Testing Instrument Validity and Identification with Invalid Instruments

David Slichter*

Submitted for SOLE, May 2014

Abstract

Instrumental variables are widespread in empirical economics, but their use is plagued by concerns that proposed instruments do not satisfy the validity condition. This paper develops a framework for testing the validity of instruments with covariates, using the motivating example of proximity to college as an instrument for the effect of college attendance on wages (Card 1995). I focus on the case of a binary instrument for a binary treatment, but the approach can be extended to continuous variables. Because this approach makes it possible to quantify the degree of invalidity, it can often allow for point and set identification of treatment effects even when the instrument is invalid. Applying the approach, I find that the proximity to college instrument is invalid and likely overestimates the returns to college.

1 Introduction

Instrumental variables are popular in empirical economics because they allow the researcher to identify the causal effect of a treatment on an outcome even when treatment is correlated with unobserved determinants of the outcome. Their use is complicated, though, by the difficulty of establishing the required condition that a proposed instrument can properly be excluded from the outcome equation.

This paper proposes a framework for testing the assumption that an instrument is valid, i.e. that it can properly be excluded from the outcome equation. Furthermore, the proposed framework allows one to quantify an instrument's direct effect for subsets of the population. In circumstances when this finding can reasonably be extended to make inferences about the direct effect in the population as a whole,

*Department of Economics, University of Rochester. E-mail: dslichte@z.rochester.edu. I am indebted to Gregorio Caetano, Nese Yildiz, Josh Kinsler, and Carolina Caetano for their guidance. I am also grateful to Mark Bils, Costi Yannelis, Mike Elsby, and seminar participants at the University of Rochester student workshop for their helpful comments. All remaining errors are my own.

this allows us to make accurate inferences about causal effects even using invalid instruments. Because any source of variation in treatment status can be considered an invalid instrument (if not a valid one), this allows for causal inferences in a broad class of situations. The required assumptions are not onerous; the requirements for set identification can be satisfied, for instance, by a selection model with comparatively agnostic functional form and distributional assumptions.

As a motivating example, I consider the problem of estimating the wage returns to college education. It is a puzzling stylized fact that instrumental variable estimates of the wage returns to schooling are often higher than ordinary least squares (OLS) estimates, since an individual's unobserved ability would intuitively seem to be positively correlated with years of schooling. There are several possible explanations; for example, Card (1999) suggests that local treatment effects might be larger than average treatment effects, and Griliches (1977) gives reasons why the direction of bias in OLS might not be so clear. Another simple explanation is that at least some of the instruments used to identify wage returns to schooling do not satisfy the required validity condition and therefore yield biased estimates.

Card (1995) suggests that growing up in a labor market containing a college can be used as an instrument for the effect of college attendance on wages, conditional on some covariates. This instrument suffers from well-known validity concerns – for example, perhaps the presence of a college in the local labor market is related to unobserved ability, or perhaps there are externalities associated with human capital such that the instrument identifies a mixture of private and social returns to schooling. But can we assemble an organized empirical case for whether these complaints have merit or not? And, if the instrument is invalid, can we still use it to rescue some information about true causal effects?

Intuition The intuition for my approach to testing validity is as follows. If Z is a valid instrument for the effect of treatment D on outcome Y, then Z has no causal connection with Y except through D. A typical causal diagram is presented in Figure 1.[1] If we could shut off Z's relevance (that is, remove any connection between Z and D), then there would no longer be any relationship between Z and Y. Then if we shut off the relevance and discover that Z and Y are still related, we can conclude that Z is not a valid instrument.

This principle leads to a simple placebo test using covariates. If covariates X shut off the relevance (that is, for some x, when $X = x$, Z is unrelated to D) without simultaneously shutting off any direct effect of Z on Y, then we can detect whether Z is valid or not by seeing whether Z and Y are related for observations with $X = x$. Some empirical papers already implement placebo tests with this underlying logic (e.g. Altonji et al. 2005a, Madestam et al. 2013), and my framework starts by formalizing this logic.

[1]Disclaimer: The diagram is illustrative of a common case, but of course not all instruments have this causal diagram. For instance, Z would still be valid if, rather than Z causing D, both Z and D had some common cause Z' which was also unconnected to Y except through D.

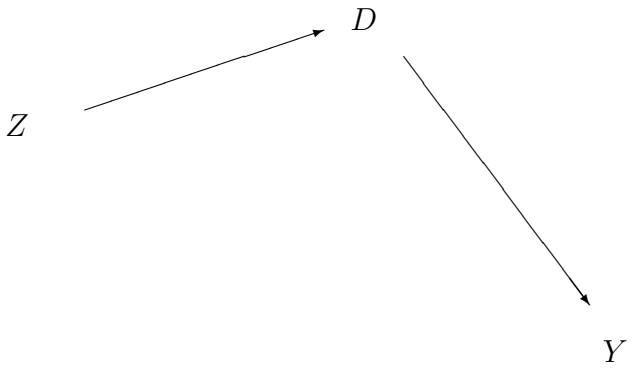

Figure 1: Example of a valid instrument

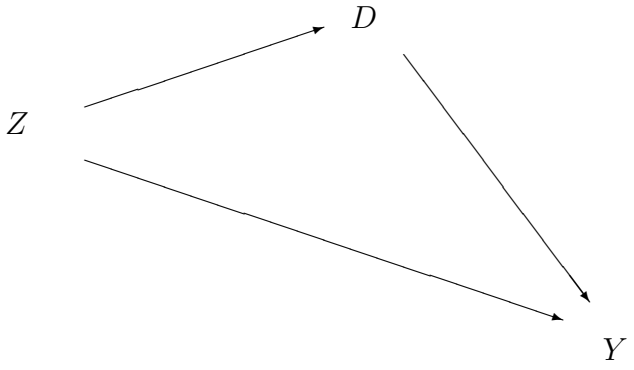

Figure 2: Example of an invalid instrument

To illustrate the placebo logic, consider how it could be applied to the proximity to college instrument. Suppose that we observe a measure of high school IQ for all the individuals in our data. We would imagine that students with very high IQ scores would almost certainly attend college regardless of where they lived. Similarly, students with very low measured IQs would seem very unlikely to attend college. These are the always takers and never takers (e.g. of Imbens and Angrist 1994); for them, the instrument has no relevance. Therefore we can test whether there is any relationship among them between their proximity to college (Z) and their log wages (Y). If the instrument is valid, we should find no such relationship. If the instrument is not valid, we will uncover such a relationship so long as the instrument is not invalid exclusively for those individuals with average IQs and GPAs.

The weakness of the placebo approach is that, in practice, we often lack covariates powerful enough to completely shut off the instrument's relevance. In the schooling example, it turns out that, even at extremely high measured IQ levels, there are still a few students not attending college, and visa versa at low measured IQ levels. Therefore, the second piece of the framework I develop in this paper will focus on cases where the covariates X merely cause *variation* in the degree to which Z and D are related – an imperfect placebo. The underlying principle is that, as we approach no relevance, we should also approach no relationship between Z and Y.

Contributions The first contribution of this paper is to build a framework for evaluating instruments by treating a placebo test as the limit of not-quite placebo tests. This imperfect placebo test starts by dividing the population by covariate values, then compares the first stage and reduced form effect of the instrument at each value of the covariate X. If the instrument is valid, then the reduced form should be close to zero whenever the first stage is close to zero. This can be tested statistically, and it is easy to present graphical evidence as well.

This approach will be applicable in many settings. There is nothing particularly special about the example of the proximity to college instrument; the variation in relevance arises simply because treatment status is binary and because measured IQ happens to be a powerful predictor of treatment status. The requirements will generally be satisfied in applications where treatment status is binary or censored and where some observed covariates are strong predictors of treatment status. The requirements may also sometimes be satisfied without bunching in the treatment variable as well.

The second contribution of this paper is in identifying causal effects using invalid instruments. Consider the possibility that Z is invalid because it directly causes Y, but we are able to measure this direct causal effect by shutting off the indirect effect of Z on Y through D using our covariates. Once we know the direct effect, we could easily determine the effect of Z on Y through D, which then allows us to recover the effect of D on Y. Difference-in-differences estimation can be thought of as an instance of this approach, where the covariate that shuts off relevance is time period (the instrument – membership in the treatment group – does not lead to treatment

in some periods), and the assumption of common trends is an assumption that time period does not modulate the direct effect of Z on Y. For many applications, it may be more reasonable to relax the assumption that we can measure the exact invalidity (which applies to every observation) to an assumption that we can bound the average invalidity, thereby yielding set identification of treatment effects. I show that the assumptions required for this set identification can be satisfied by a selection model with weak assumptions.

The third contribution of this paper is to apply the frameworks for testing instrument validity and for identification with invalid instruments to the problem of returns to schooling, using the proximity to college instrument. I find that the strength of the relationship between college proximity and college attendance within each decile of measured IQ is not a successful predictor of the strength of the relationship between college proximity and adult wages in that decile. This evidence suggests that proximity to college is not a valid instrument. The results of the set identification approach suggest that, when treated as a valid instrument, proximity to college likely overestimates the return to college.

The paper proceeds as follows. Section 2 briefly reviews the literature. Section 3 develops the econometric framework for testing instrument validity. Section 4 discusses identification of treatment effects with an instrument which is found to be invalid. Section 5 applies these methods to the proximity to college instrument. Section 6 concludes.

2 Related literature

This paper fits into large literatures on each of its three topics. There are a number of papers which develop techniques for testing instrument validity in the case where the number of instruments exceeds the number of endogenous regressors (e.g. Anderson and Rubin 1949, Sargan 1958, Hansen 1982). Kitagawa (2008) and Huber and Mellace (2011) offer tests for instrument assumptions with a just-identified instrument, using the constraint that always and never takers must receive the same outcome regardless of the value of the instrument to produce testable inequalities. Both of these papers use outcome distributions rather than covariates to place bounds on which data must represent always and never takers. A consequence of this agnosticism is that their tests can be quite undersized unless the first stage is small. In the case of testing instrument validity, an undersized test is not necessarily conservative, since many researchers who might wish to test the validity of an instrument are proposing the instrument themselves. Furthermore, their approaches encounter some difficulty testing the claim that an instrument is valid conditional on other variables. Caetano, Rothe, and Yildiz (2013) show that invalid instruments can be detected under certain kinds of censoring of treatment status. While their approach does not suffer from incorrect size, the restriction on the distribution of treatment status prevents its use in many applications.

My approach differs from all of these approaches through its reliance on covariates. This imposes a new constraint, but I will argue that the requirements for covariates in my approach are not onerous, so that my approach can be implemented with many research designs so long as the sample size is adequate.[2] The primary requirement for tests based on my approach to have power – variation in relevance of the instrument – is likely to be satisfied whenever treatment status is binary, censored, or bunched and we observe covariates which are good predictors of treatment status, though it can be satisfied in other contexts as well. My approach also offers appropriate size except in circumstances which would call into question whether it is reasonable to generalize from local average treatment effects.

This paper also fits into a literature on identifying and estimating treatment effects with invalid instruments. Flores and Flores-Lagunes (2010) give set identification of treatment effects, though their very weak assumptions may lead to very wide bounds. Other papers provide results in cases where the instrument is invalid in useful ways; for instance, Kolesar et al. (2013) considers the case where a large number of instruments are invalid individually but valid on average, and Reinhold and Woutersen (2011) provide bounds when instruments are invalid but create less problematic variation in treatment than OLS. Additionally, many selection models can also be seen as identifying treatment effects when the sources of variation in treatment are not exogenous. This can be extended to set identification, as for instance through the conditions developed in Altonji et al. (2005b).[3] My set identification results apply in somewhat different situations from Altonji et al. and the relative weakness and plausibility of the assumptions required for their and my approaches will depend on context.

