Testing instrument validity for LATE identification based on inequality moment constraints

Martin Huber and Giovanni Mellace

University of St. Gallen, Dept. of Economics

Abstract–We derive testable implications of instrument validity in just identified treatment effect models with endogeneity and consider several tests. The identifying assumptions of the local average treatment effect allow us to both point identify and bound the mean potential outcomes (i) of the always takers under treatment and (ii) of the never takers under non-treatment. The point identified means must lie within their respective bounds, which provides us with four testable inequality moment constraints. Finally, we adapt our testing framework to the identification of distributional features. A brief simulation study and an application to labor market data are also provided.

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

JEL classification: C12, C15, C21, C26.

We have benefited from comments by Alberto Abadie, Clément De Chaisemartin, Mario Fiorini, Stefan Hoderlein, Bo Honoré, Guido Imbens, Toru Kitagawa, Michael Lechner, Jeff Smith, Anthony Strittmatter, Gautam Tripathi, seminar/conference participants in Bludenz (internal research seminar, June 2011), Harvard (IZA Conference on Labor Market Policy Evaluation, October 2011), Zurich (Swiss Society of Economics and Statistics, April 2012), and Chicago (SOLE, May 2012), the editor Mark Watson and three anonymous referees. 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 heterogeneous treatment effect models with a binary treatment (which allow for effect heterogeneity across individuals), an instrument is valid if (i) the potential outcomes are mean independent of the instrument, (ii) the potential treatment states are not confounded by the instrument, 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]

[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

IV estimation is a corner stone of empirical research. Taking the estimation of the returns to education as an example, a range of instruments has been suggested to control for the endogenous choice of schooling. E.g., 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. As a second example, Card (1995) exploits geographical proximity to college (which should affect the cost of college education) as instrument for going to college. However, most instruments are far from being undisputed[3] because arguments in favor or against an instrument being valid 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.[4]

Kitagawa (2008), henceforth K08, provides the first formal test for just identified models with a binary treatment and 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

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 heterogeneous treatment effect models and show that the LATE assumptions have testable implications for IV validity.

[3] See for instance Bound, Jaeger and Baker (1995), who contest the validity of quarter of birth instruments and 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).

[4] IV tests are, however, available for overidentified models where the number of instruments exceeds the number of endogenous regressors, see for instance Sargan (1958).

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 in the mixed population of compliers and never takers. Naturally, the density of the complier outcomes under treatment and non-treatment must not be negative, which is a testable implication already observed by Balke and Pearl (1997) (for the binary outcome case).[5] K08 therefore tests whether negative densities occur in subsets of the outcome distribution and uses a bootstrap method for inference.[6]

The first contribution of this paper is the proposition of an alternative testing approach that is based on mean potential outcomes rather than densities. In the case of a binary instrument, the underlying intuition is as follows: If an instrument is valid (i.e., satisfies conditions (i) to (iii) outlined in the first paragraph), 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 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 constraint is violated, the instrument is either confounded, or it has a direct

[5] Also Heckman and Vytlacil (2005) derive a testable constraint (under possibly continuous instruments and outcomes) in their Appendix A that is equivalent to the non-negativity of complier densities under discrete instruments.

[6] In contrast to K08, Angrist and Imbens (1995) consider a multi-valued treatment and discuss a testable implication for monotonicity alone: One conditional distribution of the treatment given the instrument has to stochastically dominate the other.

effect on the mean potential outcome of the always takers, or the treatment is not monotonic in the instrument, or several problems occur at the same time. 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 of the never takers under non-treatment. For testing these constraints, we consider three different methods: A simple bootstrap test with Bonferroni adjustment, the minimum p-value test of Bennett (2009), and the smoothed indicator-based method of Chen and Szroeter (2012). As in K08, we test 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, more violations of IV validity may be detected as the bounds shrink or, put differently, as the compliers' share becomes relatively smaller to the fractions of always takers and never takers, respectively.

We therefore also demonstrate how the width of the bounds can be tightened further to increase testing power (in the sense of finding more violations of the IV assumptions) 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.

Finally, we extend 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 with compliers and always or never takers. Also the lower bounds provide two testable restrictions saying that under the null, the joint probability of being a complier and having an outcome that falls into a subset of the support must never be larger than the (unconditional) complier share in the population.[7] Similar to the tests based on mean independence, we also demonstrate how testing 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 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 detect more violations of IV validity. A generalization of the testable implications derived for the binary instrument case to discrete instruments with multiple values is provided in

[7] Note, however, that the latter restrictions are redundant if non-overlapping subsets of the outcome distribution which jointly cover the entire outcome support are used for testing as in K08. Then, they are implicitly accounted for by the non-negative complier density constraints of K08 and Balke and Pearl (1997).

Section 5. Section 6 discusses testing under the stronger joint independence assumption. Simulation results are presented in Section 7. In Section 8, we apply our methods to labor market data from 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) 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 heterogeneous 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, $Y_i^{d,z} = \varphi(d, z, U_i)$.

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
a	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
n	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 1:

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

Under Assumption 1, the share of any type conditional on the instrument is equal to its unconditional proportion in the entire population. Let $\pi_t \equiv \Pr(T = t)$, $t \in \{a, c, n\}$, 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 1 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
$P_{1\|1} \equiv \Pr(D = 1\|Z = 1)$	$\pi_a + \pi_c$
$P_{0\|1} \equiv \Pr(D = 0\|Z = 1)$	$\pi_d + \pi_n$
$P_{1\|0} \equiv \Pr(D = 1\|Z = 0)$	$\pi_a + \pi_d$
$P_{0\|0} \equiv \Pr(D = 0\|Z = 0)$	$\pi_c + \pi_n$

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 conditional on the instrument (denoted by $E(Y^{d,z}|T=t,Z=z)$), where the weights depend on the relative proportions:

$$\begin{aligned}
E(Y|D=1,Z=1) &= \frac{\pi_a}{\pi_a+\pi_c}\cdot E(Y^{1,1}|T=a,Z=1) \\
&+ \frac{\pi_c}{\pi_a+\pi_c}\cdot E(Y^{1,1}|T=c,Z=1), \qquad (1) \\
E(Y|D=1,Z=0) &= \frac{\pi_a}{\pi_a+\pi_d}\cdot E(Y^{1,0}|T=a,Z=0) \\
&+ \frac{\pi_d}{\pi_a+\pi_d}\cdot E(Y^{1,0}|T=d,Z=0), \qquad (2) \\
E(Y|D=0,Z=0) &= \frac{\pi_c}{\pi_n+\pi_c}\cdot E(Y^{0,0}|T=c,Z=0) \\
&+ \frac{\pi_n}{\pi_n+\pi_c}\cdot E(Y^{0,0}|T=n,Z=0), \qquad (3) \\
E(Y|D=0,Z=1) &= \frac{\pi_d}{\pi_n+\pi_d}\cdot E(Y^{0,1}|T=d,Z=1) \\
&+ \frac{\pi_n}{\pi_n+\pi_d}\cdot E(Y^{0,1}|T=n,Z=1). \qquad (4)
\end{aligned}$$

