Estimation of Average Treatment Effects With Misclassification

Arthur Lewbel*
Boston College

October, 2003

Abstract

This paper provides conditions for identification and estimation of the conditional or unconditional average effect of a binary treatment or policy on a scalar outcome in models where treatment may be misclassified. Misclassification probabilities and the true probability of treatment are also identified.

Misclassification occurs when treatment is measured with error, that is, some units are reported to have received treatment when they actually have not, and vice versa. Conditional outcomes, treatment probabilities, and misclassification probabilities are nonparametric. The identifying assumption is the existence of a variable that affects the decision to treat and satisfies some conditional independence assumptions. This variable could be an instrument or a second mismeasure of treatment.

Estimation takes the form of either ordinary GMM or a local GMM that is proposed, which can be used generally to nonparametrically estimate functions from conditional moment restrictions.

JEL Codes: C14, C13. Keywords: Program Evaluation, Treatment Effects, Misclassification, Contamination Bias, Measurement error, Binary Choice, Binomial Response.

This research was supported in part by the National Science Foundation through grant SES-9905010. The author wishes to thank Alberto Abadie, Francesca Molinari, and James Heckman for many helpful comments. Any errors are my own.

*Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA. Tel: (617)–552-3678. email: lewbel@bc.edu url: http://www2.bc.edu/~lewbel

1 Introduction

This paper provides conditions for nonparametric identification, and associated estimators, of the average effect (conditioned on covariates) of a binary treatment, program, or policy on a scalar outcome in models where treatment may be misclassified. Misclassification occurs when treatment is measured with error, that is, some units are reported to have received treatment when they actually have not, and vice versa. The assumptions provided also identify misclassification probabilities and the true probability of treatment (conditional on covariates), in addition to identifying the conditional average treatment effect. The proposed estimators take the form of ordinary Generalized Method of Moments (GMM) estimation when covariates are discrete, or when the dependence on covariates can be partly parameterized. To handle continuous covariates nonparametrically, a new local GMM estimator is proposed. This local GMM can be used generally to nonparametrically estimate functions that are identified from conditional moment restrictions.

The effects of misclassification are also known as contamination bias. In assessing the impact of a training program on wages, contamination or misclassification bias arises when some workers who did not enroll in the program have outside training unknown to the researcher, are thereby erroneously classified as untreated.

In a medical context, misclassification may arise when some patients fail to follow a therapy regiment that is assigned to them.

If treatment is schooling, some respondents may either lie or not know if the particular training or schooling they've had counts as higher education. Kane, Rouse, and Staiger (1999) document many different sources of misreported educational attainment. Based on comparisons between different measures of years of schooling, such as self reporting versus transcripts, they document considerable measurement errors for those completing less than twelve years of schooling, and provide evidence that both transcripts and self reports contain significant measurement errors.

In studies of returns to unionization, where treatment is union status and the outcome is wages, misreporting of union status is widely recognized. In a current survey of this literature, Hirsch (2003) shows that the impact of union status misclassification is likely to be large and increasing over time.

In a study of pension plans, Gustman and Steinmeier (2001, table 6c) and Molinari (2002, table 5) found that 15% of respondents that actually had a defined benefit plan claimed to have a defined contribution plan, and 26% that actually had a defined contribution plan claimed to have a defined benefit plan. In this example, an analysis of treatment (plan type) on outcome (e.g., retirement income) would suffer from substantial misclassification bias if respondent's data on plan type were used.

Treatment is defined to be exogeneous or unconfounded if the decision to treat or to enroll in a program is independent of potential outcomes conditional on covariates. Assuming unconfoundedness, when treatment is observed without error the average treatment effect can be estimated by matching, differencing within subpopulation averages of treated and untreated units, or by propensity score methods. Relevant models and estimators include Heckman (1974, 1976), Rubin (1974), Heckman and Robb (1985), Rosenbaum and Rubin (1985), Manski (1990, 1997), Robins, Mark, and Newey (1992), Angrist, Imbens, and Rubin (1996), Heckman, Ichimura, and Todd (1998), Hahn (1998), Abadie and Imbens (2002), and Hirano, Imbens, and Ridder (2002). It will be assumed here that the unobserved true treatment satisfies a mean unconfoundedness condition.