Finally, there is an enormous literature on the wage returns to schooling. Most papers which use U.S. data from recent years find a return on the order of 10% per additional year of education using OLS and higher returns using IV, a pattern which many researchers have found counterintuitive (e.g. Ashenfelter and Rouse 1999, Card 2001). Other papers (e.g. Keane and Wolpin 1997, Carneiro et al. 2011) have found lower average treatment effects using structural approaches. This paper helps understand the high returns to schooling measured through instrumental variables by solidifying the case that one well-known instrument is in fact invalid.

[2]My preferred implementation of my approach requires a workable sample size for estimation in each of multiple subpopulations, where the subpopulations are defined by covariate values. Therefore it becomes difficult to test the validity of instruments with any power when the sample is barely large enough to implement an IV approach with any power.

[3]Altonji et al. frame their approach as a robustness check. I describe their approach as set identification because they are effectively establishing a range of plausible values for the parameter of interest under the assumption that the degree of selection on unobservables lies between zero and the degree of selection on observables.

3 A framework for testing validity

Suppose we are interested in the effect of treatment D on outcome Y, and we also observe proposed instrument Z and covariates X. Assume that Z and D are binary. I will borrow notation from Imbens and Angrist (1994). For each individual i, let $Y_i(1)$ be i's potential outcome if i takes the treatment (that is, if $D_i = 1$) and $Y_i(0)$ be the potential outcome if i does not. Similarly, let $D_i(1)$ be i's potential treatment status if i is assigned to treatment by the instrument ($Z_i = 1$) and $D_i(0)$ be i's potential treatment status if assigned to the control group ($Z_i = 0$). We can assign labels to the types of individuals:

- i is an *always taker* if $D_i(0) = D_i(1) = 1$.
- i is a *never taker* if $D_i(0) = D_i(1) = 0$.
- i is a *complier* if $D_i(0) = 0$ and $D_i(1) = 1$.
- i is a *defier* if $D_i(0) = 1$ and $D_i(1) = 0$.

Once again following Imbens and Angrist, Z can be used to identify the local average treatment effect (LATE) under the following assumptions:

Assumption 1. *(Relevance)* $E[D_i \mid Z_i = z, X_i = x]$ is non-trivial in z for some x.

Assumption 2. *(Validity)* $(Y_i(1), Y_i(0), D_i(1), D_i(0))$ are jointly independent of Z_i conditional on X_i.

Assumption 3. *(No defiers)* There is no i such that i is a defier.

Notice that we are conditioning on X. This will be important.[4]

As the first piece of our framework, define a new function, comply(x), in the following way:

$$\text{comply}(x) \equiv E(D \mid Z = 1, X = x) - E(D \mid Z = 0, X = x)$$

Under our instrument assumptions, comply(x) captures the probability that an individual whose covariates take value x is a complier. That is, it is the size of the first stage conditional on $X = x$. Figure 3 shows a local linear fit of probability of attending college conditional on IQ and college proximity in the NLS data used by

[4]The original Imbens and Angrist paper does not condition on covariates. However, the proposition follows closely from Imbens and Angrist: At each value of X, the Imbens and Angrist assumptions are satisfied, so the LATE is identified at each x in the domain of X. Because of monotonicity, the measure of compliers at each value of X is identified, and so is the distribution of values of X. Therefore the overall LATE is identified as a weighted average of the LATEs identified at each value of X.

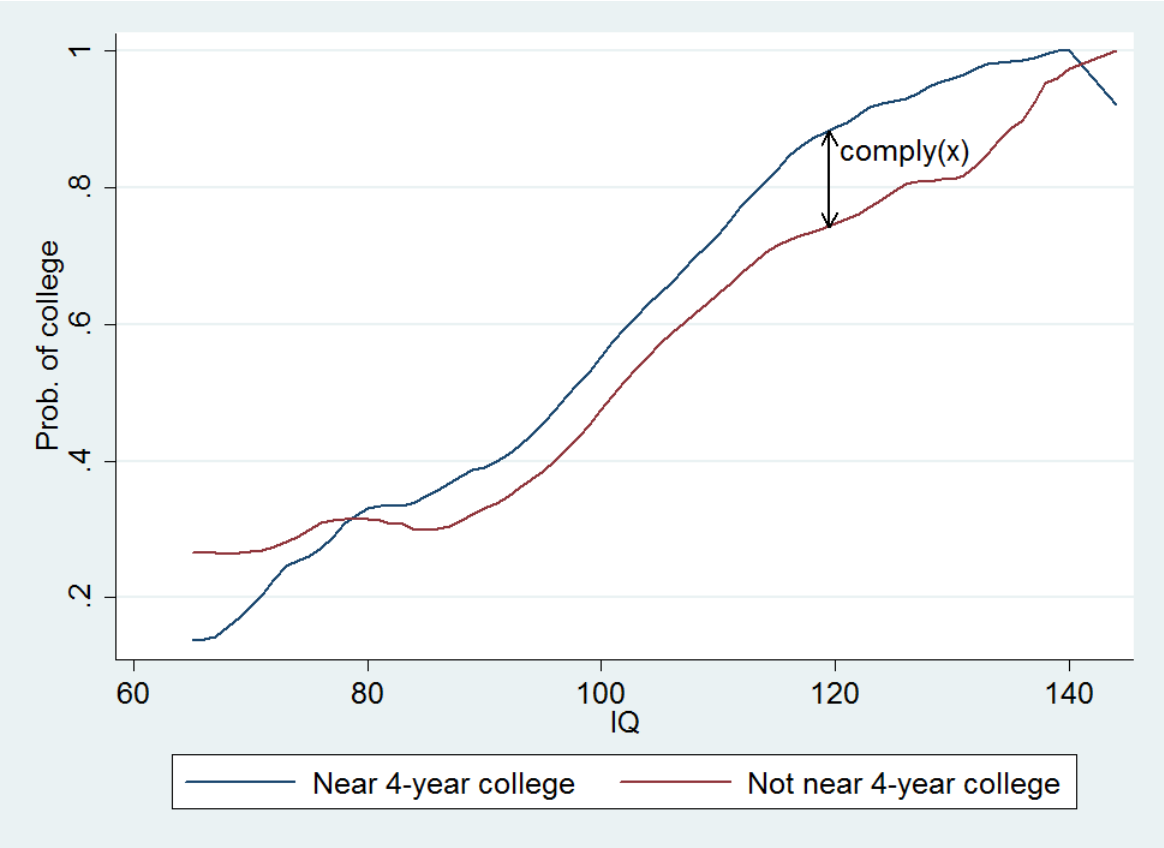

Figure 3: Estimated comply function for Card's instrument with IQ as X.

Card.[5] The function comply(IQ) can be estimated as the distance between the curves at each level of IQ. Under instrument assumptions, the true value of comply(x) should never be negative. The figure presumably shows some negative values due to estimation error; the confidence intervals become quite large at extreme values, where there are few data points.

Now define gain(x) in a parallel way:

$$\text{gain}(x) \equiv E(Y \mid Z = 1, X = x) - E(Y \mid Z = 0, X = x)$$

This function captures the extent to which the outcome Y and the instrument Z are related at each value of the covariates X. It is the size of the reduced form effect conditional on $X = x$. Figure 4 illustrates the gain function with NLS data using IQ as the covariate.[6] As with Figure 3, the confidence intervals become large at values far from 100. In the example of the proximity to college instrument, gain(x) = 0

[5]The fit is local linear with bandwidth determined by Silverman's rule. Note that the ends of the IQ distribution have been trimmed from the diagram for clarity due to excessively large confidence intervals.

[6]As with the comply function, the fit is local linear and the ends have been trimmed for clarity.

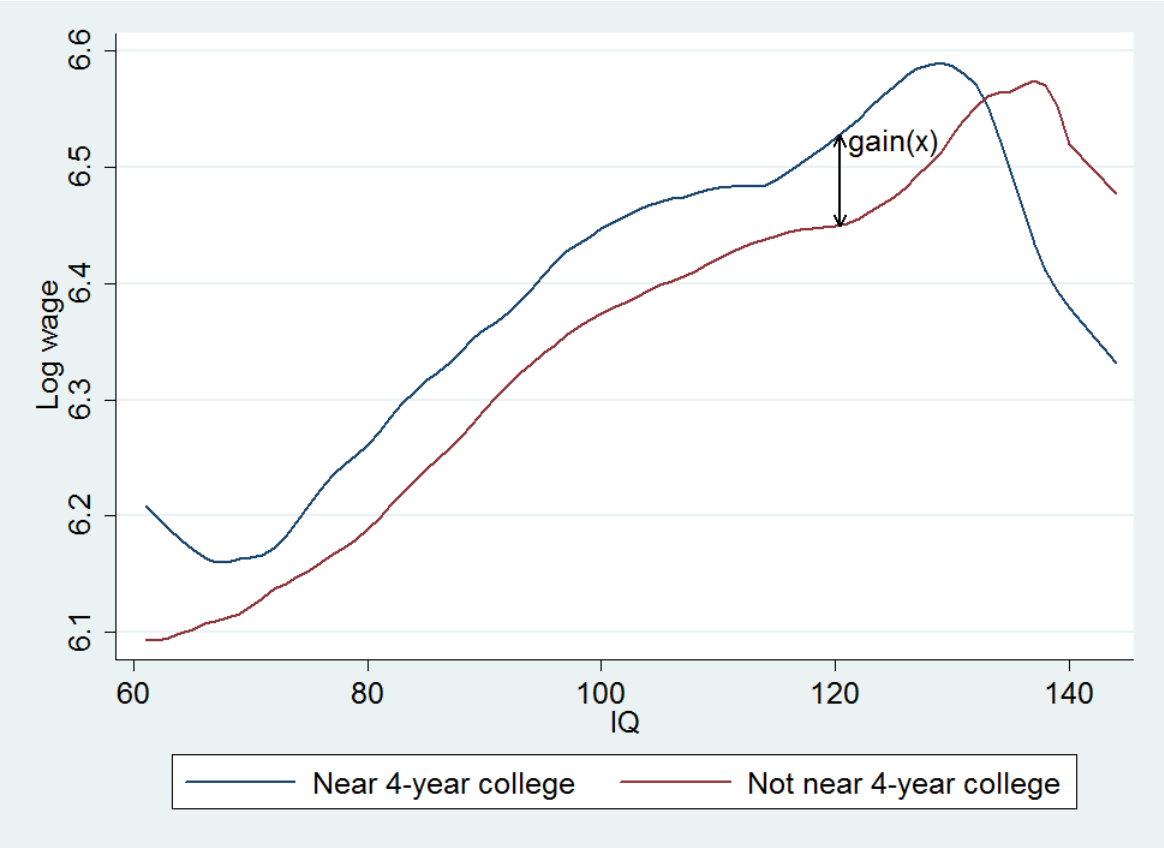

Figure 4: Estimated gain function for Card's instrument with IQ as X.

would mean that individuals whose covariates take the value x and live near a college earn the same amount on average as individuals with the same covariate value who do not live near a college.

Finally, I introduce one additional assumption.

Assumption 4. *(Finite expectations)* $E(|Y|)$ is finite.