From Table 2 and expressions (1) to (4) it becomes obvious that further assumptions are necessary to identify the LATE. Our second assumption is a mean exclusion restriction, which requires that the instrument does not exhibit an effect on the mean potential outcomes within any subpopulation (however, it may affect higher moments):

Assumption 2:

$E(Y^{d,1}|T=t,Z=1)=E(Y^{d,0}|T=t,Z=0)=E(Y^{d}|T=t)$ for $d\in\{0,1\}$ and $t \in \{a,c,d,n\}$ (mean exclusion restriction),

where the last equality makes explicit that the mean potential outcomes are not a function of the instrument. It follows that

$$E(Y^{1,1}|T=a,Z=1)=E(Y^{1,0}|T=a,Z=0)=E(Y^{1}|T=a)$$

and

$$E(Y^{0,1}|T=n,Z=1)=E(Y^{0,0}|T=n,Z=0)=E(Y^{0}|T=n),$$

which provides the base for the testable implications outlined further below. Alternatively to Assumptions 1 and 2, 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 3:

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

Assumption 4:

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

Assumption 3 states that the potential treatment state of any individual does not decrease in the instrument. This rules out the existence of defiers (type d). We shall refer to Z as a valid instrument if it satisfies Assumptions 1-3. However, LATE identification in addition requires Z to be a relevant instrument in the sense that the treatment states of (at least) some individuals react to a switch in the instrument. This is implied by Assumption 4, which postulates the existence of compliers.

As defiers do not exist, the proportions of the remaining types are identified by $P_{0|1} = \pi_n$, $P_{1|0} = \pi_a$, $P_{1|1} - P_{1|0} = P_{0|0} - P_{0|1} = \pi_c$. 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 $E(Y|D = 1, Z = 0) = E(Y^1|T = a)$ under Assumptions 1 to 3, and (4) becomes $E(Y|D = 0, Z = 1) = E(Y^0|T = n)$. Plugging $E(Y|D = 1, Z = 0)$ and $E(Y|D = 0, Z = 1)$ into (1) and (3) allows solving the equations for the mean potential outcomes of the compliers under treatment and non-treatment.[8] The difference of the latter gives the LATE and simplifies to the well known probability limit of the Wald estimator, which corresponds to the intention to treat effect divided by the share of compliers, see Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996): $E(Y^1 - Y^0|T = c) = \frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}$.

Assumptions 1-4 not only identify the LATE and the mean potential outcomes of the compliers, they also provide testable implications for IV validity based on deriving bounds on 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 $E(Y|D = 1, Z = 0) = E(Y^1|T = a)$ must lie within these bounds, otherwise either Z has a direct effect on the mean of Y, or the instrument is confounded, 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 online appendix A.4 shows how the fol-

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

lowing intuition and the tests 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): $q = \frac{\pi_a}{\pi_a + \pi_c} = \frac{P_{1|0}}{P_{1|1}}$. 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

$$E(Y|D = 1, Z = 1, Y \leq y_q) \leq E(Y|D = 1, Z = 0) \leq E(Y|D = 1, Z = 1, Y \geq y_{1-q}). \quad (5)$$

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 $r = \frac{\pi_n}{\pi_n + \pi_c} = \frac{P_{0|1}}{P_{0|0}}$, 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 $E(Y|D = 0, Z = 1) = E(Y^0|T = n)$, such that the IV assumptions require that

$$E(Y|D = 0, Z = 0, Y \leq y_r) \leq E(Y|D = 0, Z = 1) \leq E(Y|D = 0, Z = 0, Y \geq y_{1-r}). \quad (6)$$

Note that under one-sided non-compliance, only one of (5) and (6) 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

($\pi_a = 0$) 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 (6) is still useful to test the exclusion restriction on the never takers.

3 Testing

Expressions (5) and (6) provide us with four testable inequality moment constraints.[9] Under IV validity it must hold that

$$H_0 : \begin{pmatrix} E(Y|D=1,Z=1,Y\leq y_q) - E(Y|D=1,Z=0) \\ E(Y|D=1,Z=0) - E(Y|D=1,Z=1,Y\geq y_{1-q}) \\ E(Y|D=0,Z=0,Y\leq y_r) - E(Y|D=0,Z=1) \\ E(Y|D=0,Z=1) - E(Y|D=0,Z=0,Y\geq y_{1-r}) \end{pmatrix} \equiv \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \qquad (7)$$

Under a violation of IV validity 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

[9] Note that expressions (5) and (6) hold under Assumptions 1-3 alone, i.e., the existence of compliers (Assumption 4) 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.

constraints is the best one can get. It is obvious that the tests can asymptotically reject more violations of IV validity as compliers' share decreases, because the bounds on the mean outcomes of the always and never takers become tighter.

We test (7) using three different methods: A simple bootstrap test with Bonferroni adjustment, the minimum p-value test of Bennett (2009), and the smoothed indicator-based method of Chen and Szroeter (2012). Note that the parameters involved in θ can be estimated in a standard GMM framework as described in online appendix A.1, which satisfies the regularity conditions required for bootstrapping, see Horowitz (2001), as well as those in Assumption 1 of Bennett (2009). Our testing problem also meets the regularity conditions [D1], [D3], and [D4] in Chen and Szroeter (2012), whereas [D2] needs to be verified as outlined in their Appendix C.

