 1 - d)$ can be shown to be equal to $\Pr(Y \in V|D = d) \cdot \pi_c$, and thus, cannot be negative for $d \in \{0, 1\}$. The underlying intuition is that negative densities of complier outcomes must not occur in either treatment state, see Section 1. This is formally stated in Proposition 1 of K08:[10]

$$\begin{aligned}
\Pr(Y \in V, D = 1|Z = 0) &\leq \Pr(Y \in V, D = 1|Z = 1), \\
\Pr(Y \in V, D = 0|Z = 1) &\leq \Pr(Y \in V, D = 0|Z = 0) \quad \forall\ V \text{ in the support of } Y. \qquad (24)
\end{aligned}$$

Concerning the implementation of the test, K08 proposes the following bootstrap method. Let n_1, n_0 denote the numbers of observations with $Z = 1$ and $Z = 0$, respectively. Furthermore, define $P(V, d) \equiv \Pr(Y \in V, D = d|Z = 1)$ and $Q(V, d) \equiv \Pr(Y \in V, D = d|Z = 0)$. The sample analogues of these joint probabilities given Z are, respectively,

$$\begin{aligned}
P_{n_1}(V, d) &= \frac{1}{n_1} \cdot \sum_{i\ :\ Z_i=1} I\{Y_i \in V \text{ and } D_i = d\}, \\
Q_{n_0}(V, d) &= \frac{1}{n_0} \cdot \sum_{j\ :\ Z_j=0} I\{Y_j \in V \text{ and } D_j = d\}. \qquad (25)
\end{aligned}$$

[9] We thank Toru Kitagawa for a fruitful discussion on this topic.

[10] Equation (7) in Balke and Pearl (1997) and equations (12) and (13) in Richardson, Evans, and Robins (2011) provide the same constraints for the special case that Y is binary.

K08 defines the following Kolmogorov-Smirnov-type test statistic,

$$T_n = \sqrt{\frac{n_0 \cdot n_1}{n}} \cdot \max \left\{ \begin{array}{c} \sup_{V \in \mathcal{V}} \{Q_{n_0}(V,1) - P_{n_1}(V,1)\} \\ \sup_{V \in \mathcal{V}} \{P_{n_1}(V,0) - Q_{n_0}(V,0)\} \end{array} \right\}, \tag{26}$$

where $\mathcal{V}$ is a chosen collection of subsets in the support of Y. As T_n is non-pivotal, the author suggests to use a bootstrap method for inference that is analogous to Abadie (2002). I.e., B bootstrap samples of size n are drawn from the original data to compute T_b, the bootstrap analogue of T_n, in each sample. Then, the p-value is estimated by

$$\hat{p}_{\text{K08}} = B^{-1} \cdot \sum_{b=1}^{B} I\{T_b > T_n\}. \tag{27}$$

An open issue of the K08 test is the choice of $\mathcal{V}$. While a large number of subsets increases the chance to detect a violation and, thus, asymptotic power it may entail a high variance in finite samples. I.e., there exists a trade-off between the richness of $\mathcal{V}$ and the finite sample power. However, a method for optimally choosing the subsets in finite samples is currently not available.

In what follows we show that equivalent constraints to Proposition 1 of K08 plus two additional restrictions are obtained when adapting our framework to probability measures (including the pdf and cdf) rather than means. I.e., equivalent to equations (8) and (9) for the mean potential outcomes, the results of Horowitz and Manski (1995) imply the following bounds on the probabilities that the potential outcomes of the always takers under treatment and the never takers under non-treatment are in some subset V:

$$\begin{aligned} \frac{\Pr(Y \in V|D=1,Z=1) - (1-q)}{q} &\leq \Pr(Y(1) \in V|T=at) \leq \frac{\Pr(Y \in V|D=1,Z=1)}{q}, \\ \frac{\Pr(Y \in V|D=0,Z=0) - (1-r)}{r} &\leq \Pr(Y(0) \in V|T=nt) \leq \frac{\Pr(Y \in V|D=0,Z=0)}{r}, \end{aligned} \tag{28}$$

where q, r are again the shares of always or never takers in the respective mixed populations. Under Assumptions 1, 2J and 4 it follows that for all V in the support of Y,

$$\begin{aligned} \Pr(Y(1) \in V|T=at) &= \Pr(Y \in V|D=1,Z=0), \\ \Pr(Y(0) \in V|T=nt) &= \Pr(Y \in V|D=0,Z=1), \end{aligned}$$

and therefore,

$$\begin{aligned}\frac{\Pr(Y\in V|D=1,Z=1)-(1-q)}{q} &\leq \Pr(Y\in V|D=1,Z=0)\leq\frac{\Pr(Y\in V|D=1,Z=1)}{q},\\ \frac{\Pr(Y\in V|D=0,Z=0)-(1-r)}{r} &\leq \Pr(Y\in V|D=0,Z=1)\leq\frac{\Pr(Y\in V|D=0,Z=0)}{r}.\end{aligned} \quad (29)$$

This implies the inequality constraints

$$H_0:\begin{pmatrix}\frac{\Pr(Y\in V|D=1,Z=1)-(1-q)}{q}-\Pr(Y\in V|D=1,Z=0)\\ \Pr(Y\in V|D=1,Z=0)-\frac{\Pr(Y\in V|D=1,Z=1)}{q}\\ \frac{\Pr(Y\in V|D=0,Z=0)-(1-r)}{r}-\Pr(Y\in V|D=0,Z=1)\\ \Pr(Y\in V|D=0,Z=1)-\frac{\Pr(Y\in V|D=0,Z=0)}{r}\end{pmatrix}\equiv\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\\ \theta_4\end{pmatrix}\leq\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \quad (30)$$

which may be tested using the same methods as outlined in Section 3. Equivalent to the restrictions used in the K08 test, (30) allows us to construct tests with multiple constraints, depending on the definition of $\mathcal{V}$ and the number of the subsets V therein. E.g., it may be applied to the density function at various points in the outcome distribution. Note that the number of constraints obtained is four times the number of V.

Note that after some simple algebra (see Appendix A.3), (29) can be rewritten as

$$\begin{aligned}\Pr(Y\in V,D=1|Z=1)-(P_{1|1}-P_{1|0}) &\leq \Pr(Y\in V,D=1|Z=0)\leq\Pr(Y\in V,D=1|Z=1),\\ \Pr(Y\in V,D=0|Z=0)-(P_{1|1}-P_{1|0}) &\leq \Pr(Y\in V,D=0|Z=1)\leq\Pr(Y\in V,D=0|Z=0),\end{aligned}$$

$$(31)$$

which must hold for all V in the support of Y. I.e., (31) includes the constraints (24) discussed in K08, but in addition implies the following:

$$\begin{aligned}\Pr(Y\in V,D=1|Z=1)-\Pr(Y\in V,D=1|Z=0) &\leq (P_{1|1}-P_{1|0}),\\ \Pr(Y\in V,D=0|Z=0)-\Pr(Y\in V,D=0|Z=1) &\leq (P_{1|1}-P_{1|0}).\end{aligned} \quad (32)$$

The intuitive interpretation of this result is that the joint probability of being a complier and having an outcome that lies in subset V cannot be larger than the (unconditional) complier share in the population. To see this, define the probability measures in terms of densities, denoted by $f(y,D=d|Z=d)$. E.g., considering the first line of (32) it follows that $f(y, D = 1|Z = 1) - f(y, D = 1|Z = 0) \leq (P_{1|1} - P_{1|0})$ for all y in the support of Y, because

$$P_{1|1} - P_{1|0} = \int_Y [f(y, D = 1|Z = 1) - f(y, D = 1|Z = 0)]dy. \tag{33}$$

This more generally holds for $\sum_{V \in \mathcal{V}}[\Pr(Y \in V, D = 1|Z = 1) - \Pr(Y \in V, D = 1|Z = 0)]$ for any $\mathcal{V}$ with non-overlapping subsets V. Equivalent to (8) and (9), the bounds in (29) and (31) become wider as the complier share $(P_{1|1} - P_{1|0})$ grows.