This assumption is needed to prevent the possibility that a zero measure of the population might disrupt population averages, and to justify a continuity argument that allows for extrapolation from cases where there are compliers to cases where there are no compliers.

Theorem 1. *Let* $F_{Y|X,Z}(y \mid x, z)$ *be the population distribution of* Y *conditional on* $X = x$ *and* $Z = z$*. Suppose* Z *satisfies validity and no defiers conditional on* X*. Then for all* x *such that comply*$(x) = 0$*,* $F_{Y|X,Z}(y \mid x, 1) = F_{Y|X,Z}(y \mid x, 0)$*. We call such* x *a "no-relevance point".*

Proof. Suppose that x is a no-relevance point. By definition, comply$(x) = E(D \mid Z = 1, X = x) - E(D \mid Z = 0, X = x)$. Because there are no defiers, $E(D \mid Z = 1, X = x)$

is the probability that i is a complier or an always taker conditional on $X = x$, while $E(D \mid Z = 0, X = x)$ is the probability that i is an always taker conditional on $X = x$. Types are mutually exclusive, so comply(x) is the probability that i is a complier if $X_i = x$, and therefore there is a zero probability that i is a complier at x.

Vytlacil (2002) shows that the instrument assumptions are equivalent to the following latent selection model: Let the propensity score $p(x, z) \equiv P(D = 1 \mid X = x, Z = z)$. Then there exists some random variable U such that $D_i = 1\{U_i > p(x, z)\}$ for all i such that $X_i = x, Z_i = z$, and such that Z_i is jointly independent of $(U_i, Y_i(1), Y_i(0))$. I incorporate this approach for the rest of the proof.

We are to show that $F_{Y|X,Z}(y \mid x, 1) - F_{Y|X,Z}(y \mid x, 0) = 0$ for all y. Now,

$$
\begin{aligned}
&F_{Y|X,Z}(y \mid x, 1) - F_{Y|X,Z}(y \mid x, 0) \\
=&[Pr(Y_i(0) \leq y, U_i \leq p(x, z) \mid X_i = x, Z_i = 1) + Pr(Y_i(1) \leq y, U_i > p(x, z) \mid X_i = x, Z_i = 1)] \\
&- [Pr(Y_i(0) \leq y, U_i \leq p(x, z) \mid X_i = x, Z_i = 0) + Pr(Y_i(1) \leq y, U_i > p(x, z) \mid X_i = x, Z_i = 0)] \\
=&Pr(Y_i(0) \leq y \mid X_i = x, Z_i = 1) Pr(U_i \leq p(x, z) \mid Y_i(0) \leq y, X_i = x, Z_i = 1) \\
&+ Pr(Y_i(1) \leq y \mid X_i = x, Z_i = 1) Pr(U_i > p(x, z) \mid Y_i(1) \leq y, X_i = x, Z_i = 1) \\
&- Pr(Y_i(0) \leq y \mid X_i = x, Z_i = 1) Pr(U_i \leq p(x, z) \mid Y_i(0) \leq y, X_i = x, Z_i = 0) \\
&- Pr(Y_i(1) \leq y \mid X_i = x, Z_i = 1) Pr(U_i > p(x, 0) \mid Y_i(1) \leq y, X_i = x, Z_i = 0) \\
=&Pr(Y_i(0) \leq y \mid X_i = x) Pr(U_i \leq p(x, 1) \mid Y_i(0) \leq y, X_i = x) \\
&+ Pr(Y_i(1) \leq y \mid X_i = x) Pr(U_i > p(x, 1) \mid Y_i(1) \leq y, X_i = x) \\
&- Pr(Y_i(0) \leq y \mid X_i = x) Pr(U_i \leq p(x, 0) \mid Y_i(0) \leq y, X_i = x) \\
&- Pr(Y_i(1) \leq y \mid X_i = x) Pr(U_i > p(x, 0) \mid Y_i(1) \leq y, X_i = x)
\end{aligned}
$$

The first equality follows from the definitions of distributions. The second equality follows from the rule that $Pr(A, B) = Pr(A)Pr(B \mid A)$. The third equality follows from the independence of Z from $Y_i(0)$ and $Y_i(1)$.

If $X_i = x$, i must be either an always taker or a never taker. Therefore, $U_i \leq p(x, 1)$ if and only if $U_i \leq p(x, 0)$. As a result, the first and third terms above cancel, as do the second and fourth. We conclude that $F_{Y|X,Z}(y \mid x, 1) = F_{Y|X,Z}(y \mid x, 0)$. □

Corollary 1. *Suppose Z satisfies validity conditional on X, no defiers, and finite expectations. Then for all x such that comply$(x) = 0$, gain$(x) = 0$.*

Proof. By Theorem 1, if x is a no-relevance point, then $F_{Y|X,Z}(y \mid x, 1) = F_{Y|X,Z}(y \mid x, 0)$. Then $E(Y \mid X = x, Z = 1) = E(Y \mid X = x, Z = 0)$ if either expectation exists. We have assumed they exist. Therefore gain$(x) = 0$. □

The theorem provides the more general result. However, the corollary is more practical for implementation for two reasons. First, the comparison of distributions requires some statistic by which to compare them. The mean has natural appeal in part because instruments are frequently only required to be mean-independent, not independent, of an error term (e.g. in 2SLS estimation), and in part because the

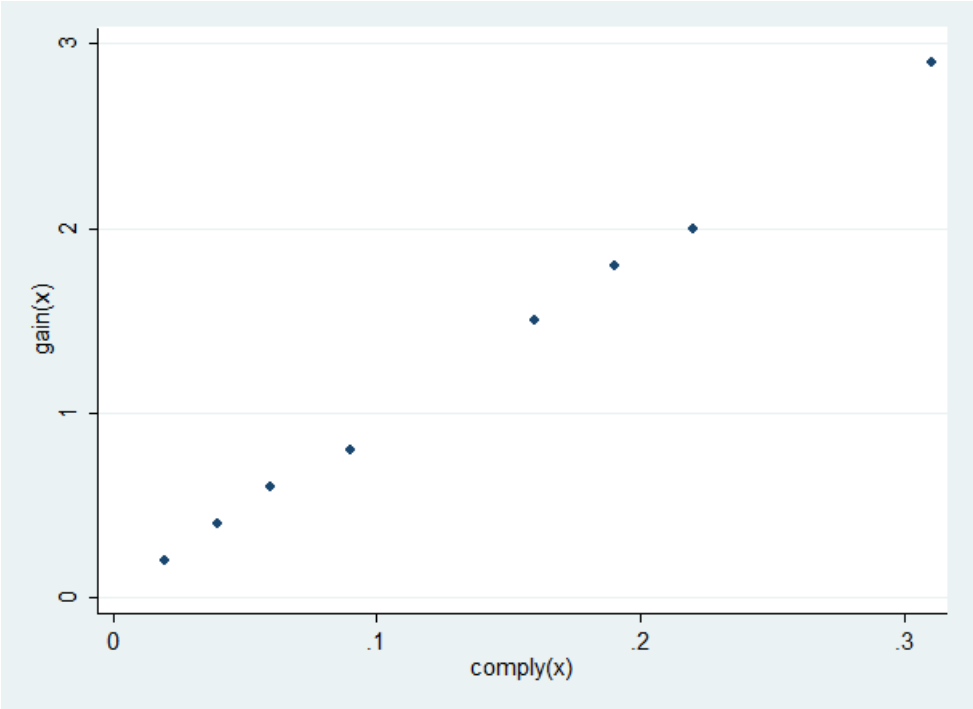

Figure 5: Sample comply-gain diagram for a likely valid instrument.

pattern of gain(x) when comply(x) $\neq 0$ will lend itself to interpretations involving treatment effects. Second, working with entire distributions can be quite burdensome with limited data.

Note that the theorem and corollary each rely on both the assumptions of validity and of no defiers. However, when x is a no-relevance point at which $E(D \mid X = x, Z) = 1$ or 0, defiers must be measure zero at x as a consequence of validity and therefore we only need to assume validity. So testing for violations of the theorem or corollary will generally be a joint test of validity and no defiers, but solely a test of validity in these special cases. (Of course, these "special cases" might be the most common no-relevance points in practice, which is why I refer to the approach in this paper as testing validity.)

Suppose now that we observed the true values of gain(x) and comply(x) for a number of x, and we plotted them on a graph, as in Figure 5. Corollary 1 requires that, if the instrument is valid, then any intercept with the gain axis must occur at the origin. If we saw values such as the values in Figure 5, we might be inclined to believe that, if we did observe some x with comply(x) $= 0$, that gain(x) would also equal zero – supporting the theory that the instrument is valid. However, if we saw values such as the values in Figure 6, we would most likely be disinclined to believe in the instrument's validity.

Is this intuition sensible? First, let's understand what this graph captures. Take the point corresponding to some x. The slope of the line connecting that point to

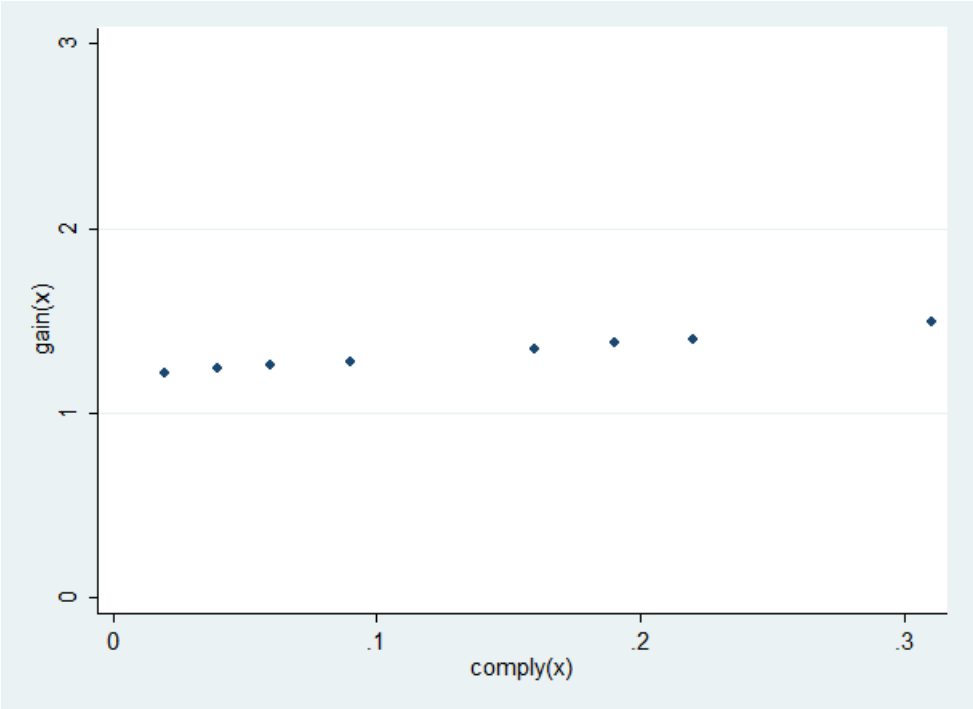

Figure 6: Sample comply-gain diagram for a likely invalid instrument.

the origin is gain(x)/comply(x). Under the assumptions of validity and no defiers, this slope is the average treatment effect on compliers at that value of x. If the average treatment effect on compliers at x is not strongly related to comply(x), then we might expect the points we observe to lie close to some line through the origin, and closer to that line as we draw closer to the origin. We should only encounter a graph like Figure 6 with a valid instrument when the treatment effects on compliers coincidentally become extremely large as comply(x) becomes small.