Starting with the bootstrap test with Bonferroni adjustment, 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. In each bootstrap sample, the recentered parameter vector $\tilde{\theta}_b = \hat{\theta}_b - \hat{\theta}$ is computed. The vector of p-values $P_{\hat{\theta}}$ is estimated as the share of bootstrap replications in which studentized versions of the recentered parameters are larger than the studentized estimates in the original sample: $P_{\hat{\theta}} = B^{-1} \sum_{b=1}^{B} I \left\{ \frac{\tilde{\theta}_b}{\hat{\sigma}} > \frac{\hat{\theta}}{\hat{\sigma}} \right\}$, where $I\{\cdot\}$ denotes the indicator function and $\hat{\sigma}$ is an estimate of the standard error of $\hat{\theta}$. While these p-values are consistent for assessing each constraint separately, we, however, are interested in the joint satisfaction of these constraints. One approach to test the latter is to use a Bonferroni adjustment, see for instance MacKinnon (2007), by which one multiplies the minimum p-value by the number of constraints, in our case four. Therefore, the p-value of a bootstrap test with Bonferroni adjustment, denoted by $\hat{p}_{\mathrm{bs}}$, is $\hat{p}_{\mathrm{bs}} = 4 \cdot \min(P_{\hat{\theta}})$.

A disadvantage of the Bonferroni adjustment is that testing power decreases as the number of non-binding constraints increases, because $\min(P_{\hat{\theta}})$ is not affected by adding irrelevant constraints, but it will be multiplied by a larger number. To increase finite sample power, several studies suggest to compute critical test values and p-values as a function of the estimated binding constraints in the data, e.g. Andrews and Jia (2008), Andrews and Soares (2010), Bennett (2009), Chen and Szroeter (2012), Hansen (2005), and Donald and Hsu (2011).[10] We therefore also consider the Bennett (2009) test, which is invariant to studentization and based on approximating the distribution of the minimum p-value $\min(P_{\hat{\theta}})$. This is obtained by two sequential bootstraps (where the second resamples from the distribution of the first bootstrap) that are computationally inexpensive compared to using the double (i.e., nested) bootstrap (see Beran, 1988) as suggested in Godfrey (2005). For computing critical values, Bennett (2009) considers both full recentering of all inequality constraints (henceforth B.f) and partial recentering (henceforth B.p) of only those constraints which are (close to being) binding in the data in order to increase power. The test algorithm 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 fully recentered vector $\tilde{\theta}_b^f \equiv \hat{\theta}_b - \hat{\theta}$ and the

[10] Further contributions to the fast evolving literature on inference in models with moment inequalities include Andrews and Guggenberger (2007), Chernozhukov, Hong and Tamer (2007), Fan and Park (2007), Guggenberger, Hahn and Kim (2008), Linton, Song and Whang (2008), and Rosen (2008).

partially recentered vector $\tilde{\theta}_b^p \equiv \hat{\theta}_b - \max(\hat{\theta}, -\delta_n \cdot \hat{\sigma})$, where δ_n is a sequence such that $\delta_n \to 0$ and $\sqrt{n} \cdot \delta_n \to \infty$ as $n \to \infty$.[11]

4. Estimate the vector of p-values under full recentering, denoted by

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

5. Compute the minimum p-values under full recentering: $\hat{p}_f = \min(P_{\tilde{\theta}^f})$.

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 B.f and B.p, denoted by $\hat{p}_{f,b_2}$ and $\hat{p}_{p,b_2}$: $\hat{p}_{f,b_2} = \min(P_{\tilde{\theta}^f,b_2})$, $\hat{p}_{p,b_2} = \min(P_{\tilde{\theta}^p,b_2})$, where $P_{\tilde{\theta}^f,b_2} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{\sqrt{n} \cdot \tilde{\theta}_b^f > \sqrt{n} \cdot \tilde{\theta}_{b_2}^f\}$, $P_{\tilde{\theta}^p,b_2} = B_1^{-1} \cdot \sum_{b=1}^{B_1} I\{\sqrt{n} \cdot \tilde{\theta}_b^f > \sqrt{n} \cdot \tilde{\theta}_{b_2}^p\}$.

8. Compute the p-values of the B.f and B.p tests by the share of bootstrapped minimum p-values that are smaller than the respective minimum p-value of the original sample:
$\hat{p}_{\text{B.f}} = B_2^{-1} \cdot \sum_{b_2=1}^{B_2} I\{\hat{p}_{f,b_2} \leq \hat{p}_f\}$, $\hat{p}_{\text{B.p}} = B_2^{-1} \cdot \sum_{b_2=1}^{B_2} I\{\hat{p}_{p,b_2} \leq \hat{p}_f\}$.

Finally, we consider the Chen and Szroeter (2012) test, which again estimates the binding constraints in the data to increase power. A particularity is that (in the spirit of Horowitz, 1992) the method is based on indicator smoothing (of the functions indicating whether the constraints are binding) at the origin of the elements in θ, where the

[11] As in Bennett (2009) and Chen and Szroeter (2012), we set $\delta_n = \sqrt{\frac{2 \cdot \ln(\ln(n))}{n}}$ in the simulations and applications further below, which has also been considered in Andrews and Soares (2010) and Chernozhukov, Hong and Tamer (2007). It is, however, not guaranteed that this choice is optimal, see for instance the discussion in Donald and Hsu (2011).

distribution of the test statistic is discontinuous. After smoothing, standard asymptotic theory applies to the test statistic, such that bootstrapping is not required to approximate its (otherwise unknown) distribution. Furthermore, Chen and Szroeter (2012) show that their test has correct asymptotic size in the uniform sense under certain conditions,[12] which is a desirable property given that the test statistic's asymptotic distribution is discontinuous w.r.t. the number of binding constraints. For smoothing based on the standard normal distribution (which in the simulations of Section 7 was most powerful among the choices offered in Chen and Szroeter, 2012), the test algorithm is as follows:

1. Estimate the vector of parameters $\hat{\theta}$ and the asymptotic variance $\hat{J}$ of $\sqrt{n} \cdot (\hat{\theta} - \theta)$.

2. Let $\hat{\eta}_i = 1/\sqrt{\hat{J}_i},\ i = 1, \ldots, 4$, where $\hat{J}_i$ is the i-th element of the main diagonal of $\hat{J}$, and compute the smoothing function $\hat{\Psi}_i(\delta_n^{-1} \cdot \hat{\eta}_i \cdot \hat{\theta}_i) = \Phi(\delta_n^{-1} \cdot \hat{\eta}_i \cdot \hat{\theta}_i)$, where Φ is the standard normal cdf and the tuning parameter δ_n is chosen in the same way as for the Bennett (2009) test.

3. Compute the approximation term $\hat{\Lambda}_i = \phi(\delta_n^{-1} \cdot \hat{\eta}_i \cdot \hat{\theta}_i) \cdot \frac{1}{\delta_n \cdot \sqrt{n}}, \quad i = 1, \ldots, 4$, with ϕ being the standard normal pdf.