A crucial question for the usefulness of the additional constraints is whether (32) may increase testing power compared to using (24) only. If relying on densities, the answer is negative, at least as far as asymptotic power is concerned. To see this, note that the prevalence of some y for which $f(y, D = 1|Z = 1) - f(y, D = 1|Z = 0) > (P_{1|1} - P_{1|0})$ necessarily implies that there exists some y' for which $f(y', D = 1|Z = 1) - f(y', D = 1|Z = 0) < 0$ such that (24) is violated, too, otherwise (33) cannot be satisfied. Gains in power might possibly only be realized in particular testing setups based on "rougher" definitions of V (where negative densities may be averaged out) which overlap.[11] To illustrate this by an example, assume a discrete outcome $Y \in \{5, 6, 7\}$ with

$$\begin{aligned}
&\text{(i) } f(5, D = 1|Z = 1) - f(5, D = 1|Z = 0) = 0.6,\\
&\text{(ii) } f(6, D = 1|Z = 1) - f(6, D = 1|Z = 0) = 0.3,\\
&\text{(iii) } f(7, D = 1|Z = 1) - f(7, D = 1|Z = 0) = -0.3,
\end{aligned}$$

such that $P_{1|1} - P_{1|0} = 0.6$. For $\mathcal{V} : V_1 = 5, V_2 = 6, V_3 = 7$ the probability measures correspond to the densities and the first constraint of (24) becomes binding for V_3. However, when defining $\mathcal{V} : V_1 = \{5, 6\}, V_2 = \{6, 7\}$, the negative density (iii) is averaged out by (ii) for V_2 such that (24) is satisfied. In contrast, (32) is violated for V_1, because (i)+(ii) is larger than $P_{1|1} - P_{1|0}$. Obviously, with a sample size going to infinity, using overlapping V is unnecessary and even potentially detrimental for the power of (24) for the reasons just discussed. This implies that at best, the additional constraints raise power in finite samples. This may be the case due to the use of overlapping subsets V, which needs to justified by an efficiency argument, and/or because violations of (32) occur in regions where estimation is more precise than in areas where (24) is binding.

[11] We are indebted to Toru Kitagawa for his helpful comments on this issue.

We conclude this section by demonstrating that the dominance or equality assumptions (to further increase power) discussed in Section 4 may analogously be imposed w.r.t. the probabilities that the potential outcomes of different subpopulations are situated in some subset V. E.g., one possible assumption is probability dominance of the potential outcomes of the compliers over those of the always takers under treatment:

Assumption 6J:

$\Pr(Y(1) \in V|T = c) \geq \Pr(Y(1) \in V|T = at)$ (probability dominance of compliers).

Assumption 6J states that compared to the always takers, a weakly higher share within complier outcomes is concentrated in a particular subset V of the potential outcome distribution under treatment (including its pdf). This is useful if we have prior knowledge about the concentration of compliers and always takers in some region of the potential outcome distribution. E.g., related to Assumption 6 (weak dominance of the mean potential outcomes of compliers), one might assume that the compliers are relatively more likely to be in the upper part of the distribution, i.e., for V covering some "upper" part of the support of $Y(1)$. Similar to the intuition of Assumption 6, Assumption 6J implies that $\Pr(Y(1) \in V|T = at) \leq \Pr(Y \in V|D = 1, Z = 1)$ (as the latter is a weighted average of both $\Pr(Y(1) \in V|T = c)$ and $\Pr(Y(1) \in V|T = at)$) such that the bounds in the first row of (29) tighten to

$$\frac{\Pr(Y \in V|D = 1, Z = 1) - (1 - q)}{q} \leq \Pr(Y \in V|D = 1, Z = 0) \leq \Pr(Y \in V|D = 1, Z = 1). \qquad (34)$$

Conversely, one may assume that in some subset V of $Y(1)$, the always takers are relatively more likely to occur than the compliers:

Assumption 7J:

$\Pr(Y(1) \in V|T = c) \leq \Pr(Y(1) \in V|T = at)$ (probability dominance of always takers).

In this case $\Pr(Y(1) \in V|T = at) \geq \Pr(Y \in V|D = 1, Z = 1)$ and the first row in (29) can be written as

$$\Pr(Y \in V|D = 1, Z = 1) \leq \Pr(Y \in V|D = 1, Z = 0) \leq \frac{\Pr(Y \in V|D = 1, Z = 1)}{q}. \qquad (35)$$

Finally, the strongest restriction is to assume that the same proportions within complier and always taker outcomes are concentrated in a particular subset:

Assumption 8J:
$\Pr(Y(1) \in V|T = c) = \Pr(Y(1) \in V|T = at)$ (equality of probabilities),

which implies that

$$\Pr(Y \in V|D = 1, Z = 1) = \Pr(Y \in V|D = 1, Z = 0). \tag{36}$$

Note that if Assumption 8J is assumed for any possible definition of the subsets V, this implies the equality of potential outcome distributions. Analogous assumptions may be imposed on the potential outcome distributions of the compliers and the never takers under non-treatment.

7 Simulations

We investigate the finite sample properties of the bootstrap tests based on inequality moment constraints by simulating IV models with both continuous and binary outcomes. For the continuous case, the data generating process (DGP) is the following:

$$\begin{aligned} Y &= D + \beta Z + U, \\ D &= I\{\alpha Z + \varepsilon > 0\}, \\ (U, \varepsilon) &\sim N(0,1), \quad \mathrm{Cov}(U, V) = 0.5, \quad Z, D \sim \mathrm{Bernoulli}(0.5). \end{aligned}$$

The treatment variable D is endogenous due to the correlation of the errors U and ε in the structural and the first stage equations, respectively. The first stage coefficient α determines the share of compliers in the population and, thus, the width of the bounds. We therefore expect testing power to decrease in the coefficient. In the simulations α is set to 0.2 and 0.6, which corresponds to shares of roughly 8 % and 23 %, respectively.[12] These figures are well in the range of complier proportions found in empirical applications, see for instance the examples presented in Section 8. Whereas monotonicity is satisfied by the linearity and additivity of our model, β gauges the violation of the exclusion restriction. The latter is satisfied for $\beta = 0$ and violated for any $\beta \neq 0$ implying a direct effect of Z on Y. Therefore, power should increase in

[12] The share of compliers is given by $\Phi(\alpha) - \Phi(0) = \Phi(\alpha) - 0.5$, where $\Phi(\cdot)$ is the cdf of the standard normal distribution.

the absolute value of β, as the probability that $E(Y|D=1,Z=0)$ and $E(Y|D=0,Z=1)$ fall outside the parameter bounds in the mixed populations increases in the magnitude of the direct effect. In the simulations, we set β to 0 and 1.