Many structural treatment models, in particular some parametric or semiparametric latent variable selection models, violate unconfoundedness or selection on observables assumptions. See, e.g., Lewbel (2002), and more generally Vytlacil (2002) and Heckman and Vytlacil (2001) for relationships between latent variable selection models and treatment effect estimators. Apparent violations of selection on observables or of other conditions used to identify and estimate treatment effects could be due to misclassification. For example, when the true assignment of individuals to either treatment or no treatment is completely determined by observables, misclassification randomness would produce the mistaken impression of randomness in the assignments.

In linear outcome models, if observed treatment satisfied classical measurement error, then ordinary two stage least squares methods could be used. However, binary regressors cannot satisfy classical measurement error assumptions (see Aigner1973). Alternatives to two stage least squares for linear outcome models with mismeasured binary regressors include Klepper (1998), Card (1996), Bollinger (1996), and Kane, Rouse, and Staiger (1999).

Choice or assignment of treatment is an example of a binary choice or binomial response model. Papers that consider estimation of binomial response models in the presence of misclassification include McFadden (1984), Chua and Fuller (1987), Brown and Light (1992), Poterba and Summers (1995), Abrevaya and Hausman (1999), Hausman, Abrevaya, and Scott-Morton (1998), and Lewbel (2000). In binomial response models a distinction is made between misclassification that can happen to any unit with some probability, versus the case where some respondents (which are unknown to the researcher) are always correctly measured while others provide ”natural responses,” e.g., always claiming to be treated or untreated regardless of the truth. See, e.g., Finney (1964). For the purposes of the present paper the distinction between these two types of misclassification is irrelevant; they will be observationally equivalent.

For treatment effects, misclassification is closely related to the problem of identification in the presence of imperfect compliance in an otherwise randomized experiment. See, e.g., Angrist, Imbens and Rubin (1996) and Balke and Pearl (1997). Also related are cases where an instrument used for identification is imperfect, as in Hotz, Mullin, and Sanders (1997), or when either covariates, treatment, or outcomes are not observed for some subjects, as in Robins (1997), Horowitz and Manski (2000), and Molinari (2001).

This paper provides sufficient conditions for nonparametric identification of probability of treatment, misclassification probabilities, and of average treatment effects, conditioned on covariates, in the presence of misclassification errors. Estimators that employ these identification conditions are provided.

The main identifying condition is the existence of an instrument, i.e., a scalar or vector of variables that is correlated with the decision to treat, but does not affect conditional misclassification probabilities or the conditional average treatment effect. An example of such an instument could be the distance to school and related measures employed by Card (1995) and others in the returns to education literature. A second mismeasure or proxy for treatment may also serve as an instrument, e.g., self reported educational attainment could be used as an instrument to deal with errors in transcript reported education level. This required identification condition is an example of an exclusion restriction, that is, a variable that affects some relevant functions and not others. Exclusion restrictions are a common method of obtaining semi and nonparametric identification in econometric models. See, e.g., Powell's (1994, section 2.5) survey.

Interestingly, even though conditional outcomes and unknown probabilities can be continuously distributed and potentially have high dimension, nonparametric identification is obtained here even with a discrete instrument. If the instrument is conditionally independent of outcomes, then the instrument need only take on two different values. Under the weaker assumption that the instrument is conditionally independent of the conditional average treament effect (but not necessarily of the outcomes themselves), then an instrument that can take on as few as three different values suffices. Instruments that can take on more than two or three values provide overidentify restrictions that can be used to test the validity of the instrument. A related result is Abadie (2003), who considers use of a binary instrument in a correctly classified treatment effect model.