4. Define the vectors

   $\hat{\Psi} = \Big(\hat{\Psi}_1(\delta_n^{-1} \cdot \hat{\eta}_1 \cdot \hat{\theta}_1), \ldots, \hat{\Psi}_4(\delta_n^{-1} \cdot \hat{\eta}_4 \cdot \hat{\theta}_4)\Big)^T$, $\hat{\Lambda} = \Big(\hat{\Lambda}_1, \ldots, \hat{\Lambda}_4\Big)^T$, $\iota_4 = (1, 1, 1, 1)^T$, $\hat{\Delta} = diag(\hat{J}_1, \ldots, \hat{J}_4)$.

5. Let $\hat{Q}_1 = \sqrt{(}n) \cdot \hat{\Psi}^T \hat{\Delta} \hat{\theta} - \iota_4^T \hat{\Lambda}$ and $\hat{Q}_2 = \sqrt{\hat{\Psi}^T \hat{\Delta} \hat{J} \hat{\Delta} \hat{\Psi}}$.

[12] As discussed in Chen and Szroeter (2012), a sufficient condition for correct asymptotic size in the uniform sense is that the first four moments exist for each of the i.i.d. data points used to estimate $\hat{\theta}$.

6. Compute the p-value of the Chen and Szroeter (2012) test as

$$\hat{p}_{\text{CS}} = \begin{cases} 1 - \Phi\left(\frac{\hat{Q}_1}{\hat{Q}_2}\right) & \text{if } \hat{Q}_2 > 0 \\ 1 & \text{if } \hat{Q}_2 = 0. \end{cases}$$

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, their use allows tightening the bounds on the mean potential outcomes of the always and/or never takers and therefore detecting more violations of IV validity.

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

Assumption 5:

$E(Y^1|T=c) \geq E(Y^1|T=a)$ (mean dominance of compliers).

Assumption 5 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 1-5, (5) becomes

$$E(Y|D=1,Z=1,Y \leq y_q) \leq E(Y|D=1,Z=0) \leq E(Y|D=1,Z=1), \tag{8}$$

which tightens the 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 6:

Assumption 6:

$E(Y^1|T = c) \leq E(Y^1|T = a)$ (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

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

Finally, the combination of Assumptions 5 and 6 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 7:

$E(Y^1|T = c) = E(Y^1|T = a)$ (equality of means).

Clearly, Assumption 7 is most powerful in finding violations of IV validity and implies that

$$E(Y|D = 1, Z = 1) = E(Y|D = 1, Z = 0), \qquad (10)$$

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 7 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 $E(Y^0|T = c) = E(Y^0|T = n)$

amounts to testing whether

$$E(Y|D=0,Z=1)=E(Y|D=0,Z=0). \tag{11}$$

5 Multivalued discrete instruments

This section generalizes the testable implications under mean independence, which were derived for binary instruments, to the case of discrete instruments with multiple values. Frölich (2007) shows that if Z has a bounded support such that $Z \in [z_{\min}, z_{\max}]$, it is possible to define and identify LATEs with respect to any distinct values or subsets in the support of Z. Let z', z'' denote two different mass points in the support of Z satisfying $z_{\min} \leq z' < z'' \leq z_{\max}$. Note that the definition and share of the complier population and all other types depend on the choice of z', z'', because in general, the pool of individuals whose treatment states respond to the instrument varies with the IV values considered. E.g., assuming an instrument $Z \in \{0,1,2\}$, setting $z'=0, z''=1$ generally yields a different proportion and population (in terms of unobservables) of compliers than considering $z'=0, z''=2$.[13] For this reason, we define the types as functions of z', z'': $T=T(z',z'')$ (with $T=T(0,1)$ in the binary case).

Instead of Assumptions 1 to 4, we now invoke Assumptions 1NB to 4NB for $z_{\min} \leq z' < z'' \leq z_{\max}$:

Assumption 1NB:

$\Pr(T(z',z'')=t|Z=z)=\Pr(T(z',z'')=t)$ $\forall$ z in the support of Z and $t \in \{a,c,d,n\}$ (unconf. type).

[13] This is the case because compliers under $z'=0, z''=2$ are those individuals with $D^0=0$ and $D^2=1$ and may also include some individuals that would be never takers if considering $z'=0, z''=1$ ($D^0=0$ and $D^1=0$).

Assumption 2NB:

$E(Y^{d,z}|T(z',z'') = t) = E(Y^d|T(z',z'') = t)$ $\forall$ z in the support of Z, $d \in \{0,1\}$, and $t \in \{a,c,d,n\}$ (mean exclusion restriction).

Assumption 3NB:

$\Pr(D^{z''} \geq D^{z'}) = 1$ (monotonicity).

Assumption 4NB:

$\Pr(D^{z''} > D^{z'}) > 0$ (existence of compliers).

Then, the LATE among compliers ($T(z',z'') = c$) is identified by $E(Y^1 - Y^0|T(z',z'') = c) = \frac{E(Y|Z=z'')-E(Y|Z=z')}{E(D|Z=z'')-E(D|Z=z')}$.

Theorem 8 in Frölich (2007) shows that the LATE on the largest complier population possible is identified by choosing $z' = z_{\min}$ and $z'' = z_{\max}$. Therefore, if Assumption 4NB is not satisfied for $z_{\min}, z_{\max}$ (such that the largest share of compliers possible is zero), it cannot hold for any other pair z', z'' either. On the other hand, if Assumption 4NB holds for $z' = z_{\min}$ and $z'' = z_{\max}$, this does not automatically mean that it also holds for all or any other 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 the absence of compliers maximizes the asymptotic power to find violations of IV validity.

To generalize the testable implications to the case of discrete instruments with multiple values, we somewhat redefine z', z''. The latter may now either refer to particular mass points of Z as before (e.g. $z' = 0$ and $z'' = 1$) or alternatively, to non-overlapping subsets of the support of Z which satisfy the restriction that the largest value in z' is strictly smaller than the smallest in z'' (e.g., z'={0,1} and z"={2,3}). Furthermore, define $\tilde{Z}$ as

$$\tilde{Z} = \begin{cases} 0 & \text{if } Z \in z' \\ \\ 1 & \text{if } Z \in z'' \end{cases}, \tag{12}$$

Under Assumptions 1NB to 4NB, 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:

$$\begin{pmatrix} E(Y|\tilde{Z}=1, D=1, Y \leq y_q) - E(Y|\tilde{Z}=0, D=1), \\ E(Y|\tilde{Z}=0, D=1) - E(Y|\tilde{Z}=1, D=1, Y \geq y_{1-q}), \\ E(Y|\tilde{Z}=0, D=0, Y \leq y_r) - E(Y|\tilde{Z}=1, D=0), \\ E(Y|\tilde{Z}=1, D=0) - E(Y|\tilde{Z}=0, D=0, Y \geq y_{1-r}) \end{pmatrix} \leq 0. \tag{13}$$