Table 3 reports the rejection frequencies of the various tests at the 5% level of significance for sample sizes $n=250,1000$ and 1000 simulations. The first and second columns indicate the level of α and β, respectively. The third column (st.dist1) gives $\max(\hat{\theta}_1,\hat{\theta}_2)/\text{st.dev.}(Y)$, i.e., the maximum distance between the estimate $E(Y|D=1,Z=0)$ and the bounds in the mixed population, standardized by the standard deviation of Y. A positive value implies that the point estimate of the always takers' mean potential outcome falls outside the bounds, i.e., is either smaller than the lower bound or higher than the upper bound. The fourth column (st.dist0) provides the distance parameter for the never takers: $\max(\hat{\theta}_3,\hat{\theta}_4)/\text{st.dev.}(Y)$. Columns 5 and 6 report the bias of the LATE and of the mean difference in Y of treated and non-treated individuals (which ignores endogeneity), respectively. The LATE estimator is heavily biased whenever $\beta \neq 0$ and clearly more so than taking mean differences. But even under the null with $n=250$ and $\alpha=0.2$, the estimator performs poorly, suggesting that we should be cautious when using IV estimation in small samples when the instrument is weak.

Table 3: Simulations - continuous outcome

n=250						rejection frequencies						
						mean-based tests			probability-based tests			
α	β	st.dist1	st.dist0	b.LATE	b.diff	bs	mP.p	mP.f	mP.p(2)	mP.f(2)	mP.p(4)	mP.f(4)
0.2	0.0	-0.090	-0.103	-1.440	0.796	0.007	0.017	0.007	0.031	0.013	0.028	0.004
0.6	0.0	-0.223	-0.313	-0.110	0.774	0.000	0.000	0.000	0.001	0.000	0.003	0.000
0.2	1.0	0.494	0.458	27.841	0.886	0.933	0.960	0.916	0.982	0.961	0.621	0.534
0.6	1.0	0.259	0.096	4.825	1.009	0.380	0.506	0.373	0.794	0.691	0.380	0.245
n=1000						rejection frequencies						
						mean-based tests			probability-based tests			
α	β	st.dist1	st.dist0	b.LATE	b.diff	bs	mP.p	mP.f	mP.p(2)	mP.f(2)	mP.p(4)	mP.f(4)
0.2	0.0	-0.118	-0.138	-0.282	0.795	0.001	0.003	0.001	0.005	0.003	0.015	0.003
0.6	0.0	-0.243	-0.357	-0.009	0.772	0.000	0.000	0.000	0.000	0.000	0.001	0.000
0.2	1.0	0.505	0.482	17.121	0.877	1.000	1.000	1.000	1.000	1.000	0.992	0.982
0.6	1.0	0.249	0.094	4.491	1.010	0.930	0.965	0.928	0.998	0.994	0.982	0.945

Note: Rejection frequencies at the 5% level. All tests are based on 499 bootstrap draws.

Columns 7 to 9 display the rejection frequencies for the tests based on the constraints under mean independence in (10), namely the bootstrap test with Bonferroni adjustment (bs), the Bennett (2009) test based on minimum p-values with partial (mP.p) and full (mP.f) recentering. Also columns 10 to 13 refer to versions of the Bennett (2009) test, however, using the probability-based constraints of (30) under joint independence. The partially and fully recentered statistics mP.p(2) and mP.f(2) rely on two subsets V which are obtained by cutting the distribution of Y in each simulation into two. I.e., the breakpoint between the subsets is half the difference of the maximum and minimum values of the simulated outcome $((\max(Y) - \min(Y))/2)$. By considering just two subsets we sacrifice asymptotic power, but gain finite sample power. Finally, mP.p(4) and mP.f(4) use four subsets based on the following partition: $V_1 = (-\infty, -1), V_2 = [-1, 0), V_3 = [0, 1), V_4 = [1, \infty)$. As for the K08 test, the optimal choice of the subsets V is an unsolved issue. From this perspective, the tests based on the constraints under mean independence may appear more readily applicable than those based on joint independence. Note that for all tests, the number of bootstrap draws is set to 499.

Under the null hypothesis ($\beta = 0$) the rejection frequencies of any method are quite low and clearly smaller than 5%. As expected, the empirical size decreases in α, because the bounds become wider due to a higher share of compliers, and in the sample size, which makes the estimation of $\hat{\theta}$ more precise. Under the violation of IV validity ($\beta = 1$) all tests gain power as the sample size grows and lose power as the share of compliers becomes larger. The most powerful approach in the given scenario appears to be the partially recentered minimum p-value test based on the probability constraints (mP.p(2)), which dominates any other method whenever the null does not hold. Note that also the fully recentered version (mP.f(2)) is more powerful than all tests based on the mean constraints. In contrast, mP.p(4) is less powerful than (mP.p) for $n = 250$ as well as for $\alpha = 0.2$ and $n = 1000$, whereas the converse is true for $\alpha = 0.6$ and $n = 1000$. This demonstrates that the choice of subsets affects the relative performance of the tests and that the appropriateness of a particular choice is a function of both the sample size and the features of the DGP.

Table 4: Simulations - binary outcome

n=250						rejection frequencies				
						mean-based tests			prob.-based tests	
α	β	st.dist1	st.dist0	b.LATE	b.diff	bs	mP.p	mP.f	mP.p(2)	mP.f(2)
0.2	0.0	-0.017	-0.082	-0.557	0.225	0.008	0.025	0.009	0.032	0.010
0.6	0.0	-0.082	-0.441	-0.032	0.212	0.000	0.001	0.000	0.001	0.000
0.2	1.0	0.132	0.715	6.031	0.073	0.795	0.848	0.527	0.849	0.769
0.6	1.0	0.123	0.106	0.856	0.113	0.186	0.339	0.124	0.338	0.185
n=1000						rejection frequencies				
						mean-based tests			prob.-based tests	
α	β	st.dist1	st.dist0	b.LATE	b.diff	bs	mP.p	mP.f	mP.p(2)	mP.f(2)
0.2	0.0	-0.040	-0.126	-0.088	0.226	0.003	0.009	0.002	0.010	0.004
0.6	0.0	-0.110	-0.522	0.000	0.212	0.000	0.000	0.000	0.000	0.000
0.2	1.0	0.131	0.759	3.685	0.071	1.000	1.000	1.000	1.000	1.000
0.6	1.0	0.121	0.142	0.800	0.115	0.817	0.902	0.792	0.899	0.813

Note: Rejection frequencies at the 5% level. All tests are based on 499 bootstrap draws.

Table 4 presents the rejection frequencies when the outcome is binary (probability-based tests with four subsets are therefore not considered). The DGP is identical to the first one with the exception that

$$Y = I\{D + \beta Z + U > 0\}.$$

In this case, the true treatment effect depends on the parameter α and is 0.386 for $\alpha = 0.2$ and 0.403 for $\alpha = 0.6$. Also for the binary outcome, the LATE estimator is severely biased for $\beta \neq 0$. Concerning testing, note that the mean tests need to be modified in a innocuous way to be suitable for binary outcomes.