Identification is obtained by deriving conditional moment restrictions of the form

$$E[g(q(Z), W) \mid Z = z] = 0, \tag{1}$$

where g is a known vector valued function, W is a vector of observed functions of outcomes and (misclassified) treatments, Z is a vector of covariates, and $q(Z)$ is a vector of unknown functions including

the conditional average treatment effect and correctly classified conditional treatment probabilities. For discretely distributed Z, or when the functions $q(z)$ can be finitely parameterized, Hansen's (1982) Generalized Method of Moments (GMM) estimator can be used. To nonparametrically estimate $q(z)$ with continuously distributed Z, a local GMM estimator is proposed.

This local GMM is a generic estimator of models in the form of equation (1), and so may have more general applications. Other, nontreatment examples where local GMM could be used are provided, such as semiparametric binary choice. Estimators related to local GMM include Gozalo and Linton (2000) and Newey and Powell (2003).

2 Identification

Sufficient conditions for identification are provided here under three different scenarios. First is the simple case of identification when misclassification probabilities are known from some outside source, such as aggregate data or a validation sample. Second is the case where identification is obtained by an instrument that is conditionally independent of the treatment effect and can take on at least three values. Third is identification using an instrument that is conditionally independent of outcomes but need only take on two values. In this third case the instrument could be a second mismeasure of treatment.

2.1 Identification by Known Misclassification Probabilities

Let Y be an observed outcome, T^* index the actual, unobserved treatment, and T index the reported treatment. Let $t = 1$ denote receiving treatment or enrolling in a program, and $t = 0$ denote no treatment. Let $Y(t)$ denote the outcome from treatment $T^* = t$, and X be a vector of observable covariates. The goal is estimation of the conditional average treatment effect $E[Y(1) - Y(0) \mid X = x]$. Define

$$\tau^*(x) = E(Y \mid X = x, T^* = 1) - E(Y \mid X = x, T^* = 0) \tag{2}$$

ASSUMPTION A1: $E[Y(t) \mid T^*, X] = E[Y(t) \mid X]$

Assumption A1 is a weak version of the standard unconfoundedness assumption, which is with respect to the true treatment T^*. Heckman, Ichimura, and Todd (1998) show that this version of unconfoundedness

implies that the conditional average treatment effect satisfies

$$E[Y(1) - Y(0) \mid X = x] = \tau^*(x)$$

If T^* were observed without error, then equation (2) would provide an estimator for $\tau^*(x)$, by replacing expectations with nonparametric regressions. Other estimators, e.g. those based on matching or propensity score methods could be used instead given unconfoundedness.

ASSUMPTION A2: $E(Y \mid X, T^*, T) = E(Y \mid X, T^*)$.

Assumption A2 says that, conditional on X and on the actual treatment T^*, the measurement of treatment does not affect the expected outcome. This is analogous to the classical measurement error assumption that actual outcomes be independent of measurement errors made by the researcher. This could be a substantive assumption if the misclassification is due to misperception or deceit on the part of the subject, for example, if T indicates the treatment that the subject thinks he or she had, then Assumption A2 would rule out placebo effects. This assumption could also be violated if an individual's propensity to lie about treatment is related to outcomes, e.g., individuals who erroneously claim to have a college degree might also be more aggressive job or wage seekers in general.

Make the following definitions.

$$r^*(x) = E(T^* \mid X = x)$$

$$b_t(x) = E[I(T = 1 - t) \mid X = x, T^* = t] = \Pr(T = 1 - t \mid X = x, T^* = t)$$

Conditioning on $X = x$, the function $r^*(x)$ is the probability of receiving treatment, while $b_1(x)$ is the probability of misclassifying the treated and $b_0(x)$ is the probability of misclassifying the untreated.

ASSUMPTION A3: $b_0(x) + b_1(x) < 1$, $E(T^* \mid X = x, T = 1] \neq E(T^* \mid X = x, T = 0]$, and $0 < r^*(x) < 1$ for all $x \in supp(X)$.