Let $n_{\tilde{Z}}$ be the number of possible choices of $\tilde{Z}$ with neighboring mass points or non-overlapping subsets of the support of Z. 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 number of possible definitions of $\tilde{Z}$ with neighboring z', z'' is 4:

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

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 (and thus, finite sample power) and entails a weakly higher complier share than any other neighboring pair (and thus, weakly lower asymptotic power). In large samples, defining the neighboring z', z'' as mass points or rather small subsets of the support appears preferable, because this increases asymptotic power by minimizing the complier share in each pair and maximizing the number of pairs to be tested. However, in small samples, we face a trade-off between asymptotic power and finite sample power due to few observations per pair when defining z', z'' as mass points or small subsets.

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 1 and 2, see for instance Imbens and Angrist (1994):

Assumption 1J:

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

where '⊥' denotes statistical independence. Assumption 1J 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.[14] The subsequent review of the K08 test and the adaptation of our method to joint independence make it obvious that Assumption 1J allows constructing asymptotically more powerful tests based on probability measures (such as density functions) rather than means only. However, it is not yet clear 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.

Henceforth assuming a binary instrument, the testing approach proposed in K08 exploits the fact that under IV validity (now relying on Assumption 1J instead of 1 and 2) 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:[15]

$$\begin{aligned} \Pr(Y \in V, D = 1|Z = 0) &\leq \Pr(Y \in V, D = 1|Z = 1), \\ \Pr(Y \in V, D = 0|Z = 1) &\leq \Pr(Y \in V, D = 0|Z = 0) \ \forall\ V \text{ in the support of } Y. \end{aligned} \tag{14}$$

For testing, K08 makes use of a two sample Kolmogorov-Smirnov-type statistic on the supremum of $\Pr(Y \in V, D = 1|Z = 0) - \Pr(Y \in V, D = 1|Z = 1)$ and $\Pr(Y \in V, D = 0|Z = 1) - \Pr(Y \in V, D = 0|Z = 0)$, respectively, across a predefined collection of

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

[15] 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.

subsets V, henceforth denoted as $\mathcal{V}$. As the test statistic does not converge to any known distribution, the author proposes a bootstrap (or permutation) method which is analogous to Abadie (2002) to compute critical values. An open question of the K08 test is the choice of $\mathcal{V}$, i.e., the definition and number of subsets 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.

In what follows we adapt our testing framework to probability measures (including the pdf and cdf) rather than means and compare the resulting constraints to K08. Analogous to equations (5) and (6) 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:

$$\begin{aligned}
\frac{\Pr(Y \in V|D=1,Z=1)-(1-q)}{q} &\leq \Pr(Y^1 \in V|T=a) \\
&\leq \frac{\Pr(Y \in V|D=1,Z=1)}{q}, \\
\frac{\Pr(Y \in V|D=0,Z=0)-(1-r)}{r} &\leq \Pr(Y^0 \in V|T=n) \\
&\leq \frac{\Pr(Y \in V|D=0,Z=0)}{r}, \qquad (15)
\end{aligned}$$

where q, r are again the shares of always or never takers in the respective mixed populations. Under Assumptions 1J and 3 it follows ($\forall$ V in the support of Y) that $\Pr(Y^1 \in V|T=a) = \Pr(Y \in V|D=1,Z=0)$, $\Pr(Y^0 \in V|T=n) = \Pr(Y \in V|D=0,Z=1)$

and therefore,

$$\begin{aligned}
\frac{\Pr(Y \in V|D=1,Z=1)-(1-q)}{q} &\leq \Pr(Y \in V|D=1,Z=0) \\
&\leq \frac{\Pr(Y \in V|D=1,Z=1)}{q}, \\
\frac{\Pr(Y \in V|D=0,Z=0)-(1-r)}{r} &\leq \Pr(Y \in V|D=0,Z=1) \\
&\leq \frac{\Pr(Y \in V|D=0,Z=0)}{r}. \qquad (16)
\end{aligned}$$

We obtain the following inequality constraints:

$$H_0 : \begin{pmatrix} \frac{\Pr(Y\in V|D=1,Z=1)-(1-q)}{q} - \Pr(Y \in V|D=1,Z=0) \\ \Pr(Y \in V|D=1,Z=0) - \frac{\Pr(Y\in V|D=1,Z=1)}{q} \\ \frac{\Pr(Y\in V|D=0,Z=0)-(1-r)}{r} - \Pr(Y \in V|D=0,Z=1) \\ \Pr(Y \in V|D=0,Z=1) - \frac{\Pr(Y\in V|D=0,Z=0)}{r} \end{pmatrix} \equiv \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \qquad (17)$$

(17) allows us to construct tests with multiple constraints, depending on the definition of $\mathcal{V}$ (with the number of constraints being four times the number of V).

To see how our inequality constraints compare to K08, we use simple algebra (see online appendix A.2) to rewrite (16) as

$$\begin{aligned}
\Pr(Y \in V, D=1|Z=1) - (P_{1|1} - P_{1|0}) &\leq \Pr(Y \in V, D=1|Z=0) \\
&\leq \Pr(Y \in V, D=1|Z=1), \\
\Pr(Y \in V, D=0|Z=0) - (P_{1|1} - P_{1|0}) &\leq \Pr(Y \in V, D=0|Z=1) \\
&\leq \Pr(Y \in V, D=0|Z=0), \qquad (18)
\end{aligned}$$

which must hold for all V in the support of Y. It is easy to see that (18) includes the constraints (14) of K08, but in addition the following:

$$\begin{aligned}
\Pr(Y \in V, D=1|Z=1) - \Pr(Y \in V, D=1|Z=0) &\leq (P_{1|1} - P_{1|0}), \\
\Pr(Y \in V, D=0|Z=0) - \Pr(Y \in V, D=0|Z=1) &\leq (P_{1|1} - P_{1|0}). \qquad (19)
\end{aligned}$$

The 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, consider any $\mathcal{V}$ with non-overlapping subsets V that jointly cover the entire support of Y. By the law of total probability, for instance the first line of (19) must hold because

$$P_{1|1} - P_{1|0} = \sum_{V \in \mathcal{V}} [\Pr(Y \in V, D = 1|Z = 1) - \Pr(Y \in V, D = 1|Z = 0)]. \tag{20}$$

It is worth noting that if $\mathcal{V}$ is defined in the way just described (non-overlapping V that jointly cover the entire support of Y), the constraints (19) are already taken into account by (14) and thus redundant. The prevalence of some $\Pr(Y \in V, D = 1|Z = 1) - \Pr(Y \in V, D = 1|Z = 0) > (P_{1|1} - P_{1|0})$ then necessarily implies the existence of at least one distinct subset V' for which $\Pr(Y \in V', D = 1|Z = 1) - \Pr(Y \in V', D = 1|Z = 0) < 0$ so that (14) is violated, too, otherwise (20) cannot be satisfied.[16] Power gains based on (19) might possibly only be realized if $\mathcal{V}$ contains overlapping V (so that negative densities may be averaged out) and/or does not cover the entire support of Y. As an illustration, a hypothetical example for such a case is provided in online appendix A.5.