As before, all tests are quite conservative under the null and even more so for the larger share of compliers and/or sample size. Under the violation of IV validity, the minimum p-value-based test with partial recentering is again most powerful. However, in contrast to the continuous outcome case, the mean- and probability-based versions of the test have similar power. In conclusion, mP.p(2) appears to be the preferred choice in the simulations considered as it is competitive under either definition of the outcome.

8 Application

This section presents an application to the labor market data of Card (1995), who evaluates the returns to college education based on the 1966 and 1976 waves of the U.S. National Longitudinal Survey of Young Men (NLSYM) (3,010 observations). Among others, he uses a dummy for proximity to a 4-year college in 1966 as an instrument for the potentially endogenous decision of going to college. Proximity should induce some individuals (in particular those from low income families) to strive for a college degree who would otherwise not, for instance due to costs associated with not living at home. However, the instrument may well be correlated with factors like local labor market conditions or family background (e.g. parents' education, which could shape preferences for particular residential areas) which might be related to the outcome (log of weekly earnings in 1976). This has been acknowledged by Card (1995) himself, who for this reason includes a range of control variables in his estimations. For testing, we follow K08 (who also considers this data set) and define the educational level as binary treatment which indicates one's education to be 16 years or more such that it roughly corresponds to a four year college degree. Again similar to K08, we test IV validity both in the entire sample (i.e., unconditionally) and in a subsample. The latter only includes white individuals living in an urban area not located in the south whose fathers have at least 12 years of education (554 observations), in order to control for factors that are potentially correlated with both the instrument and the outcome.

Table 5 presents the results of the tests on IV validity. The first column gives the estimated complier proportion, which is crucial for the power of the tests, the second and third columns report the standardized maximum distances $\max(\hat{\theta}_1, \hat{\theta}_2)/\text{st.dev.}(Y)$, $\max(\hat{\theta}_3, \hat{\theta}_4)/\text{st.dev.}(Y)$. The remaining columns contain the p-values of the bootstrap test with Bonferroni adjustment (bs) and the Bennett (2009) test with partial and full recentering, based on both the constraints on the means (mP.p, mP.f) and on the probabilities with 2 and 4 subsets V (mP.p(2), mP.f(2),mP.p(4), mP.f(4)), respectively, defined by an equidistant grid over the support of Y.When considering the results for the full sample, we see that the point estimate of the mean potential outcome of the never takers falls well outside its bounds. This violation is highly significant, as all tests reject the null at the 1% level. Therefore, proximity does not appear to be an unconditionally valid instrument. In the subsample, however, IV validity cannot be rejected. None of the constraints is binding and accordingly, all tests yield very large p-values. This is in line with the the results of K08 and demonstrates the importance of carefully considering potential confounders, i.e., variables that are both related with the instrument and the outcome, in empirical applications.

Table 5: Application to Card (1995) - IV validity tests

				p-values						
				mean-based tests			probability-based tests			
Sample	compliers	st.dist1	st.dist0	bs	mP.p	mP.f	mP.p(2)	mP.f(2)	mP.p(4)	mP.f(4)
full sample	6.9 %	-0.203	0.224	0.000	0.001	0.001	0.001	0.002	0.002	0.006
subsample	13.2 %	-0.419	-0.302	1.000	0.787	1.000	0.979	0.735	0.907	0.997

Note: All tests are based on 1999 bootstrap draws.

Finally, we test the mean equality constraints (20) and (21) using two sample t-tests. Table 6 reports the sample analogues of $E(Y|D = d, Z = z)$ (where $z, d \in \{0, 1\}$), denoted by $\bar{Y}_{D=d,Z=z}$, the differences (diff) $\bar{Y}_{D=1,Z=1} - \bar{Y}_{D=1,Z=0}$ and $\bar{Y}_{D=0,Z=0} - \bar{Y}_{D=0,Z=1}$, and the respective (asymptotic) p-values (p-val). Not surprisingly, the tests yield low p-values for the full sample of Card (1995), which did not even satisfy the weaker inequality constraints. In contrast, the subsample also passes the stricter difference of means tests at any conventional level of significance. This suggests that IV validity and homogeneity of the mean potential outcomes of compliers and always takers under treatment and of compliers and never takers under non-treatment hold.

Table 6: Application to Card (1995) - difference of means tests

Sample	$\bar{Y}_{D=1,Z=1}$	$\bar{Y}_{D=1,Z=0}$	diff	p-val	$\bar{Y}_{D=0,Z=0}$	$\bar{Y}_{D=0,Z=1}$	diff	p-val
full sample	6.449	6.369	0.081	0.012	6.094	6.254	-0.160	0.000
subsample	6.465	6.483	-0.018	0.806	6.348	6.390	-0.043	0.569

9 Conclusion

The LATE framework of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) with a binary treatment and instrument implies that the mean potential outcome of the always takers under treatment and that of the never takers under non-treatment can be both point identified and bounded. As the points must lie within their respective bounds, this provides four testable inequality moment constraints for instrument validity, a fact apparently neglected in the literature. For this reason we propose bootstrap tests, of which the minimum p-value-based method with partial recentering of Bennett (2009) appears to be most appropriate in terms of finite sample behavior. As a further contribution, it is shown how testing power might be increased by imposing restrictions on the order of the mean potential outcomes of different subpopulations which has also been considered in Huber and Mellace (2010), among others. Moreover, we demonstrate that IV validity and homogeneity in mean potential outcomes across particular subpopulations can be tested jointly by simple difference of means tests.

We also relate our work to the approach of Kitagawa (2008) who tests for the incidence of negative densities of complier outcomes to verify instrument validity. Interestingly, by adapting our framework to probabilities rather than means of potential outcomes, one obtains the same testable implications as Kitagawa (2008) plus two additional constraints not considered before. The latter might increase testing power in finite samples under particular conditions. Finally, we briefly investigate the finite sample properties of our tests and consider an empirical application to the labor market data of Card (1995).

The testing problem discussed in this paper raises the question of what can be done about identification if instrument validity is rejected. Obviously, the most appropriate solution would be to search for better instruments, but this may not always be feasible in practice. As an alternative, one could relax some of the IV assumptions. Then, point identification is lost, but the LATE might still be partially identified within reasonable bounds in the spirit of Manski (1989). E.g., Flores and Flores-Lagunes (2010) derive bounds on the LATE when the exclusion restriction is violated, but monotonicity of the treatment in the instrument holds, while Huber and Mellace (2010) consider the violation of monotonicity, but maintain the exclusion restriction.