I suggest three approaches to testing validity, in order of the weakness of the required assumptions. The first is to observe a no-relevance point. As discussed before, this is simply placebo testing. The second is to suppose that, at least for some subsets of the population, there exists a relationship between comply(x) and gain(x) which satisfies conditions for nonparametric estimation. Local linear estimation is likely to be an appropriate approach because of its performance on boundaries. Then we fit the observed values of comply and gain to estimate what value of gain(x) would be expected if x were a no-relevance point. The third approach is to assume a parametric relationship between the comply and gain functions. A simple example is to assume that the average treatment effect on compliers at x is unrelated to comply(x) (which holds, for instance, if treatment effects are homogeneous in the population).

The first approach is trivial; conditional on observing a no-relevance point, we can test the claim of Theorem 1 with a Kolmogorov-Smirnov test, or we could test

Corollary 1 with a t-test for the equality of means. The next two subsections will briefly discuss the second and third approaches outlined above, starting with the parametric approach. Because the comply and gain functions must be estimated from data, error in this estimation will be a concern in finite samples for the second and third approaches. This estimation error will be discussed after the initial discussion of the two approaches.

Before proceeding any further, though, one point deserves brief additional discussion. As pointed out before, the theorem and the corollary both give conditions which must be true if Z is a successful instrument *conditional on* X. However, it is not always the case that conditioning on a covariate will preserve the validity of Z. In particular, we should be cautious in conditioning on any X which is caused by Z. As an illustration, suppose that, in testing the proximity to college instrument for college attendance, students who do not apply to college do not attend. Then we could condition on whether students applied to college, and we would observe that the instrument (proximity to college) is unrelated to treatment status (college attendance) among those students who do not apply, so that we would have a no-relevance point. But if the instrument is valid unconditional on application status, it is unlikely to be valid conditional on application status, since some of the people who do not live near a college and choose not to apply to college might be compliers, and would have a different realization of the covariate (i.e., they would apply to college) if they had a different realization of the instrument. These compliers presumably have higher ability than the never takers, and therefore, under the theory of the instrument's unconditional validity, we would expect students who did not apply to college to have higher earnings if they did not grow up near a college, yielding a negative value of gain(x) at the no-relevance point.

With that disclaimer out of the way, let us proceed to developing approaches.

3.1 Parametric testing

Suppose that we do not observe a no-relevance point but we do observe a collection of values of X which produce different values of the comply function. We wish to predict what would happen if we observed a no-relevance point. To estimate this relationship, we will fit a curve to the collection of measured values of comply and gain, and we will test for whether the curve intersects the gain axis at the origin. The true curve we will try to estimate can be described by the function g, defined such that, drawing an x at random,

$$E\left[\text{gain}(x) \mid \text{comply}(x)\right] = g(\text{comply}(x)) \ a.s.$$

This approach assumes that comply(X) is a random variable. Our test, from Corollary 1, will be whether or not $g(0) = 0$. (This means 0 should be contained in the support of comply(X), so that $g(0)$ is defined, even if we happen not to observe any no-relevance x.) I will outline an infeasible estimator but I use the language of expectations for two reasons. First, there might be multiple values of gain(x)

attained at one value of comply(x). For instance, proximity to college might have no impact on college attendance for both very high IQ and very low IQ individuals, so that comply(x) $= 0$ for both subpopulations, but perhaps low IQ individuals earn 10% more by living near a college while high IQ individuals earn only 5% more. Second, I use expectations rather than deterministic values of gain(x) to acknowledge that we must ultimately implement a feasible estimator.[7]

Nonparametric estimation of g is in some sense agnostic but is unnecessary if we already know the true functional form of g. The simplest example is if treatment effects are homogeneous. If the treatment effect $\beta_i = Y_i(1) - Y_i(0) = \beta$ for all i, then for all x, under the instrument assumptions, gain(x) $= \beta$ comply(x). In fact, the same functional form, $g(\text{comply}(x)) = \beta$ comply(x), will occur under the instrument assumptions whenever the treatment effect on compliers at x is independent of comply(x).

A simple test of instrument validity, then, is to estimate the line of fit between the gain and comply functions allowing for a constant term, and then test whether the constant term is equal to zero. That is, we take a collection of values of X for which we observe gain and comply, and then estimate the regression

$$E\left[\text{gain}(x) \mid \text{comply}(x)\right] = \alpha_1 + \alpha_2 \text{ comply}(x)$$

Under the null hypothesis of validity, $\alpha_1 = 0$. Under the alternative hypothesis of invalidity, α_1 will not in general equal zero.[8] One simple approach is to estimate α_1 using ordinary least squares regression. If standard OLS assumptions hold, $\widehat{\alpha_1}$ (our estimate of α_1) would follow the usual distribution of a constant term in OLS with n observations, and confidence intervals can be estimated using standard statistical packages.

Of course, we will observe estimates $\widehat{\text{comply}}(x)$ and $\widehat{\text{gain}}(x)$ rather than their population analogues. A standard set of OLS assumptions, adapted to the problem of fitting a relationship between $\widehat{\text{comply}}(x)$ and $\widehat{\text{gain}}(x)$, is the following:

OLS 1 Let $\varepsilon \equiv \widehat{\text{gain}}(x) - \alpha_1 - \alpha_2 \widehat{\text{comply}}(x)$. Then $E(\varepsilon \mid \text{comply}(x)) = 0$.

[7] If effects are smooth, so that we might have every reason to believe that the true values of gain lie exactly on g, and if we need that gain(x) should be exactly equal to zero when comply(x) $= 0$ (not just equal to zero on average), why should we use the notation of expectation? Because estimation error in comply and gain can be thought of as introducing noise into the observed relationship between comply and gain. The issue that we might be tricked into believing an instrument is valid when gain(x) is symmetric around zero for several no-relevance x's can be largely resolved by allowing for different g functions on different subsets of the data and jointly testing that $g(0) = 0$ for each subset.

[8] It is theoretically possible to construct a case in which there is a zero intercept despite invalidity. For example, it might be that the direct effect of Z on Y is proportional to comply(X). There are additional knife edge cases, but it seems to me that the simplest rule of thumb for intuiting whether this approach is likely to have power is to ask whether suspected invalidity is likely to approach zero as the relevance approaches zero.

OLS 2 We observe more than one value of $\widehat{\text{comply}}(x)$.

OLS 3 Observations of $\left(\widehat{\text{comply}}(x), \widehat{\text{gain}}(x)\right)$ are i.i.d.

OLS 2 is presumably satisfied, or we wouldn't be pursuing an approach based on variation in relevance. OLS 1 is possibly problematic in finite samples, however, because estimation error in comply(x) is potentially correlated with estimation error in gain(x), for reasons which will be discussed shortly. Similarly, to estimate appropriate standard errors for $\widehat{\alpha}_1$, we must estimate the comply and gain functions in a way which preserves OLS 3. A more complete discussion of appropriate assumptions accompanies the discussion of estimation error.

It is easy to see how this approach could be generalized. Suppose we imagine that the true relationship between comply and $E(\text{gain})$ is characterized by a function h_θ with some parameters θ, with $h(0;\theta) = 0$. Then we estimate θ and α in the equation

$$E\left[\text{gain}(x) \mid \text{comply}(x)\right] = \alpha + h(\text{comply}(x);\theta)$$

and test whether $\alpha = 0$.

3.2 Nonparametric testing

In principle, we would like to impose as few restrictions on g as possible. We can do so with nonparametric estimation. Suppose that g is continuous on the interval $[0, \epsilon)$ for some $\epsilon > 0$. Then it follows from the corollary that, if the instrument assumptions and finite expectations are satisfied, then

$$\lim_{c \downarrow 0} E\left[\text{gain}(x) \mid \text{comply}(x) = c\right] = 0$$

Because we are interested in estimating a boundary value, local linear estimation is likely to be appropriate. The most significant assumption that local polynomial requires for our purposes is that, for local polynomial estimation of degree p, we require that g be $p+1$ times continuously differentiable in comply(x). Therefore, for local linear estimation, we need that g must be twice differentiable.

This assumption may sound innocuous, but it is likely to have bite in some applications. Consider, for example, if there are two subsets of the domain of X, A and B. Treatment effects are large for individuals in A relative to treatment effects in B. Suppose in addition that the lowest observed values of the comply function for A are above the lowest values for B. Then we may find ourselves in a situation like the one illustrated in Figure 9, where g is discontinuous at the lowest value of comply in A.

When we encounter this situation, we can still attempt estimation by considering subsets of the data individually. It might not be the case that our data as a whole fulfills the local polynomial assumptions, but we may be able to divide our data into subsets such that, instead of a single function g satisfying the appropriate conditions

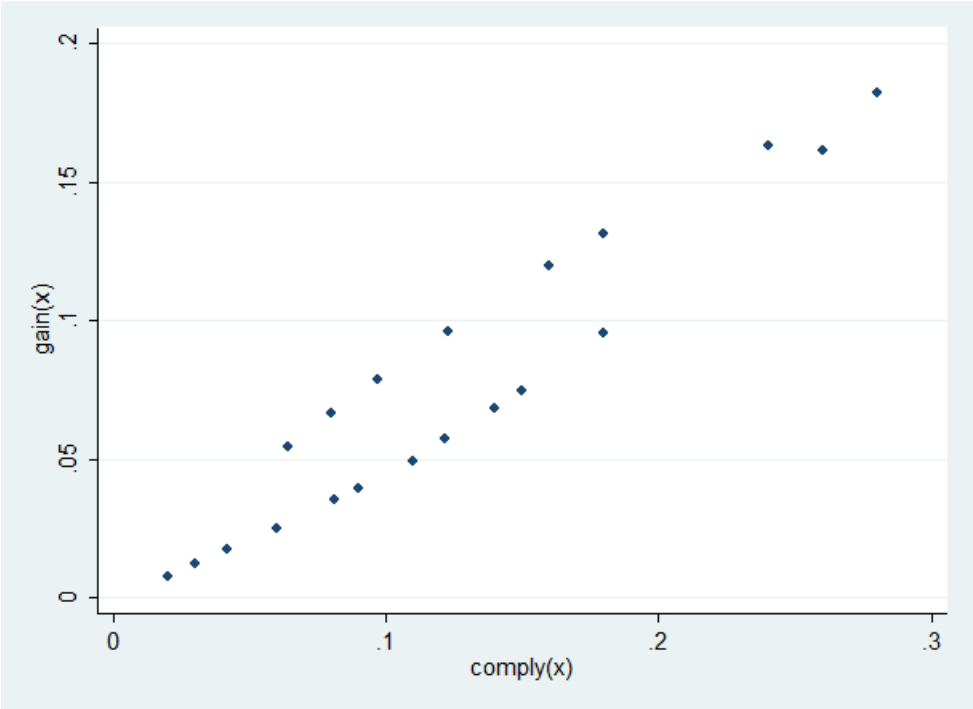

Figure 7: Sample comply-gain diagram with discontinuous function g.

for the enter dataset, we instead have functions g_j for each subset j which satisfies local polynomial conditions. Then we can perform a joint test that $g_j(0) = 0$ for each such subset j. Naturally, we could follow the same approach if a parametric assumption were appropriate only for subsets of the data.

The other significant assumption is that draws of $\left(\widehat{\text{comply}}(x), \widehat{\text{gain}}(x)\right)$ should be i.i.d. This consideration also crops up in the parametric case, and risks not being satisfied if we pursue an inappropriate approach to estimating the comply and gain functions.