In the simulations and the application, we use the methods of Bennett (2009) and Chen and Szroeter (2012) to test (17). These differ in two important ways from the K08 procedure. Firstly, the latter derives the critical value of the test statistic under the least favorable condition for which the null is rejected with the highest probability ($\Pr(Y \in V, D = 1|Z = 0) = \Pr(Y \in V, D = 1|Z = 1)$ and $\Pr(Y \in V, D = 0|Z = 1) = \Pr(Y \in V, D = 0|Z = 0)$). In contrast, in the partial recentering approach of Bennett (2009) and in Chen and Szroeter (2012), the asymptotic null distribution of the respective test statistic is based on pre-estimating for which V the moment inequalities are binding,

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

which may increase finite sample power if $\mathcal{V}$ is confined to a finite class of subsets. On the other hand, these tests cannot deal with an infinite number of inequality constraints, while the K08 test can allow $\mathcal{V}$ to have an infinite number of subsets and therefore may asymptotically screen out more alternatives.[17]

We conclude this section by pointing out that dominance or equality constraints which are in the spirit of the mean restrictions 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. This allows detecting more violations of IV validity and is discussed in online appendix A.3.

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:

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

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 the power to reject violations of IV validity 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

[17] We are grateful to an anonymous referee for pointing this out.

23 %, respectively.[18] These figures are well in the range of complier proportions found in empirical applications, see for instance the example presented in Section 8. Note that in our model, the unconfounded type restriction (Assumption 1) holds by construction because ε and Z are independent. Furthermore, monotonicity (Assumption 3) is satisfied because αZ and ε are linearly separable and α is constant (i.e., the same for everyone).[19] The simulations therefore focus on the validity of the exclusion restriction (which has traditionally received most attention). The latter is satisfied for $\beta = 0$ and violated for any $\beta \neq 0$ implying a direct effect of Z on Y and, therefore, a violation of Assumption 2. For this reason, we suspect the power to reject violations of IV validity to increase in 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

[18] 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.

[19] Non-monotonicity could be introduced for instance by considering a random (rather than a constant) first stage coefficient that takes both positive and negative values. Then, the instrument weakly increases the treatment for some units while it weakly decreases it for others, such that both compliers and defiers may be present (depending on the distribution of the random coefficient and ε).

parameter for the never takers: $\max(\hat{\theta}_3, \hat{\theta}_4)/\text{st.dev.}(Y)$. Column 5 reports the bias of the LATE estimator, which is heavily biased whenever $\beta \neq 0$. 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			
α	β	st.dist1	st.dist0	b.LATE	bs	B.p	B.f	CS
0.2	0.0	-0.090	-0.103	-1.440	0.011	0.017	0.007	0.005
0.6	0.0	-0.223	-0.313	-0.110	0.000	0.000	0.000	0.000
0.2	1.0	0.494	0.458	27.841	0.943	0.960	0.916	0.983
0.6	1.0	0.259	0.096	4.825	0.404	0.506	0.373	0.450
		n=1000			rejection frequencies mean-based tests			
0.2	0.0	-0.118	-0.138	-0.282	0.003	0.003	0.001	0.000
0.6	0.0	-0.243	-0.357	-0.009	0.000	0.000	0.000	0.000
0.2	1.0	0.505	0.482	17.121	1.000	1.000	1.000	1.000
0.6	1.0	0.249	0.094	4.491	0.931	0.965	0.928	0.939

n=250		rejection frequencies probability-based tests					
α	β	B.p(2)	B.f(2)	B.p(4)	B.f(4)	CS(2)	CS(4)
0.2	0.0	0.031	0.013	0.028	0.004	0.001	0.000
0.6	0.0	0.001	0.000	0.003	0.000	0.000	0.000
0.2	1.0	0.982	0.961	0.621	0.534	0.975	0.782
0.6	1.0	0.794	0.691	0.380	0.245	0.712	0.456
n=1000		rejection frequencies probability-based tests					
0.2	0.0	0.005	0.003	0.015	0.003	0.000	0.000
0.6	0.0	0.000	0.000	0.001	0.000	0.000	0.000
0.2	1.0	1.000	1.000	0.992	0.982	1.000	1.000
0.6	1.0	0.998	0.994	0.982	0.945	0.997	1.000

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

Columns 6 to 9 display the rejection frequencies for the tests based on the constraints under mean independence in (7), namely the bootstrap test with Bonferroni adjustment (bs), the Bennett (2009) test with partial (B.p) and full (B.f) recentering, and the Chen and Szroeter (2012) test (CS) using $\Psi(\cdot) = \Phi(\cdot)$ as smoothing function.[20] The results for testing the probability constraints (17) (under the stronger IV assumptions of Section 6) using the methods of Bennett (2009) and Chen and Szroeter (2012) are reported in columns 10 to 15. B.p(2), B.f(2), and CS(2) are based on two subsets V by cutting the distribution of Y in each simulation into two at the value that corresponds to half the difference of the maximum and minimum of the simulated outcome $((\max(Y) - \min(Y))/2)$. B.p(4), B.f(4), and CS(4) use four subsets based on the following partition: $V_1 = (-\infty, -1), V_2 = [-1, 0), V_3 = [0, 1), V_4 = [1, \infty)$. For all tests, the number of bootstrap draws is set to 499. Furthermore, note that the bootstrap test with Bonferroni adjustment and the method of Chen and Szroeter (2012) are based on studentized $\hat{\theta}$, while the Bennett (2009) tests use the original moment constraints (as the results are invariant to studentization).