A Appendix

A.1 GMM framework

In order to show that the tests described in Section 3 are valid it is sufficient to show that all the parameters involved in estimating the test statistics vector θ can be consistently estimated in a GMM framework. The latter satisfies Assumption 1 of Bennett (2009), which encounters standard conditions such as i.i.d. sampling, uniformly bounded moments of the moment functions up to a particular order, Lipschitz continuity, and an asymptotically linear representation of the testing problem. In the spirit of Lee (2009), it can be shown that $\mu_{01} \equiv E(Y|D = 1, Z = 0)$, $\mu_{10} \equiv E(Y|D = 0, Z = 1)$, $\mu_{11}^{lb} \equiv E(Y|D = 1, Z = 1, Y \leq y_q)$, $\mu_{11}^{ub} \equiv E(Y|D = 1, Z = 1, Y \geq y_{1-q})$, $\mu_{00}^{lb} \equiv E(Y|D = 0, Z = 0, Y \leq y_r)$ and $\mu_{00}^{ub} \equiv E(Y|D = 0, Z = 0, Y \geq y_{1-r})$ can be estimated as the unique solution to an just identified GMM problem, where the moment function is defined as

$$g(\vartheta, W_i) = \begin{pmatrix} (Y_i - \mu_{11}^{lb}) \cdot D_i \cdot Z_i \cdot I\{Y_i \leq y_q\} \\ (I\{Y_i > y_q\} - (1 - \frac{P_{1|0}}{1-P_{0|1}})) \cdot D_i \cdot Z_i \\ (Y_i - \mu_{11}^{ub}) \cdot D_i \cdot Z_i \cdot I\{Y_i \geq y_{1-q}\} \\ (I\{Y_i < y_{1-q}\} - (1 - \frac{P_{1|0}}{1-P_{0|1}})) \cdot D_i \cdot Z_i \\ (Y_i - \mu_{00}^{lb}) \cdot (1 - D_i) \cdot (1 - Z_i) \cdot I\{Y_i \leq y_r\} \\ (I\{Y_i > y_r\} - (1 - \frac{P_{0|1}}{1-P_{1|0}})) \cdot (1 - D_i) \cdot (1 - Z_i) \\ (Y_i - \mu_{00}^{ub}) \cdot (1 - D_i) \cdot (1 - Z_i) \cdot I\{Y_i \geq y_{1-r}\} \\ (I\{Y_i < y_{1-r}\} - (1 - \frac{P_{0|1}}{1-P_{1|0}})) \cdot (1 - D_i) \cdot (1 - Z_i) \\ (Y_i - \mu_{01}) \cdot (D_i) \cdot (1 - Z_i) \\ (Y_i - \mu_{10}) \cdot (1 - D_i) \cdot (Z_i) \\ (D_i - P_{1|0}) \cdot (1 - Z_i) \\ ((1 - D_i) - P_{0|1}) \cdot Z_i \end{pmatrix},$$

where $\vartheta = (\mu_{11}^{lb}, y_q, \mu_{11}^{ub}, y_{1-q}, \mu_{00}^{lb}, y_r, \mu_{00}^{ub}, y_{1-r}, \mu_{01}, \mu_{10}, P_{1|0}, P_{0|1})^T$ is the (12×1)-vector of parameters and $W_i = (Y_i, Z_i, D_i)$.

The GMM objective function for this just identified model is given by

$$G(\vartheta, W_i) = \sum_{i=1}^{n} g(\vartheta, W_i)^T \sum_{i=1}^{n} g(\vartheta, W_i).$$

A consistent estimator of the parameter vector ϑ can be obtained by solving the following minimization problem

$$\hat{\vartheta} = \min_{\vartheta} G(\vartheta, W_i). \tag{A.1}$$

$\sqrt{n}$-consistency and asymptotic normality of $\hat{\vartheta}$ follows directly from Propositions 2 and 3 in Lee (2009). In particular, from Theorem 7.2 of Newey and McFadden (1994) it follows that

$$\sqrt{n} \cdot (\hat{\vartheta} - \vartheta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, (H(\vartheta_0))^{-1}\Omega(\vartheta_0)(H(\vartheta_0)^T)^{-1}),$$

where $H(\vartheta_0) \equiv \nabla_{\vartheta_0} E(g(\vartheta, W_i))$ is the derivative of $E(g(\vartheta, W_i))$ at ϑ_0, the true value of ϑ, and $\Omega(\vartheta_0)$ is the asymptotic variance-covariance matrix of $g(\vartheta_0, W_i)$. Theorem 7.2 of Newey and McFadden (1994) also ensures that the regularity conditions listed in Assumption 1 of Bennett (2009) are met and that the limiting distribution of the test statistics θ exists and is continuous.

We close the discussion by providing the analytical representations of $H(\vartheta_0)$ and $\Omega(\vartheta_0)$: [13]

$$H(\vartheta_0) = \begin{pmatrix} H_{\mu_{11}^{lb}, y_q} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & H_{y_q, P_{d|z}} \\ \mathbf{0}_{2\times2} & H_{\mu_{11}^{ub}, y_{1-q}} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & H_{y_q, P_{d|z}} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & H_{\mu_{00}^{lb}, y_r} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & H_{y_r, P_{d|z}} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & H_{\mu_{00}^{ub}, y_{1-r}} & \mathbf{0}_{2\times2} & H_{y_r, P_{d|z}} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & H_{\mu_{01}, \mu_{10}} & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & H_{P_{1|0}, P_{0|1}} \end{pmatrix}$$

where $\mathbf{0}_{r\times c}$ is a $r \times c$-matrix of zeros. Denoting by f_{dz} the pdf of Y given $D = d$ and $Z = z$, the non-zero components of $H(\vartheta_0)$ are

$$H_{\mu_{11}^{lb}, y_q} = E(D \cdot Z) \cdot \begin{pmatrix} -q & (y_q - \mu_{11}^{lb}) \cdot f_{11}(y_q) \\ 0 & -f_{11}(y_q) \end{pmatrix},$$

$$H_{\mu_{11}^{ub}, y_{1-q}} = E(D \cdot Z) \cdot \begin{pmatrix} -q & -(y_{1-q} - \mu_{11}^{ub}) \cdot f_{11}(y_{1-q}) \\ 0 & f_{11}(y_{1-q}) \end{pmatrix},$$

$$H_{\mu_{00}^{lb}, y_r} = E[(1 - D) \cdot (1 - Z)] \cdot \begin{pmatrix} -r & (y_r - \mu_{00}^{lb}) \cdot f_{00}(y_r) \\ 0 & -f_{00}(y_r) \end{pmatrix},$$

$$H_{\mu_{00}^{ub}, y_{1-r}} = E[(1 - D) \cdot (1 - Z)] \cdot \begin{pmatrix} -r & -(y_{1-r} - \mu_{00}^{ub}) \cdot f_{00}(y_{1-r}) \\ 0 & f_{00}(y_{1-r}) \end{pmatrix},$$

[13] As suggested by Newey and McFadden (1994) $\Omega(\vartheta_0)$ can be consistently estimated by

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (g(\hat{\vartheta}, W_i) \cdot g(\hat{\vartheta}, W_i)^T).$$

Alternatively one can take the sample counterpart of the closed form solution of $\Omega(\vartheta_0)$ provided belove.

$$H_{\mu_{01},\mu_{10}} = \begin{pmatrix} -E[(D)\cdot(1-Z)] & 0 \\ 0 & -E[(1-D)\cdot(Z)] \end{pmatrix},$$

$$H_{P_{1|0},P_{0|1}} = \begin{pmatrix} -E(1-Z) & 0 \\ 0 & -E(Z) \end{pmatrix},$$

$$H_{y_q,P_{d|z}} = E(D\cdot Z)\cdot \begin{pmatrix} 0 & 0 \\ \frac{1}{1-P_{0|1}} & \frac{P_{1|0}}{(1-P_{0|1})^2} \end{pmatrix},$$

$$H_{y_r,P_{d|z}} = E[(1-D)\cdot(1-Z)]\cdot \begin{pmatrix} 0 & 0 \\ \frac{P_{0|1}}{(1-P_{1|0})^2} & \frac{1}{1-P_{1|0}} \end{pmatrix}.$$