3.3 Estimation error

The estimators described in the previous two sections were infeasible because the true values of the comply and gain functions are not known for each x. Estimating these functions from the data poses some challenges, though, because our errors in estimating one function are likely to be correlated with our errors in estimating the other. This is analogous to the problem that two-stage least squares has a central tendency in the direction of OLS (Nagar 1959) and produces a similar problem of bias.

The intuition is the same as the intuition behind why IV is biased in the direction of OLS. Presumably the error term in the first stage equation is correlated with the error term in the second stage equation, or we wouldn't need an instrument.

The randomness of realizations of these error terms in the sample are the source of estimation error in finite samples. Because the error terms are correlated, then the estimation errors will have some correlation as well.

Estimation errors are a potential issue for the two approaches outlined before. Provided we observe the true values of the comply and gain functions, the desirable properties of the approaches outlined in the previous two parts follow easily from the existing literature on regression and nonparametric regression. But using incorrect values for the comply and gain functions will lead to incorrect size for validity tests, as well as biased estimates of the direct effects of Z for use in the set identification approach of the next section. For example, in the case of homogeneous effects, in which we wished to estimate g using OLS, our parameter estimates will be biased because the error term in the gain equation is correlated with $\widehat{\text{comply}}(x)$.

As an additional cautionary note, estimating comply(x) and gain(x) using data not at x introduces an additional complication: Errors in our estimate of, say, comply(x) will be correlated with errors in our estimate of comply(x') for x' near x! In the context of the validity test I've outlined, this becomes a violation of the i.i.d. assumption used in regression.

Preliminary Monte Carlo evidence suggests that these concerns are ignorable in ideal cases but become important in small samples when the correlation in error terms is strong and the covariate X is not a particularly powerful predictor of treatment status.

Perhaps future analysis will uncover a simple criteria for judging whether estimation error is likely to be serious problem. In the meantime, one technique for getting around these issues is to use an approach parallel to the split-sample instrumental variables (SSIV) approach outlined by Angrist and Krueger (1995). Because we wish to use the values of the comply and gain functions in a regression of gain(x) on comply(x), our core problem is that the estimation errors are correlated, for the reasons described above. If estimation errors were uncorrelated, then the problem for the regression of gain on comply would only be attenuation bias. But the correlation in estimation errors can be broken by splitting the sample in half at each x, using one half to estimate comply(x) and the other half to estimate gain(x). Because we can compute standard errors for the estimation of comply(x), the degree of noise in our estimate of the comply function can be explicitly measured and we can adjust appropriately for attenuation bias using error-in-variables regression with the appropriate reliability ratio. Noise in our estimates of gain(x) will not bias the regression results, though such noise does reduce the power of the test to detect invalid instruments. By estimating the comply and gain functions separately at each value of X, we also avoid violation of the i.i.d. assumption.

This approach will produce a test with correct size, but I have not shown that this is the optimal approach to estimating the comply and gain functions.[9] If data at

[9]In fact, I would suppose it is not; it seems inefficient to throw out half of your data for each step of the estimation.

each x is sparse, it becomes practical to group multiple values of X together so that standard errors are not excessive at each X. At this time, I do not have guidance for how this grouping should be accomplished. Future work may uncover a more efficient approach to estimating the comply and gain functions so that tests of validity based on the comply-gain framework can have maximal power conditional on size.

3.4 Commentary

I have outlined three approaches to testing an instrument's validity: Finding a no-relevance point, guessing what would happen at a no-relevance point without parametric assumptions, and guessing what would happen at a no-relevance point using parametric assumptions. What are the relative strengths and weaknesses of these three approaches?

Finding a no-relevance point is potentially the most credible. However, it needs to be the case that the point is actually no-relevance. Suppose we find some x such that our estimate $\widehat{\text{comply}}(x) = 0$ when, in reality, $\text{comply}(x) = k > 0$. For the reasons outlined in the previous section, we might expect that $\bar{u}_1(x) - \bar{u}_0(x) \neq 0$, so that we would estimate $\text{gain}(x)$ to be non-zero under validity. So we are not free from concerns about estimation error in this case. Claims based on testing at a no-relevance points will be strongest when we have strong theoretical reasons to believe that $\text{comply}(x) = 0$. Finding such points will frequently be a challenge in empirical applications.

The second approach was nonparametric estimation of the g function describing the comply-gain relationship. The primary advantage of this approach is the weak requirement on treatment effects. It suffices that treatment effects are continuous over $\text{comply}(x)$ for some subsets of the data. However, the test will have low power unless there are some points at which the instrument induces little compliance. That is, we may not need covariates so strong that they can completely shut off the relevance, but we do require that our covariates should be quite strong.

The third approach was parametric estimation of the g function. The parametric assumption can give the test some power even when the covariates are not so strong that they can nearly shut off the relevance. Furthermore, a parametric assumption may be more desirable than nonparametric estimation when the data is sparse and noisy. This would occur, for example, with small samples. However, it may be easy to disbelieve a parametric assumption in many applications. It is difficult to imagine applications where, should a linear specification for the g function appear to be inappropriate to the data, another parametric assumption would be satisfying.

An advantage of the comply-gain framework is that it allows for the presentation of visual evidence about an instrument's validity. In the same way that researchers employing a regression discontinuity design can easily demonstrate the robustness of their findings to reasonable alternate specifications using a graph, so too is it possible for researchers with appropriate data and adequate sample size to demonstrate with a graph that validity appears to hold (or not to hold) under any reasonable functional

form assumptions.

This paper focuses on the case of binary instruments. However, it is easy to see how the framework can be extended to continuous instruments. With continuous Z, comply(x) becomes the partial of D with respect to Z for the population at $X = x$, and gain(x) becomes the partial of Y with respect to Z at $X = x$. I focus on the binary case in this paper for simplicity.

4 Identification with invalid instruments

Imagine that we have discovered using the techniques from the previous section that our instrument is invalid. This might be disappointing, but it may still possible to recover some information about treatment effects. This section describes conditions under which set identification of local average treatment effects is still possible, and compares these conditions to a selection model approach.

Suppose that our proposed instrument is found to be invalid, such that we can write the true (structural) outcome equation in the following way:

$$Y_i = g(X_i) + \beta_i D_i + h(X_i)Z_i + \epsilon_i$$

with $h(X_i) \neq 0$ and ϵ uncorrelated with the other terms (including Z) by definition. We wish to identify the average value of β_i across the population of compliers. Suppose we knew $h(X_i)$. Define $Y_i^* \equiv Y_i - h(X_i)Z_i$. Then we could re-write the equation as

$$Y_i^* = g(X_i) + \beta_i D_i + \epsilon_i$$

while preserving the desirable property that the left-hand side is observed for each i. Then we can use Z as an instrument for the effect of D on Y^*, and we will still satisfy the 2SLS requirement that Z be mean-independent of the error term ϵ. Since the causal effect of D on Y^* is the same as the effect of D on Y, this means we can identify the LATE of interest.

The challenge, then, is to identify $h(X_i)$. One approach would be to assume that $h(X_i) = k$ for some constant k, and that our task is to identify k. Then each of the approaches outlined in the previous section can identify k. (If the instrument is valid, then $h(X_i) = 0$.)

This assumption of constant invalidity has sometimes been used in practice. As mentioned before, difference-in-differences estimation can be thought of as using time period as the covariate X, with the pre-treatment period as the no-relevance x. Then the common trends assumption is an assumption that h takes the same value for pre- and post-treatment periods.

However, assuming that h takes a single value is clearly restrictive in certain ways. Consider the example of the proximity to college instrument. Suppose the labor market is segmented between skilled and unskilled workers, and suppose that,

because of moving costs, the higher probability of attending college if you grew up near a college leads there to be an unusually high proportion of skilled workers (and an unusually low proportion of unskilled workers) in the places where the students who grew up near a college live in adulthood. By simple supply and demand,[10] this should depress wages for the always takers and improve wages for the never takers when they grew up near a college relative to if they had not. Therefore, we would expect to find a different value of h if we estimated it by allowing the measure of always takers to go to 1 rather than estimating it by allowing the measure of never takers to go to 1.

If h is not constant, can we say anything? In fact, all we need to know is $\bar{h} \equiv E\left[h(X_i) \mid Z = 1\right]$. Define $Y_i' \equiv Y_i - \bar{h}Z_i$. Then

$$Y_i' = g(X_i) + \beta_i D_i + \left[h(X_i) - \bar{h}\right] Z_i + \epsilon_i$$

Because the function h is unknown, the term $\left[h(X_i) - \bar{h}\right] Z_i$ is unobserved, and therefore would be part of the error term if we estimated this equation. Then we can estimate the LATE in the above equation using Z as an instrument provided that Z is uncorrelated with $h(X_i)Z_i - \bar{h}Z_i + \epsilon_i$. Some algebra confirms that this is the case. (See Appendix A.) Once again, since the treatment effect of D on Y' is the same as the treatment effect of D on Y, we can therefore identify the LATE of interest if we know $\bar{h}$.

But if we don't know h, why would we expect to know $\bar{h}$? We might not know exactly the true value of $\bar{h}$, but a reasonable assumption might be that our observations allow us to bound $\bar{h}$. For instance, in the returns to schooling example, it might be reasonable to suppose that the invalidity for the population as a whole lies somewhere in between the invalidity for the high-IQ, high-income always takers and the low-IQ, low-income never takers. This assumption might make sense, for example, if we think of the invalidity among always takers as reflecting a single wage premium for skilled workers to living in an area with a college, the invalidity among never takers as reflecting a single wage premium for unskilled workers, and we imagine that the population of individuals living near colleges are some mixture of skilled and unskilled workers.[11] Or it may simply be intuitive that the compliers in this case seem to be an intermediate case between the always takers and the never takers.

Suppose we use the techniques of the previous section to come up with an upper and a lower bound for $\bar{h}$ by estimating h_J and h_K, the invalidity from subsets of the data J and K. Assume without loss of generality that $h_J > h_K$. Then let $\underline{Y}_i \equiv Y_i - h_J Z_i$, and let $\bar{Y}_i \equiv Y_i - h_K Z_i$. Provided we observe h_J and h_K, then $\bar{Y}$ and $\underline{Y}$ are observed too. So we can estimate the LATE for D on $\bar{Y}$, which we can call

[10] For arguments that this kind of supply and demand framework applies, see, for instance, Goldin and Katz 2008.

[11] Of course, this example model is more specific than we need, and would yield a point estimate of $\bar{h}$ when all we require here is a set containing the true $\bar{h}$.

$\bar{\beta}$, by treating Z as a valid instrument for D. We can also estimate another LATE – call it $\underline{\beta}$ – by treating Z as a valid instrument for the effect of D on $\underline{Y}$. Then the true LATE of interest, $\beta = E\left[Y_i(1) = Y_i(0) \mid i \text{ complier}\right]$ lies between $\bar{\beta}$ and $\underline{\beta}$.[12]

The bounds obtained in this way will vary in their width. In general, we will find narrower bounds when (i) the invalidity is similar in our bounding subsets J and K, (ii) our covariates are powerful enough to give accurate (low standard error) estimates of h_J and h_K, and (iii) when the true reduced form effect is large, so that the relationship between Z and Y which occurs through D is large relative to our uncertainty about the true value of $\bar{h}$.