Under IV validity, the rejection frequencies of any method are quite low and clearly smaller than 5%. The reason for this is that in our simulations with $\beta = 0$, θ lies in the interior of possible parameter values for which the moment inequalities are satisfied, but not at the boundaries (where the asymptotic size is exactly 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 a violation of IV validity ($\beta = 1$), all tests gain power as the sample size grows and lose

[20] In our simulations, we also considered $\Psi(\cdot) = I\{\cdot \geq -1\}$ and $\Psi(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))^{-1}$ as smoothing functions for the Chen and Szroeter (2012) tests. As either choice reduced testing power compared to $\Psi(\cdot) = \Phi(\cdot)$, the results are not reported.

power as the share of compliers becomes larger. Overall, the partially recentered Bennett (2009) test with two subsets (B.p(2)) is most powerful in the given scenario. Whenever the null is violated, it either has the highest rejection frequencies or comes very close to the best performing test. Also CS(2) has comparably high power. Interestingly, the probability-based tests based on four subsets are generally less powerful than those based on two subsets, which demonstrates that the choice of $\mathcal{V}$ may importantly affect the properties of the tests.

Table 4: Simulations - binary outcome

		n=250			mean-based tests			
α	β	st.dist1	st.dist0	b.LATE	bs	B.p	B.f	CS
0.2	0.0	-0.017	-0.082	-0.557	0.012	0.025	0.009	0.016
0.6	0.0	-0.082	-0.441	-0.032	0.000	0.001	0.000	0.000
0.2	1.0	0.132	0.715	6.031	0.798	0.848	0.527	0.935
0.6	1.0	0.123	0.106	0.856	0.192	0.339	0.124	0.326

		n=1000			mean-based tests			
α	β	st.dist1	st.dist0	b.LATE	bs	B.p	B.f	CS
0.2	0.0	-0.040	-0.126	-0.088	0.004	0.009	0.002	0.003
0.6	0.0	-0.110	-0.522	0.000	0.000	0.000	0.000	0.000
0.2	1.0	0.131	0.759	3.685	1.000	1.000	1.000	1.000
0.6	1.0	0.121	0.142	0.800	0.820	0.902	0.792	0.921

		n=250			n=1000		
		prob.-based tests			prob.-based tests		
α	β	B.p(2)	B.f(2)	CS(2)	B.p(2)	B.f(2)	CS(2)
0.2	0.0	0.032	0.010	0.000	0.010	0.004	0.000
0.6	0.0	0.001	0.000	0.000	0.000	0.000	0.000
0.2	1.0	0.849	0.769	0.334	1.000	1.000	1.000
0.6	1.0	0.338	0.185	0.126	0.899	0.813	0.890

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

Table 4 presents the rejection frequencies for binary outcomes, where the DGP is identical to the previous one with the exception that $Y = I\{D+\beta Z+U > 0\}$. Therefore, the true treatment effect depends on the parameter α and is 0.386 for $\alpha = 0.2$ and 0.403 for $\alpha = 0.6$. Concerning testing, note that the mean tests need to be modified in a innocuous way to be appropriate for binary outcomes, see online appendix A.4. Furthermore, probability-based tests with four subsets are not considered, because the outcome can only take two values. As before, all tests are quite conservative when $\beta = 0$ and even more so for the larger share of compliers and/or sample size. Under $\beta = 1$, the CS (mean) test has the highest power overall, but also B.p and B.p(2) perform decently.

8 Application

This section presents an application to labor market data from 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. We test IV validity in the full sample (i.e., unconditionally) and in four different subsamples (consisting of white individuals not living in the South) in order to control for covariates that are potentially correlated with both the instrument and the outcome. The subsamples differ in terms of father's education (12 or mode years or less than 12 years) and living in an urban or rural area.

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

				p-values mean-based tests			
Sample	compl.	st.dist1	st.dist0	bs	B.p	B.f	CS
full sample	6.9%	-0.203	0.224	0.002	0.001	0.001	0.001
w,non.S.,f.edu≥12,u	13.2%	-0.419	-0.302	1.000	0.787	1.000	1.000
w,non.S,f.edu<12,u	3.6%	-0.051	-0.016	1.000	0.665	0.866	0.919
w,non.S,f.edu≥12,r	16.3%	-0.061	-0.396	1.000	0.602	0.936	0.986
w,non.S,f.edu<12,r	6.7%	-0.092	-0.164	1.000	0.777	0.976	0.989

p-values probability-based tests

Sample	B.p(2)	B.f(2)	B.p(4)	B.f(4)	CS(2)	CS(4)
full sample	0.001	0.002	0.002	0.006	0.000	0.003
w,non.S.,f.edu≥12,u	0.979	0.735	0.907	0.997	1.000	1.000
w,non.S,f.edu<12,u	0.762	0.873	0.008	0.040	0.941	0.627
w,non.S,f.edu≥12,r	0.699	0.894	0.556	0.723	0.962	0.952
w,non.S,f.edu<12,r	0.103	0.187	1.000	1.000	0.443	0.872

Note: Tests are based on 1999 bootstrap draws. w=white, non.S=non-South, f.edu=father's education, u=urban, r=rural.

Table 5 presents the results. The first column gives the estimated complier proportion, which is crucial for the tests' power to detect violations of IV validity, 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) as well as the Bennett (2009) and Chen and Szroeter (2012) tests (using $\Psi(\cdot) = \Phi(\cdot)$ as smoothing function) for the mean constraints (B.p, B.f, CS) and the probability constraints (B.p(2), B.f(2), B.p(4), B.f(4), CS(2), CS(4)) with 2 and 4 subsets V. The latter are defined by an equidistant grid over the support of Y. In the full sample, the point estimate of the mean potential outcome of the never takers falls well outside its bounds and all tests reject the null at the 1% level. In contrast, none of the mean constraints is binding in any of the subsamples and with the exception of B.p(4) and B.f(4) in the third subsample, all tests yield rather large p-values. This leads us to conclude that while college proximity is most likely not an unconditionally valid instrument, IV validity conditional on covariates is generally not refuted.[21]

Finally, we test the mean equality constraints (10) and (11) 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, all subsamples pass the stricter difference of means tests at the 5% level of significance. This suggests that IV validity and homogeneity of the mean

[21] As IV validity is required to hold in any subpopulation defined on the covariates, testing should ideally include all possible combinations of covariates to maximize asymptotic power. However, this may be infeasible in empirical applications due to small samples problems. A method specifically designed for testing conditional on covariates is yet to be developed.

potential outcomes of compliers and always takers under treatment and of compliers and never takers under non-treatment is likely to hold given the covariates considered.