$\Omega(\vartheta_0)$ is given by the block diagonal variance-covariance matrix

$$\Omega(\vartheta_0) \quad = \quad \begin{pmatrix} \Omega_{11} & \Omega_{12} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} \\ \Omega_{12} & \Omega_{13} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \Omega_{21} & \Omega_{22} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \Omega_{22} & \Omega_{23} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \Omega_{31} & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \Omega_{32} \end{pmatrix},$$

where the non-zero elements are

$$\Omega_{11} = \begin{pmatrix} \int_{-\infty}^{y_q} (y-\mu_{11}^{lb})^2\cdot f_{11}\cdot dy & 0 \\ 0 & q\cdot(1-q) \end{pmatrix}\cdot E(D\cdot Z),$$

$$\Omega_{12} = \begin{pmatrix} I\{y_{1-q} < y_q\}\cdot \int_{y_{1-q}}^{y_q} (y-\mu_{11}^{lb})\cdot(y-\mu_{11}^{ub})\cdot f_{11}\cdot dy & \int_{-\infty}^{\min(y_q,y_{1-q})} (y-\mu_{11}^{lb})\cdot f_{11}\cdot dy \\ \int_{\max(y_q,y_{1-q})}^{\infty} (y-\mu_{11}^{ub})\cdot f_{11}\cdot dy & \int_{-\infty}^{\infty} I(y>y_q)\cdot I(y<y_{1-q})\cdot dy-(1-q)^2 \end{pmatrix}\cdot E(D\cdot Z),$$

$$\Omega_{13} = \begin{pmatrix} \int_{y_{1-q}}^{\infty} (y-\mu_{11}^{ub})^2\cdot f_{11}\cdot dy & 0 \\ 0 & q\cdot(1-q) \end{pmatrix}\cdot E(D\cdot Z),$$

$$\Omega_{21} = \begin{pmatrix} \int_{-\infty}^{y_r} (y-\mu_{00}^{lb})^2\cdot f_{00}\cdot dy & 0 \\ 0 & r\cdot(1-r) \end{pmatrix}\cdot E[(1-D)\cdot(1-Z)],$$

$$\Omega_{22} = \begin{pmatrix} I\{y_{1-r} < y_r\} \cdot \int_{y_{1-r}}^{y_r} (y - \mu_{00}^{lb}) \cdot (y - \mu_{00}^{ub}) \cdot f_{00} \cdot dy & \int_{-\infty}^{\min(y_r, y_{1-r})} (y - \mu_{00}^{lb}) \cdot f_{00} \cdot dy \\ \int_{\max(y_r, y_{1-r})}^{\infty} (y - \mu_{00}^{ub}) \cdot f_{00} \cdot dy & \int_{-\infty}^{\infty} I(y > y_r) \cdot I(y < y_{1-r}) dy - (1-r)^2 \end{pmatrix} \cdot E[(1-D) \cdot (1-Z)],$$

$$\Omega_{23} = \begin{pmatrix} \int_{y_{1-r}}^{\infty} (y - \mu_{00}^{ub})^2 \cdot f_{00} \cdot dy & 0 \\ 0 & r \cdot (1-r) \end{pmatrix} \cdot E[(1-D) \cdot (1-Z)],$$

$$\Omega_{31} = \begin{pmatrix} \int_{-\infty}^{\infty} (y - \mu_{01})^2 \cdot f_{10} \cdot dy \cdot E[D \cdot (1-Z)] & 0 \\ 0 & \int_{-\infty}^{\infty} (y - \mu_{10})^2 \cdot f_{01} \cdot dy \cdot E[(1-D) \cdot Z] \end{pmatrix},$$

$$\Omega_{32} = \begin{pmatrix} P_{1|0} \cdot P_{0|0} \cdot E(1-Z) & 0 \\ 0 & P_{0|1} \cdot P_{1|1} \cdot E(Z) \end{pmatrix}.$$

A consistent estimator of the asymptotic variance of $\hat{\vartheta}$ is given by $n^{-1} \cdot (H(\hat{\vartheta}))^{-1} \Omega(\hat{\vartheta}) (H(\hat{\vartheta})^T)^{-1}$. Then, it is straightforward to construct an estimator for the variance of the test statistic $\hat{\theta}$, see Appendix A.2.

A.2 Wolak (1989, 1991) test

In order to describe the Wolak test it is useful to rewrite the test statistic θ as $\theta = C\vartheta$, where C is

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

We would like to test the null hypothesis $H_0 : C\vartheta \leq 0$ against $H_1 : C\vartheta > 0$. Theorem 7.2 of Newey and McFadden (1994) implies that ϑ satisfies Assumption 1 of Donald and Hsu (2010). Therefore, it follows that

$$\sqrt{n} \cdot (C\hat{\vartheta} - C\vartheta_0) \xrightarrow{d} \mathcal{N}(0, J(\vartheta_0)),$$

where $J(\vartheta_0) \equiv C(H(\vartheta_0))^{-1} \Omega(\vartheta_0) (H(\vartheta_0)^T)^{-1} C^T$. We consider the following test statistic proposed by Wolak (1989a):[14]

$$V_n = \min_{\{\vartheta | C\vartheta \leq 0\}} n \cdot (C\hat{\vartheta} - C\vartheta)^T J(\hat{\vartheta})^{-1} (C\hat{\vartheta} - C\vartheta)$$

[14] Wolak (1989a) proposes three asymptotically equivalent test statistics.

Let ϑ^* be the LFC (i.e., the parameter configuration for which the null is rejected with the lowest probability). As Wolak (1991) shows, $\vartheta^* \in \mathcal{B}$ where $\mathcal{B}$ is the set of ϑ which satisfy the null hypothesis such that there are at least two constraints that hold as equality. Moreover, for any $\vartheta \in \mathcal{B}$,

$$\lim_{n\to\infty} \Pr(V_n \geq c|\vartheta) = \sum_{i=0}^{m} \Pr(\chi_i^2 \geq c) \cdot \psi(m, m-i, C_m(H(\vartheta_0))^{-1}\Omega(\vartheta_0)(H(\vartheta_0)^T)^{-1}C_m^T),$$

where m is the number of constraints that hold as equality, C_m is the corresponding sub-matrix of C and $\psi(m, m-i, \Sigma)$ is a weighting function. Let s be a m-dimensional vector of normally distributed elements with mean zero and covariance Σ and $\tilde{s}$ be the solution to

$$\arg\min_{\gamma\geq 0}(s-\gamma)^T\Sigma(s-\gamma).$$

Then, $\psi(m, m-i, \Sigma)$ is the probability that $\tilde{s}$ has exactly $m-i$ positive elements. Therefore, one has to solve

$$\sum_{i=0}^{m} \Pr(\chi_i^2 \geq c(\vartheta, \alpha)) \cdot \psi(m, m-i, C_m(H(\vartheta_0))^{-1}\Omega(\vartheta_0)(H(\vartheta_0)^T)^{-1}C_m^T) = \alpha, \tag{A.2}$$

for a given confidence level α and for each $\vartheta \in \mathcal{B}$ in order to construct a test with asymptotically exact size. The critical value is

$$c(\alpha) = \max_{\vartheta\in\mathcal{B}} c(\vartheta, \alpha).$$

Since we have four inequality constraints, the number of configurations of ϑ such that $\vartheta \in \mathcal{B}$ is 11 (all possible combinations of the four constraints such that at least two of them hold as equality). This implies that (A.2) must be evaluated 11 times in order to obtain the LFC.