4.1 Comparison with selection models

As mentioned before, selection models can potentially identify a causal effect of D on Y using invalid instruments, i.e. any sources of variation in D which also have a direct impact on Y. For example, the following model is identified under the assumption of joint normality of the error terms u and v:

$$Y = Z\delta + \beta D + u$$
$$D = 1\left\{Z\gamma + v > 0\right\}$$

However, it's important to note that the identification of β is coming from both functional form (linearity) and distributional assumptions. In practice, we have little way of knowing whether it is reasonable to believe that u and v are jointly normal. Functional form assumptions are particularly problematic in the second equation, because the underlying variable determining D is not observed. Frequently, the only uncontroversial assumption to be made about the density of v is that it is low in the tails.

The assumptions required for the set identification approach can be satisfied by a similar selection model. We can extend the set identification model to include an equation describing selection for D:

$$Y_i = g(X_i) + \beta_i D_i + h(X_i)Z_i + \epsilon_i$$
$$D_i = 1\left[f(X_i) + Z_i + \xi_i > 0\right]$$

Here, f can be any function. We may have little idea if ξ should be normal or not (it would be unlikely to be standard normal, since we have normalized the scale by setting the coefficient of Z equal to one) but perhaps we feel confident that the distribution of ξ will have low density in the tails. This is a weakening of the assumption of normality. Suppose in addition that f takes large values for large X and small values for small X. This is a weakening of a linearity assumption. Finally, assume that h is monotonic in X. This is weaker than allowing Z and an interaction of Z with X to appear in the outcome equation. Then we can approach groups

[12] This statement follows from the result of Choi and Lee 2012.

of always takers (never takers) by taking X in the direction of positive (negative) infinity. This allows us to estimate an upper and a lower bound for $\bar{h}$. Therefore we can set identify the LATE using the approach outlined earlier, and using only very weak assumptions on the distribution of ξ and on the functional forms of both equations.

In this sense, the set identification approach can often be equivalent to a selection model with very weak assumptions. Of course, the example outlined above is not the only setup which would satisfy the assumptions required for the set identification result; it is a sufficient but not a necessary model.[13] I include it to demonstrate that, conditional on being willing to accept the most basic functional form restrictions, like linearity, then the set identification approach can give credible estimates of treatment effects.

5 Application to schooling returns

Card (1995) uses proximity to college as an instrument for the effect of college attendance on wages. I will first review the data and research design used by Card. I will then test the exogeneity of college proximity using each of the three approaches outlined before: finding a no-relevance point, parametric estimation of the g function, and nonparametric estimation. Each of these approaches rejects the validity of the Card instrument. I then implement my set identification approach, estimating bounds of invalidity from the populations of students least and most likely to attend college. These bounds are wide, but they are suggestive that the true returns to college lie somewhere in the range from zero to the estimates derived from OLS.

5.1 Data

Card uses data from the National Longitudinal Survey of Young Men (NLSYM). The NLSYM began with a sample of 5525 American men between the ages of 14 and 24 in 1966 and periodically surveyed participants until 1981. There is some oversampling of minorities and southerners. He collects data on wages and completed education from the 1976 survey, at which time respondents are between the ages of 24 and 34. 29 percent of the original sample dropped out before the 1976 survey and 83 percent of the sample in 1976 reports a valid wage. For further descriptive statistics, including those comparing the 1976 sample to the original population, see Card's

[13] As an example of a model which satisfies the set identification assumptions but not the selection model assumptions, imagine a diff-in-diffs approach in which we observe a difference between the outcomes of treatment and control groups both before and after a period in which a policy is implemented, assuming that the policy has no lasting effect. X would be time period and Z would be assignment to the treatment group, and we would have two no-relevance values of X – the period before the experiment and the period after. Then it doesn't make sense to use a standard discrete choice model for the determination of D, as we could write D as being strictly determined by X and Z, with no error term.

paper. I use data taken directly from Card's website.[14] Frequencies of each level of education are shown in Table 1 to give the reader some sense of the typical levels of education in the sample. Conditional on completing a year of education beyond high school, individuals complete an average of 15.4 years of schooling. Individuals who do not complete at least 13 years of schooling complete an average of 11.1.

Card defines the key variables as follows: The outcome, Y, is the log of wage in 1976. The treatment, D, is the number of years of education reported in 1976. Note that this is not binary. In my own analysis, I will use some specifications where D is allowed to be continuous, but I will primarily define D to be a binary variable which is equal to one for individuals who report any years of education completed beyond 12th grade.[15] The instrument Z is equal to one if there is an accredited 4-year college in the respondent's local labor market area in 1966. Card uses additional instruments which I drop from my analysis.[16]

Several variables are available for use as covariates X in my model. There are two measures of cognitive ability: Knowledge of the World of Work (KWW) scores, which are measured in 1966, and a measure of IQ from school records, taken in 1966 but only available for a subset of respondents. There are also measures of parental education, taken in 1966, some variables describing family conditions during the respondent's youth, and indicators for the region where the respondent lived in 1966. I select the measure of IQ for the primary analysis because it generates significant variation in relevance and because Griliches (1977) suggests it is more reliable than KWW.

Card estimates several models. His baseline OLS result shows a return to schooling of approximately 7% per year of schooling, controlling for experience and experience squared, living in the south in 1966, race, and residence in a metropolitan statistical area (SMSA) in 1976. I replicate this specification and two others in Table 2, as well as the same specifications with a dummy for college attendance as the treatment variable. The estimates decline somewhat with this set of controls, but substantial differences persist.

Card also estimates instrumental variables models using the presence of a 4-year college and the presence of a 2-year college in the local labor market in 1966 as

[14]Card has generously made his cleaned data available at `http://davidcard.berkeley.edu/data_sets.html`.

[15]I am not alone in making a binary variable out of Card's continuous variable. Kitagawa (2008) also constructs a binary variable for college attendance, though Kitagawa defines the variable equal to one only for those students who complete at least four years of education beyond high school. Kitagawa's definition is actually problematic; imagine a student who completes high school if he does not live near a college and who drops out of college if he does. Then this student will be a never taker by Kitagawa's definition, yet we would expect this student to earn higher wages if he lives near a college even if college proximity affects wage only through years of education, as required for Card's instrument. I encounter a parallel problem for always takers under my definition of D, so I resort to the continuous case as a robustness check.

[16]Card also uses proximity to 2-year colleges as an instrument. I do not include this instrument in my analysis because I find it is no longer statistically significant at conventional levels in the first stage equation once IQ and proximity to a 4-year college are included. Card also uses an interaction term between proximity and parental education as a robustness check, which I omit for simplicity.

Years of Education	No.	Cumulative %
0	3	0.1
1	2	0.1
2	2	0.2
3	4	0.3
4	6	0.5
5	13	0.8
6	22	1.4
7	40	2.5
8	90	5.0
9	91	7.6
10	146	11.6
11	193	16.9
12	1,194	50.0
13	323	58.9
14	309	67.5
15	206	73.2
16	532	87.9
17	179	92.9
18	258	100.0
Total	3,613.0	

Table 1: Frequency Table: Years of Education

Variables	(1) lwage76	(2) lwage76	(3) lwage76	(4) lwage76	(5) lwage76	(6) lwage76
ed76	0.0932***	0.0743***	0.0691***			
	(0.00368)	(0.00366)	(0.00509)			
college				0.302***	0.231***	0.171***
				(0.0182)	(0.0174)	(0.0208)
exper76	0.0898***	0.0845***	0.0940***	0.0853***	0.0803***	0.0730***
	(0.00708)	(0.00674)	(0.00923)	(0.00747)	(0.00707)	(0.00928)
expersq76	-0.00249***	-0.00229***	-0.00270***	-0.00326***	-0.00285***	-0.00210***
	(0.000341)	(0.000319)	(0.000467)	(0.000356)	(0.000330)	(0.000473)
south66		-0.0975***	-0.0593***		-0.125***	-0.0614***
		(0.0159)	(0.0189)		(0.0163)	(0.0194)
black		-0.190***	-0.139***		-0.224***	-0.136***
		(0.0182)	(0.0274)		(0.0187)	(0.0280)
smsa76r		0.168***	0.156***		0.185***	0.160***
		(0.0152)	(0.0186)		(0.0158)	(0.0191)
IQ			0.00256***			0.00453***
			(0.000757)			(0.000756)
Constant	4.469***	4.712***	4.471***	5.666***	5.679***	5.254***
	(0.0703)	(0.0703)	(0.110)	(0.0403)	(0.0405)	(0.0966)
Observations	3,010	3,010	2,061	3,010	3,010	2,061
R-squared	0.196	0.284	0.222	0.096	0.224	0.173

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 2: OLS results

instruments for schooling, finding returns to an additional year of education which range between 12 and 14%, depending on the exact specification.

I replicate a version of Card's results using only the presence of a 4-year college first as an instrument for (binary) college attendance, then as an instrument for total years of education. IQ is included as a control in the 2SLS specification since it will be used to generate variation in relevance, which is a test of the validity of college proximity *conditional on IQ*. Also shown are the results of specifications with a number of covariates (including experience and squared experience, residence in a metropolitan statistical area in 1976, race, and region of residence in 1966) as Card prefers specifications with these covariates included. The results are shown in Table 3. Recall that my choice of instruments is lightly altered from Card's.[17]

Because of the small sample (only 2470 individuals have an observation of IQ) and because of the curse of dimensionality and the importance of keeping estimation error manageable, I am not currently able to meaningfully test the validity of the instrument using the full set of covariates to induce variation in relevance. For now, we focus on the validity of the instrument solely conditional on IQ. While this case does not capture Card's preferred specification, it is nonetheless somewhat informative about the likely size of treatment effects. Subsequent versions of this paper will attempt to develop more efficient approaches to testing instrument validity in order to make it possible to test instrument validity conditional on a richer set of covariates.

5.2 Testing instrument validity

This section applies a variety of tests to determine whether the Card instrument is plausibly valid conditioning only on IQ. I begin by producing scatterplots of the relationship between the values of the comply and gain functions, using IQ to induce variation in the instrument's relevance. Visual inspection of the scatterplots is enough to induce skepticism about the instrument's validity. I then fit these points both parametrically and nonparametrically. In addition, I test a group where the instrument's relevance to the treatment of attending college is plausibly zero: high school dropouts. Because this is testing conditional on educational attainment, I will need to argue that conditioning on dropout status would not turn a valid instrument into an invalid one. The no-relevance point approach also suggests that the instrument is invalid.

First, the comply-gain scatterplot. Figure 8 shows a scatterplot, with each point representing a decile of IQ. The sample has been split within each decile, with half of the points randomly selected to estimate the comply function and the other half used to estimate the gain function. The standard errors of the estimates of the comply component range between .049 and .098, and the standard errors of the estimates of the gain component range between .067 and .105. Recall that we are interested in

[17] See previous footnote.