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
full sample	6.449	6.369	0.081	0.012
w,non.S,f.edu≥12,u	6.465	6.483	-0.018	0.806
w,non.S,f.edu<12,u	6.431	6.568	-0.137	0.193
w,non.S,f.edu≥12,r	6.443	6.258	0.184	0.051
w,non.S,f.edu<12,r	6.518	6.442	0.076	0.390
Sample	$\bar{Y}_{D=0,Z=0}$	$\bar{Y}_{D=0,Z=1}$	diff	p-val
full sample	6.094	6.254	-0.160	0.000
w,non.S,f.edu≥12,u	6.348	6.390	-0.043	0.569
w,non.S,f.edu<12,u	6.383	6.336	0.047	0.447
w,non.S,f.edu≥12,r	6.212	6.217	-0.005	0.950
w,non.S,f.edu<12,r	6.261	6.259	0.001	0.985

Note: w=white, non.S=non-South, f.edu=father's education, u=urban, r=rural.

9 Conclusion

The LATE framework of Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) 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 implications for IV validity, which can be formulated as inequality moment constraints. We consider various methods for testing the latter: a simple bootstrap test with Bonferroni adjustment, the minimum p-value based test of Bennett (2009), and the Chen and Szroeter (2012) test using indicator smoothing. Furthermore, we point out that power against violations of IV validity may be increased by imposing restrictions on the order of the mean potential outcomes across subpopulations. If one is even willing to assume equality in mean potential outcomes across subpopulations, simple difference of means tests can be used for testing. We also relate our work to Kitagawa (2008), who tests for the non-negativity of complier outcomes' densities in subsets of the outcome support to verify IV validity. By adapting our framework to probability measures rather than means of potential outcomes, we obtain testable implications that are equivalent to those in Kitagawa (2008), given that the probability measures used for testing are defined by non-overlapping subsets that jointly cover the entire outcome support. Finally, we briefly investigate the finite sample properties of our tests and consider an empirical application to labor market data from 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.

References

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

Andrews, Donald and Patrik Guggenberger, "The Limit of Finite-Sample Size and a Problem with Subsampling," *Cowles Foundation Discussion Paper 1605R* (2007).

Andrews, Donald and Panle Jia, "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *Cowles Foundation Discussion Paper 1676* (2008).

Andrews, Donald and Gustavo Soares, "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica* 78 (2010), 119–157.

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

Angrist, Joshua, Guido Imbens and Donald Rubin, "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association* 91 (1996), 444–472.

Angrist, Joshua and Alan Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics* 106 (1991), 979–1014.

Balke, Alexander and Judea Pearl, "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association* 92 (1997), 1171–1176.

Bennett, Christopher, "Consistent and Asymptotically Unbiased MinP Tests of Multiple Inequality Moment Restrictions," Working Paper 09-W08, Department of Economics, Vanderbilt University, 2009.

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

Bound, John, David Jaeger and Regina Baker, "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 (1995), 443–450.

Cameron, Adrian Colin and Pravin Trivedi, *Microeconometrics* (Cambridge: Cambridge Univ. Press, 2005).

Card, David, "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," (pp. 201–222), in Loizos Christofides, Kenneth Grant and Robert Swidinsky, eds., *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp* (Toronto: University of Toronto Press, 1995).

———, "The Causal Effect of Education on Earnings," (pp. 1802–1863), in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics* (Amsterdam: North-Holland, 1999).

Chen, Le-Yu and Jerzy Szroeter, "Testing Multiple Inequality Hypotheses: A Smoothed Indicator Approach," *CeMMAP working paper 16/12* (2012).

Chernozhukov, Victor, Han Hong and Elie Tamer, "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica* 75 (2007), 1243–1284.

Donald, Stephen and Yu-Chin Hsu, "A New Test for Linear Inequality Constraints When the Variance Covariance Matrix Depends on the Unknown Parameters," *Economics Letters* 113 (2011), 241–243.

Fan, Yanqin and Sang Soo Park, "Confidence Sets for Some Partially Identified Parameters," *mimeo* (2007).

Flores, Carlos and Alfonso Flores-Lagunes, "Nonparametric Partial Identification of Causal Net and Mechanism Average Treatment Effects," *mimeo, University of Florida* (2010).

Frölich, Markus, "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics* 139 (2007), 35–75.

Frölich, Markus and Blaise Melly, "Unconditional Quantile Treatment Effects under Endogeneity," *IZA DP No. 3288* (2008).

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

Greene, William, *Econometric analysis*, 6 edition (Upper Saddle River, NJ: Pearson Prentice Hall, 2008).

Guggenberger, Patrik, Jinyong Hahn and Kyooil Kim, "Specification testing under moment inequalities," *Economics Letters* 99 (2008), 375–378.

Hansen, Peter Reinhard, "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics* 23 (2005), 365–380.

Heckman, James and Edward Vytlacil, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica* 73 (2005), 669–738.

Horowitz, Joel, "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica* 60 (1992), 505–531.

———, "The Bootstrap," (pp. 3159–3228), in James Heckman and Edward Leamer, eds., *Handbook of Econometrics* (North-Holland, 2001).

Horowitz, Joel and Charles Manski, “Identification and Robustness with Contaminated and Corrupted Data,” *Econometrica* 63 (1995), 281–302.

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

Imbens, Guido and Joshua Angrist, “Identification and Estimation of Local Average Treatment Effects,” *Econometrica* 62 (1994), 467–475.

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

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

Linton, Oliver B., Kyungchul Song and Yoon-Jae Whang, “Bootstrap tests of stochastic dominance with asymptotic similarity on the boundary,” *CeMMAP working paper 8/08* (2008).

MacKinnon, James, “Bootstrap Hypothesis Testing,” Working Paper 1127, Queen’s University, Department of Economics, 2007.

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

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

Richardson, Thomas, Robin Evans and James Robins, “Transparent parameterizations of models for potential outcomes,” (pp. 569–610), in James Berger A. P. Dawid David Heckerman Adrian Smith José Bernardo, M. J. Bayarri and Mike West, eds., *Bayesian Statistics 9* (Oxford University Press, 2011).

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

Rubin, Donald, “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology* 66 (1974), 688–701.

Sargan, John D., “The Estimation of Economic Relationships using Instrumental Variables,” *Econometrica* 26 (1958), 393–415.

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

Zhang, Junni and Donald Rubin, “Estimation of causal effects via principal stratification when some outcome are truncated by death,” *Journal of Educational and Behavioral Statistics* 28 (2003), 353–368.