A.3 Proof of equation (31)

First of all, notice that

$$\begin{aligned}\frac{\Pr(Y \in V|D=1, Z=1) - (1-q)}{q} &= \frac{\Pr(Y \in V, D=1|Z=1)}{q \cdot \Pr(D=1|Z=1)} - \frac{(1-q)}{q} \\ &= \frac{\Pr(Y \in V, D=1|Z=1)}{P_{1|0}} - \frac{P_{1|1} - P_{1|0}}{P_{1|0}}.\end{aligned} \tag{A.3}$$

Moreover

$$\begin{aligned}\frac{\Pr(Y \in V|D=1, Z=1)}{q} &= \frac{\Pr(Y \in V, D=1|Z=1)}{q \cdot \Pr(D=1|Z=1)} \\ &= \frac{\Pr(Y \in V, D=1|Z=1)}{P_{1|0}}.\end{aligned} \tag{A.4}$$

and

$$\Pr(Y \in V|D = 1, Z = 0) \quad = \quad \frac{\Pr(Y \in V, D = 1|Z = 0)}{P_{1|0}}. \tag{A.5}$$

Therefore the first line of 29 can be written as

$$\begin{aligned} \frac{\Pr(Y \in V, D = 1|Z = 1)}{P_{1|0}} - \frac{P_{1|1} - P_{1|0}}{P_{1|0}} &\leq \frac{\Pr(Y \in V, D = 1|Z = 0)}{P_{1|0}} \leq \frac{\Pr(Y \in V, D = 1|Z = 1)}{P_{1|0}}, \\ &\Rightarrow \\ \Pr(Y \in V, D = 1|Z = 1) - (P_{1|1} - P_{1|0}) &\leq \Pr(Y \in V, D = 1|Z = 0) \leq \Pr(Y \in V, D = 1|Z = 1). \end{aligned} \tag{A.6}$$

An equivalent argument can be used to obtain the second line.

A.4 Discrete outcomes

In the main text we have shown how to bound $E(Y(1)|T = at)$ and $E(Y(1)|T = nt)$ when the outcome is continuous, here we will consider the case of discrete outcomes. For the sake of brevity we will just consider the always takers' bounds, as an analogous argument applies to the never takers. If the outcome has a discrete mass distribution, the shares of individuals for which $Y \leq y_q|D = 1, Z = 1$ and $Y \geq y_{1-q}|D = 1, Z = 1$ may differ from q. This is due to the presence of ties in the outcome, i.e. the occurrence of mass points with equal outcome values, which entails a non-unique quantile function such that a particular outcome value is observed at several ranks. Therefore, the trimming rule based on the quantile functions described above does not provide the sharp bounds of $E(Y(1)|T = at)$ in general. In the presence of mass points in the outcome we have to replace the non-unique quantile functions, which give equal ranks to all ties, by modified versions which account for ties in the trimming rule.

To this end, we denote by $n_{1,1}$ the number of observations for which $Z = 1$ and $D = 1$ and by $Y_{1,1}$ the outcome variable in the respective observed group. Let $Y_{1,1}^{(1)}, \ldots, Y_{1,1}^{(n_{1,1})}$ be the order statistic of $Y_{1,1}$ such that $Y_{1,1}^{(1)} \leq Y_{1,1}^{(2)} \ldots \leq Y_{1,1}^{(n_{1,1})}$. Note that this implies that for all observations $Z = 1$ and $D = 1$ with the same outcome value the order may be randomly assigned as it does not play any role for the results as it will become apparent further below. Since the number of always takers in the subsample with $Z = 1$ and $D = 1$ is exactly $q \cdot n_{1,1}$, the lower bound of $E(Y(1)|T = at)$ is obtained by averaging over the $q \cdot n_{1,1}$ observations of $Y_{1,1}^{(1)}, \ldots, Y_{1,1}^{(n_{1,1})}$ at the lowest ranks, which we denote as $Y_{1,1}^{\min} = (Y_{1,1}^{(1)}, \ldots, Y_{1,1}^{(q \cdot n_{1,1})})$. For the upper bound of $E(Y(1)|T = at)$, we average over the $q \cdot n_{1,1}$ observations of $Y_{1,1}^{(1)}, \ldots, Y_{1,1}^{(n_{1,1})}$ at the highest ranks, which we denote as $Y_{1,1}^{\max} = (Y_{1,1}^{(q \cdot n_{1,1})}, \ldots, Y_{1,1}^{(n_{1,1})})$.

If $q \cdot n_{1,1}$ is not an integer one can replace it with its integer part denoted by $\widetilde{q \cdot n_{1,1}}$. Note that we take the integer part of $q \cdot n_{1,1}$ because the number of always takers cannot be bigger than $q \cdot n_{1,1}$. It easy to see that the share of individuals of $Y_{1,1}$ in $Y_{1,1}^{\min}$ and $Y_{1,1}^{\max}$, which is given by $\tilde{q} = \frac{\widetilde{q \cdot n_{1,1}}}{n_{1,1}}$, is approximately q (and is exactly q if $q \cdot n_{1,1}$ is an integer). Notice that $q - \tilde{q} \to 0$ as $n_{1,1} \to \infty$.

To give an example assume that $q = 0.6$ and $n_{1,1} = 11$. Moreover, suppose that in the subsample with $Z = 1$ and $D = 1$ we have five observations for which $Y = 0$, three observations with $Y = 1$, while the remaining three take the values 1.1, 1.2, and 1.3. Therefore, the order statistic of $Y_{1,1}$ is given by $(0, 0, 0, 0, 0, 1, 1, 1, 1.1, 1.2, 1.3)'$. Since the always takers are $0.6 \times 11 = 6.6$ in the subsample where $Z = 1$ and $D = 1$, we approximate $q \cdot n_{1,1} = 6.6$ by its integer part $\widetilde{q \cdot n_{1,1}} = 6$. Then, the lower bound of $E(Y(1)|T = at)$ is obtained by choosing $Y_{1,1}^{\min} = (0, 0, 0, 0, 0, 1)'$ and the upper bound is relies on $Y_{1,1}^{\max} = (1, 1, 1.1, 1.2, 1.3)'$. In this example, $\tilde{q} = \frac{6}{11} = 0.5455$ which is not too far from $q = 0.6$ even though $n_{1,1}$ is small. If we applied the standard trimming rule, we would get $y_q = 1$ and $y_{1-q} = 0$. Therefore, the share of observations for which $Y \leq y_q | D = 1, Z = 1$ is $\frac{8}{11} = 0.7273$, while $Y | D = 1, Z = 1 \geq y_{1-q}$ is $\frac{11}{11} = 1$. This demonstrates that the bounds of $E(Y(1)|T = at)$ based on the standard trimming rule are not valid.

References

Abadie, A. (2002): “Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models,” *Journal of the American Statistical Association*, 97, 284–292.