Variables	(1) lwage76	(2) lwage76	(3) lwage76	(4) lwage76	(5) lwage76	(6) lwage76	(7) lwage76	(8) lwage76
ed76	0.188***		0.333***		0.117*		0.108*	
	(0.0261)		(0.129)		(0.0677)		(0.0635)	
college		1.279***		1.401**		0.492		0.432
		(0.220)		(0.554)		(0.300)		(0.264)
IQ			-0.0193**	-0.0132*	0.000157	0.00176	0.000554	0.00217
			(0.00971)	(0.00756)	(0.00346)	(0.00272)	(0.00329)	(0.00249)
exper76					0.117***	0.0947***	0.113***	0.0914***
					(0.0335)	(0.0225)	(0.0317)	(0.0205)
expersq76					-0.00319***	-0.00227***	-0.00311***	-0.00227***
					(0.000852)	(0.000521)	(0.000819)	(0.000511)
smsa76r					0.146***	0.139***	0.139***	0.132***
					(0.0238)	(0.0277)	(0.0234)	(0.0269)
black					-0.141***	-0.138***	-0.149***	-0.147***
					(0.0282)	(0.0293)	(0.0285)	(0.0292)
south66					-0.0622***	-0.0732***	0.0574	0.0341
					(0.0196)	(0.0229)	(0.0497)	(0.0588)
Constant	3.767***	5.616***	3.673***	6.858***	3.905***	5.199***	3.933***	5.147***
	(0.347)	(0.112)	(0.815)	(0.453)	(0.801)	(0.112)	(0.743)	(0.107)
Region dummies	N	N	N	N	N	N	Y	Y
Observations	3,010	3,010	2,061	2,061	2,061	2,061	2,061	2,061
R-squared					0.186	0.081	0.209	0.124

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3: IV results

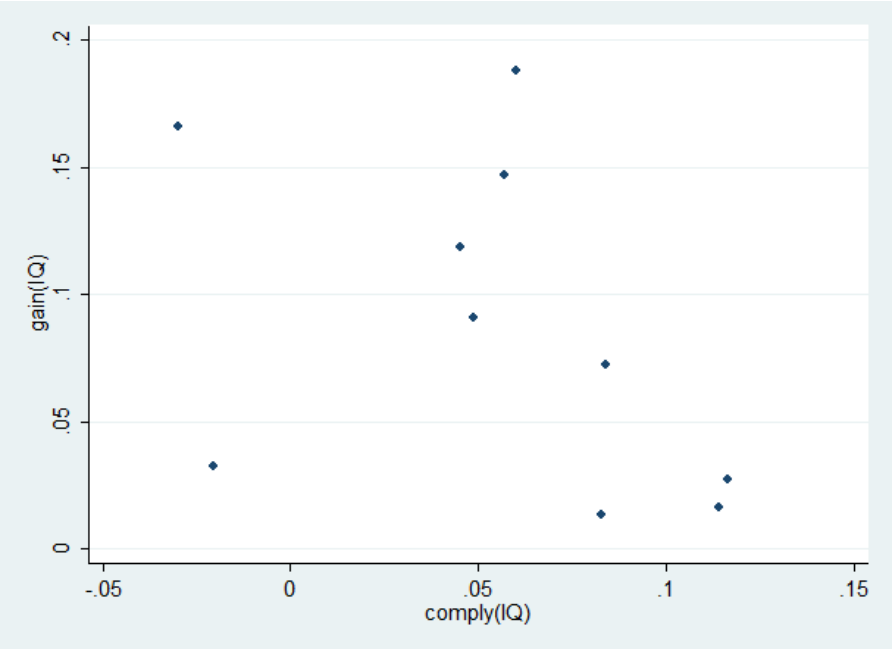

Figure 8: Comply-gain diagram by IQ decile

assessing whether, if some hypothetical true (i.e. population) point lay on the gain axis, it would in fact lie at the origin. Furthermore, if the instrument is valid, given the 2SLS results, we would expect to find that the points in the scatterplot appear to rise to the right, reflecting that, the more likely living near a college is to induce individuals to attend college, the higher wages of individuals who lived near a college in 1966 should be relative to their peers who live far from colleges. Yet in the data, the pattern of points has a slope close to zero (in fact, negative), and the intercept does not appear to lie at the origin.

Figure 9 shows the same diagram with a linear fit. This fit reflects the assumption that treatment effects are not higher at higher values of compliance. In fact, this assumption is likely to be violated, for reasons which I will discuss shortly. I assume classical error-in-variables to compute standard errors for coefficients in the linear fit case, using a reliability ratio of .33. The estimated intercept is .185, with a standard error of .053. The p-value for the null hypothesis that the intercept is zero is .008, allowing us to reject the null hypothesis at conventional significance levels.

Interestingly, the slope coefficient in this linear fit is negative, and we can even reject that it is positive at a 90% confidence level. Remarkably, that means that, if there is a single direct effect of proximity to college on wage, and if the treatment effect is unrelated to comply(x), then the effect of attending college on wages is likely to be *negative*. Of course, these assumptions are unlikely to hold, but they are in some sense not so much stronger than the assumptions which led to IV estimates larger than OLS estimates. The results are available in Table 4.

Figure 10 shows the scatterplot with a local linear fit. A nonparametric fit makes

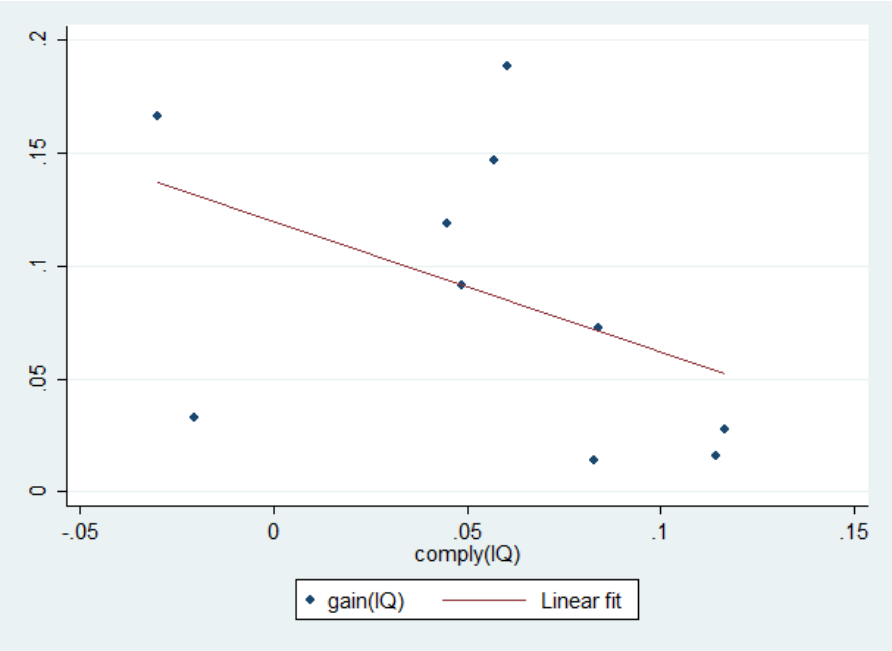

Figure 9: Comply-gain diagram with linear fit

little sense with this few points of support, so I omit hypothesis testing.

Of course, as discussed previously, we can only interpret these diagrams as evidence against validity if treatment effects are not getting arbitrarily large as comply(x) becomes small. In this case, there is a very simple reason to suspect that treatment effects are not identical at all levels of comply(x), which is that not all students who are induced to attend college by living near one wind up attending college for the same number of years. Figure 11 allows for intensive differences, testing validity under the assumption that log wages are linear in years of education by comparing the first stage (still notated as comply(x), though treatment is now continuous) to the reduced form effect (as gain(x)) at different deciles of IQ. The standard errors of the estimates of comply(IQ) range from .334 to .434. Using a reliability ratio of .5, the constant is estimated to be .089, with a standard error of .025. The p-value under the null hypothesis of a zero intercept is .007, once again allowing us to reject the null at conventional significance levels. The slope coefficient is still negative, but is quite close to zero. Results can be seen in Table 4.

Finally, we can assess the validity of the proximity to college instrument for the effect of attending college on wages conditioning not on IQ, but on completing no more than 11 years of education. This clearly represents a no-relevance point, since no person completing fewer than 12 years of education attends college. However, since the number of years of education completed is clearly a causal result of the instrument, this conditioning is suspicious. But under the theory of the instrument's validity, proximity to college increases years of education by inducing students to attend college, not by inducing them to complete high school. This assumes that the

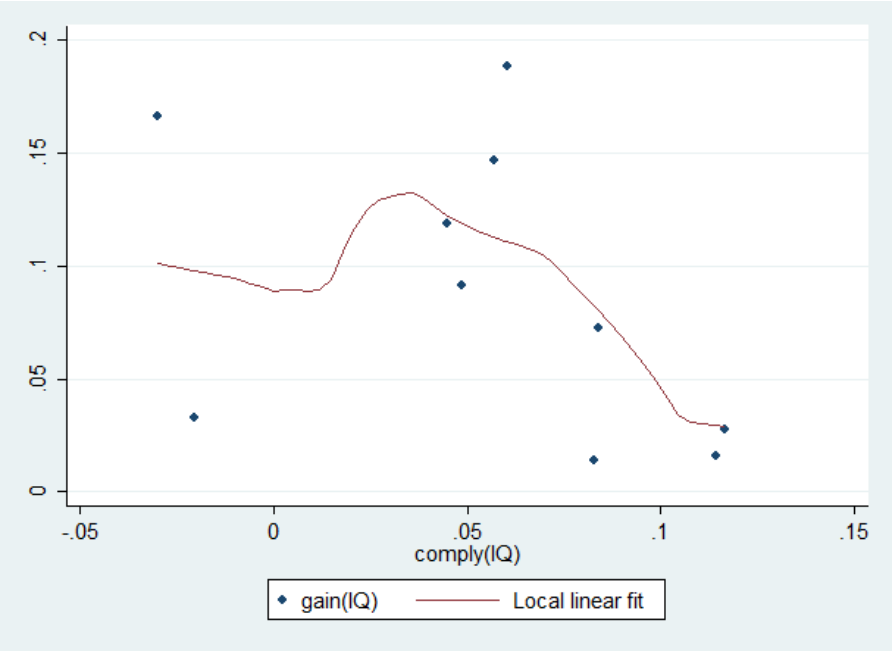

Figure 10: Comply-gain diagram with local linear fit

option value of completing high school in order to be able to attend a 4-year college is unlikely to convince students to complete high school; this is believable if students at the margin of dropping out of high school are unlikely to be at the margin of attending a 4-year college.

This comparison can be made with a simple comparison of means. Students completing 11 or fewer years of education who live near a 4-year college earn an average log wage of 6.076 in 1976 (standard error=.026). Their peers who do not live near a 4-year college earn an average log wage of 5.89 (standard error=.027). The t-statistic for the equality of these means is 5.05, and we reject that college proximity is a valid instrument for college attendance conditional on dropout status. Our point estimate of the difference is .189, with a standard error of .038. Intuition and further tests suggest that this discrepancy is unlikely to occur because of option value.[18]

The balance of this evidence is enough to suggest that, conditioning on IQ, proximity to a four year college is unlikely to be a valid instrument for attending college

[18]If option value induces students to complete more years of high school in areas near 4-year colleges, then the remaining students who drop out despite living near a college are likely to fall lower in the distribution of ability than students who drop out and do not live near a college. This would produce a negative, not the observed positive, wage premium for dropouts who lived near colleges in 1966 relative to their peers who did not live near a college.

As an additional robustness check to consider the possibility of contamination through option value, I performed the same test, conditioning on completing 10 or fewer and 9 or fewer years of education, and obtained even larger discrepancies.

Variables	(1) gain	(2) gain
comply	-1.744*	-0.0154
	(0.917)	(0.120)
Constant	0.185***	0.0889***
	(0.0531)	(0.0248)
D binary or cont.?	Binary	Cont.
Observations	10	10
R-squared	0.578	0.004

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: Error-in-variables estimates of parametric case

or for years of education. Additionally, conditioning on dropout status rather than IQ, proximity to a 4-year college is unlikely to be a valid instrument for attending college.