Andrews, D. W., and P. Guggenberger (2007): “The Limit of Finite-Sample Size and a Problem with Subsampling,” *Cowles Foundation Discussion Paper 1605R.*

Andrews, D. W., and P. Jia (2008): “Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure,” *Cowles Foundation Discussion Paper 1676.*

Andrews, D. W. K., and G. Soares (2010): “Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,” *Econometrica*, 78, 119–157.

Angrist, J. (1990): “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records,” *American Economic Review*, 80, 313–336.

Angrist, J., and W. Evans (1998): “Children and their parents labor supply: Evidence from exogeneous variation in family size,” *American Economic Review*, 88, 450–477.

Angrist, J., G. Imbens, and D. Rubin (1996): “Identification of Causal Effects using Instrumental Variables,” *Journal of American Statistical Association*, 91, 444–472 (with discussion).

Angrist, J., and A. Krueger (1991): “Does Compulsory School Attendance Affect Schooling and Earnings?,” *Quarterly Journal of Economics*, 106, 979–1014.

Balke, A., and J. Pearl (1997): “Bounds on Treatment Effects From Studies With Imperfect Compliance,” *Journal of the American Statistical Association*, 92, pp. 1171–1176.

Bennett, C. J. (2009): “Consistent and Asymptotically Unbiased MinP Tests of Multiple Inequality Moment Restrictions,” Working Paper 09-W08, Department of Economics, Vanderbilt University.

Beran, R. (1988): “Prepivoting test statistics: A bootstrap view of asymptotic refinements,” *Journal of the American Statistical Association*, 83, 687–697.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): “Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak,” *Journal of the American Statistical Association*, 90, 443–450.

Cameron, A., and P. Trivedi (2005): *Microeconometrics.* Cambridge Univ. Press, Cambridge.

Card, D. (1995): “Using Geographic Variation in College Proximity to Estimate the Return to Schooling,” in *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. Christofides, E. Grant, and R. Swidinsky, pp. 201–222. University of Toronto Press, Toronto.

——— (1999): “The Causal Effect of Education on Earnings,” in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1802–1863. North-Holland, Amsterdam.

Chen, L.-Y., and J. Szroeter (2009): “Hypothesis testing of multiple inequalities: the method of constraint chaining,” *CeMMAP working paper 13/09.*

Chernozhukov, V., H. Hong, and E. Tamer (2007): “Estimation and Confidence Regions for Parameter Sets in Econometric Models1,” *Econometrica*, 75, 1243–1284.

Donald, S. G., and Y.-C. Hsu (2010): “A New Test for Linear Inequality Constraints When the Variance Covariance Matrix Depends on the Unknown Parameters,” Working paper, Department of Economics, Vanderbilt University.

Efron, B. (1979): “Bootstrap Methods: Another Look at the Jackknife,” *The Annals of Statistics*, 7, 1–26.

Fan, Y., and S. Park (2007): “Confidence Sets for Some Partially Identified Parameters,” *mimeo.*

Flores, C. A., and A. Flores-Lagunes (2010): “Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects,” *mimeo, University of Florida.*

Frölich, M. (2007): “Nonparametric IV Estimation of Local Average Treatment Effects with Covariates,” *Journal of Econometrics*, 139, 35–75.

Frölich, M., and B. Melly (2008): “Unconditional Quantile Treatment Effects under Endogeneity,” *IZA DP No. 3288.*

Godfrey, L. G. (2005): “Controlling the overall significance level of a battery of least squares diagnostic tests,” *Oxford Bulletin of Economics and Statistics*, 67, 263–279.

Greene, W. H. (2008): *Econometric analysis.* Pearson Prentice Hall, Upper Saddle River, NJ, 6 edn.

Guggenberger, P., J. Hahn, and K. Kim (2008): “Specification testing under moment inequalities,” *Economics Letters*, 99, 375–378.

Hansen, P. R. (2005): “A Test for Superior Predictive Ability,” *Journal of Business & Economic Statistics*, 23, 365–380.

Horowitz, J. L., and C. F. Manski (1995): “Identification and Robustness with Contaminated and Corrupted Data,” *Econometrica*, 63, 281–302.

Huber, M., and G. Mellace (2010): “Sharp IV bounds on average treatment effects under endogeneity and noncompliance,” *University of St Gallen, Dept. of Economics Discussion Paper no. 2010-31.*

Imbens, G. W., and J. Angrist (1994): “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62, 467–475.

Imbens, G. W., and D. Rubin (1997): “Estimating outcome distributions for compliers in instrumental variables models,” *Review of Economic Studies*, 64, 555–574.

Kitagawa, T. (2008): “A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model,” *mimeo.*

Kodde, D. A., and F. C. Palm (1986): “Wald Criteria for Jointly Testing Equality and Inequality Restrictions,” *Econometrica*, 54, 1243–1248.

Kudo, A. (1963): “A multivariate analogue of the one-sided test,” *Biometrika*, 50, 403–418.

Lee, D. S. (2009): “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects,” *Review of Economic Studies*, 76, 1071–1102.

Linton, O. B., K. Song, and Y.-J. Whang (2008): “Bootstrap tests of stochastic dominance with asymptotic similarity on the boundary,” *CeMMAP working paper 8/08.*

MacKinnon, J. G. (2007): “Bootstrap Hypothesis Testing,” Working Paper 1127, Queen’s University, Department of Economics.

Manski, C. F. (1989): “Anatomy of the Selection Problem,” *Journal of Human Resources*, 24, 343–360.

Newey, W. K., and D. McFadden (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden. Elsevier, Amsterdam.

Peracchi, F. (2001): *Econometrics.* Wiley, New York.

Perlman, M. D. (1969): “One-Sided Testing Problems in Multivariate Analysis,” *The Annals of Mathematical Statistics*, 40, 549–567.

Richardson, T. S., R. J. Evans, and J. M. Robins (2011): *Bayesian Statistics 9*chap. Transparent parameterizations of models for potential outcomes. Oxford University Press.

Romano, J., A. Shaikh, and M. Wolf (2008): "Control of the false discovery rate under dependence using the bootstrap and subsampling," *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 17, 417–442.

Rosen, A. M. (2008): "Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities," *Journal of Econometrics*, 146, 107–117.

Rosenzweig, M. R., and K. I. Wolpin (2000): "Natural "Natural Experiments" in Economics," *Journal of Economic Literature*, 38, 827–874.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1990): "Formal Modes of Statistical Inference For Causal Effects," *Journal of Statistical Planning and Inference*, 25, 279–292.

Sargan, J. D. (1958): "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 26, 393–415.

Wolak, F. A. (1987): "An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model," *Journal of the American Statistical Association*, 82, 782–793.

——— (1989a): "Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models," *Econometric Theory*, 5, 1–35.

——— (1989b): "Testing inequality constraints in linear econometric models," *Journal of Econometrics*, 41, 205–235.

——— (1991): "The Local Nature of Hypothesis Tests Involving Inequality Constraints in Nonlinear Models," *Econometrica*, 59, 981–995.

Wooldridge, J. M. (2002): *Econometric analysis of cross section and panel data*. MIT Press, Cambridge and London.

Zhang, J., and D. B. Rubin (2003): "Estimation of causal effects via principal stratification when some outcome are truncated by death," *Journal of Educational and Behavioral Statistics*, 28, 353–368.