5.3 Set identification

The previous section offered evidence that the Card instrument is invalid. This section attempts to place bounds on the direct effect of attending college on log wages.

Because of general equilibrium effects, there are theoretical reasons to believe that the direct effect of proximity to a 4-year college on log wages will be the highest for unskilled workers and lowest for skilled workers. If this is the case, we can bound the invalidity by measuring direct effects for individuals with low and with high academic achievement, on the premise that individuals with low academic achievement are likely to become unskilled workers, and visa versa for high-achieving students.

To estimate a bound for unskilled workers, we can use the simple no-relevance point of high school dropouts. This yields an estimated direct impact of .189. To estimate the bound for skilled workers, I perform parametric estimation of invalidity with linear fit on the top half of the IQ distribution, obtaining an estimated intercept of .069.[19]

Applying these two numbers as bounds produces an estimate of the effects of attending college on log wages. Table 5 shows the IV estimates at the lower and upper bounds. These bounds include a wide range of values, including the possibility that the wage returns to college are zero. The bound estimated from $\bar{Y}$ still shows

[19]I broke observations in the top half of the IQ distribution into deciles by IQ within the top half of the distribution, and I used a reliability ratio of .25, following the same approach as before.

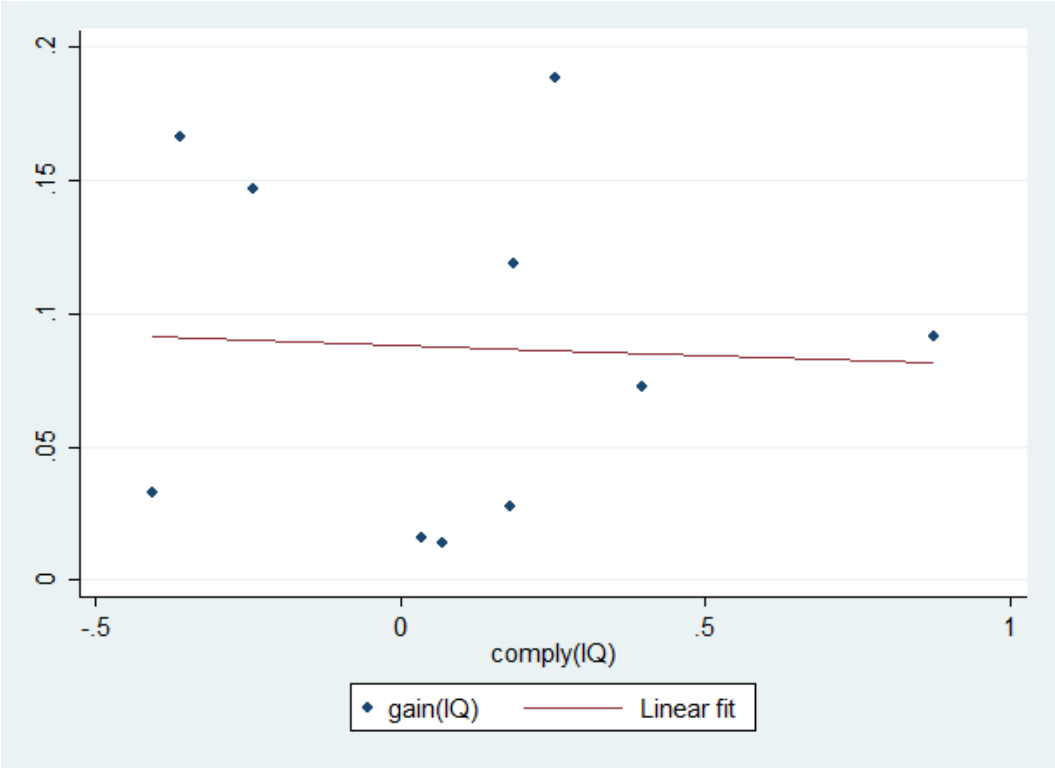

Figure 11: Comply-gain diagram with continuous treatment

a positive effect, but it falls below the effect estimated treating proximity to college as a valid instrument. An important caveat is that different results may arise if additional covariates are added. Because of the curse of dimensionality, I am unable to produce results with further controls at this time. These results will be updated once Monte Carlo simulations are completed to determine finite sample properties of potential approaches to including additional covariates. In the meantime, these empirical results should be taken to be suggestive rather than definitive.

6 Conclusion

This paper introduces a framework for testing the validity assumption used in instrumental variables estimation, treating a placebo test as the limit of not-quite placebo tests. Tests in this framework have power to detect invalid instruments when there are covariates which can induce variation in the instrument's relevance without causing the invalidity to go to zero as the instrument's relevance goes to zero. The requirement of variation in relevance is satisfied in many applications – for example, when treatment is binary and we observe covariates which are good predictors of treatment status. Both parametric and nonparametric approaches to testing are possible. Furthermore, when instruments are found to be invalid, point and set

Variables	(1) $\underline{Y}$	(2) $\bar{Y}$
college	-0.274*	0.716***
	(0.155)	(0.162)
Constant	6.271***	5.853***
	(0.0791)	(0.0824)
Observations	3,010	3,010

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5: Set identification bounds

identification of causal effects are possible under some additional assumptions.

The application to the Card instrument highlights both the value of the approach and some of its current shortcomings. On the positive side, it is possible to construct arguments suggesting that the instrument is invalid, including arguments which do not require a pure placebo test. On the negative side, though it turns out that the test is powerful enough in this case to reject the null with a single covariate, the limited size of the sample can be an obstacle to testing with a large number of covariates. A point of emphasis for future work will be to develop techniques based on this framework which make more efficient use of the available data to test the hypothesis of instrument validity.

References

[1] Altonji J.G., Elder T.E., and C.R. Taber (2005a). An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling, *Journal of Human Resources*, 40, 791-821.

[2] Altonji J.G., Elder T.E., and C.R. Taber (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools, *Journal of Political Economy*, 113, 151-184.

[3] Anderson, T. W. and H. Rubin (1949). Estimation of the Parameters of a Single Equation in a Complete Set of Stochastic Equations, *Annals of Mathematical Statistics*, 20, 46-68.

[4] Angrist, J.D. and A.B. Krueger (1995). Split-Sample Instrumental Variables Estimates of the Return to Schooling, *Journal of Business & Economic Statistics*, 13, 225-235.

[5] Ashenfelter O. and C. Rouse (1999). Schooling, Intelligence and Income in America: Cracks in the Bell Curve, NBER working paper, no. 6902.

[6] Caetano, M.C., Rothe C., and N. Yildiz (2013). Discontinuity Test of the Validity of an Instrumental Variable: A Control Function Approach, unpublished manuscript.

[7] Card, D. (1995). "Using Geographic Variation in College Proximity to Estimate the Return to Schooling". In L.N. Christofides, E.K. Grant, and R. Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*. Toronto: University of Toronto Press.

[8] Card, D. (1999). "The Causal Effect of Education on Earnings." In O. Ashenfelter and D. Card, editors, *Handbook of Labor Economics* Volume 3A. Amsterdam: Elsevier.

[9] Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistence Econometric Problems, *Econometrica*, 69, 1127-1160.

[10] Carneiro, P., Heckman, J.J., and E.J. Vytlacil (2011). Estimating Marginal Returns to Education, *American Economic Review*, 101, 2754-81.

[11] Choi J. and M. Lee (2012). Bounding Endogenous Regressor Coefficients Using Moment Inequalities and Generalized Instruments, *Statistica Neerlandica*, 66, 161-182.

[12] Dufour J. (1997). Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models, *Econometrica*, 65, 1365-1388.

[13] Evans, W.N. and R.M. Schwab (1995). Finishing High School and Starting College: Do Catholic Schools Make a Difference?, *Quarterly Journal of Economics*, 110, 941-974.

[14] Flores, C.A. and A. Flores-Lagunes (2010). Partial Identification of Local Average Treatment Effects with an Invalid Instrument, unpublished manuscript.

[15] Hansen, L.P. (1982). Large Sample Properties of Generalised Method of Moments Estimators, *Econometrica*, 50, 1029-1054.

[16] Huber, M. and Mellace, G. (2011). Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints, unpublished manuscript.

[17] Imbens, G.W. and J.D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62, 467-475.

[18] Goldin C. and L.F. Katz (2008). The Race Between Education and Technology, Harvard University Press.

[19] Keane, M.P. and K.I. Wolpin (1997). The Career Decisions of Young Men, *Journal of Political Economy*, 105, 473-522.

[20] Kitagawa, T. (2008). A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model, unpublished manuscript.

[21] Kolesar M., Chetty R., Friedman J.N., Glaeser E.L., Imbens G.W. (2011). Identification and Inference with Many Invalid Instruments, unpublished manuscript.

[22] Madestam A., Shoag D., Veuger S., D. Yanagizawa-Drott (2013). Do Political Protests Matter? Evidence from the Tea Party Movement, unpublished manuscript.

[23] Nagar, A. (1959) The Bias and Moment Matrix of the General k-class estimators of the Parameters in Simultaneous Equations, *Econometrica*, 27, 575-595.

[24] Nelson, C.R., and R. Startz (1990). Some further results on the small sample properties of the instrumental variable estimator, *Econometrica*, 58, 967976.

[25] Reinhold S. and T. Woutersen (2011). Endogeneity and Imperfect Instruments: Estimating Bounds for the Effect of Early Childbearing on High School Completion, unpublished manuscript.

[26] Sargan, J. D. (1958). The Estimation of Economic Relationships Using Instrumental Variables, *Econometrica*, 26, 393-415.

[27] Stock, J.H., Wright, J.H., and M. Yogo (2002). A Survey of Weak Instruments and Weak Identification in GMM, *Journal of Business & Economic Statistics*, 20, 518-529.

[28] Vytlacil, E.J. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result, *Econometrica*, 70, 331-341.

Appendix

A Proof: Z uncorrelated with error term

Claim: Z and the error term are uncorrelated once we have adjusted for $\bar{h}$.

The covariance of Z and $\left[h(X) - \bar{h}\right] Z + \epsilon$ is equal to

$$E\left[Z\left(\left[h(X) - \bar{h}\right] Z + \epsilon\right)\right] - E(Z)E\left(\left[h(X) - \bar{h}\right] Z + \epsilon\right)$$

Factoring, this gives

$$E\left(Z^2\left[h(X) - \bar{h}\right]\right) + E(Z\epsilon) - E(Z)E\left(Z\left[h(X) - \bar{h}\right]\right) - E(Z)E(\epsilon)$$

Because ϵ is defined as an error term with mean zero and orthogonal to Z, the second and fourth terms are equal to zero. Applying the law of iterated expectations to the two remaining terms gives

$$E\left[E\left(Z^2\left[h(X)-\bar{h}\right] \mid Z\right)\right] - E(Z)E\left[E\left(Z\left[h(X)-\bar{h}\right] \mid Z\right)\right]$$

Z takes only two values, 1 and 0, and each of the terms which is now conditioned on Z takes the value 0 when $Z=0$. Then we can write this expression as

$$E(Z)E\left[h(X)-\bar{h} \mid Z=1\right] - E(Z)^2E\left[h(X)-\bar{h} \mid Z=1\right]$$

This equals zero when $E\left[h(X)-\bar{h} \mid Z=1\right]=0$. But $\bar{h}$ is a constant, so $E\left[h(X)-\bar{h} \mid Z=1\right] = E\left[h(X) \mid Z=1\right]-\bar{h}=\bar{h}-\bar{h}=0$.

Therefore, the covariance of Z and the error term is zero, so they are uncorrelated.