Assumption A3 says first that the sum of misclassification probabilities is less than one, meaning that, on average, observations of T are more accurate than pure guesses. In a binomial response model with misclassification, this assumption is what Hausman, Abrevaya, and Scott-Morton (1998) call the monotonicity condition. Given failure to observe T^*, without an assumption like this, by symmetry one could never tell if the roles of $t = 0$ and $t = 1$ were reversed, and so for example one could not distinguish whether any estimate of $\tau^*(x)$ corresponded to the treatment effect or the negative of the treatment effect. This assumption

can be relaxed to $b_0(x) + b_1(x) \neq 1$ if we only wish to identify the magnitude but not the sign of treatment effects, which may be useful in applications where the sign of treatment effects are not in doubt and large misclassification probabilities cannot be ruled out.

The second condition of Assumption A3 says that T provides some information beyond what x contains regarding the probability of treatment. Assumption A3 also requires that for any x we may condition on, there is a nonzero probability of treatment and a nonzero probability of nontreatment, which is needed because a conditional treatment effect cannot be identified if everyone is treated or if no one is treated.

Define the following functions.

$$r(x) = E(T \mid X = x)$$

$$\tau(x) = E(Y \mid X = x, T = 1) - E(Y \mid X = x, T = 0)$$

ASSUMPTION A4: Assume $r(x)$ and $\tau(x)$ are identified.

The functions $r(x)$ and $\tau(x)$ are conditional expectations of observable data, so Assumption A4 will hold given any data set that permits consistent estimation of these conditional expectations. If X is discretely distributed, then only consistency of sample averages is required.

Note that $r(x)$ and $\tau(x)$ are the same as $r^*(x)$ and $\tau^*(x)$, except defined in terms of the observed treatment T instead of the true treatment T^*, so if treatment were observed without error, then $r(x)$ would be the conditional probability of treatment and, by Assumption A1, $\tau(x)$ would equal the conditional average treatment effect.

Define the function m by

$$m(b_0, b_1, r) = \left(\frac{1}{1 - b_1 - b_0}\right)\left(1 - \frac{(1 - b_1)b_0}{r} - \frac{(1 - b_0)b_1}{1 - r}\right). \tag{3}$$

THEOREM 1: Let Assumptions A1, A2, A3, and A4 hold. Then

$$r^*(x) = \frac{r(x) - b_0(x)}{1 - b_0(x) - b_1(x)}, \tag{4}$$

$$\tau^*(x) = \tau(x)/m[b_0(x), b_1(x), r(x)] \tag{5}$$

and if $b_0(x)$ and $b_1(x)$ are identified, the probability of treatment $r^*(x)$ and conditional average treatment effect $E[Y(1) - Y(0) \mid X = x]$ are also identified.

Theorem 1 shows that if the misclassification probabilities $b_t(x)$ are known to the researcher or can be identified (for example from a validation sample or from known aggregate population proportions), then the true conditional average treatment effect $\tau^*(x)$ and the true probability of treatment $r^*(x)$ are directly identified using equations (4) and (5). Results similar to those of Theorem 1 have been used to construct bounds on treatment effects. See, e.g., Hotz, Mullin, and Sanders (1997).

One immediate implication of Theorem 1 is that $\tau^*(x) = 0$ if and only if $\tau(x) = 0$. Therefore, if we wish to test whether true treatment effects are nonzero, it suffices to test if the mismeasured treatment effects $\tau(x)$ are nonzero. Given Assumptions A1 to A4, if we only want to test whether treatment effects are nonzero, the presence of misclassification can be ignored.

If $b_0(x)+b_1(x)$ equals one, then identification breaks down because equation (4) then reduces to $r(x) = b_0(x)$, so in that case $r(x)$ provides no information regarding the true selection probability $r^*(x)$.

2.2 Identification by a Three Valued Instrument

Now consider identification without external knowledge of the misclassification probabilities $b_t(x)$. Partition X into two subvectors V and Z, so $X = (V, Z)$.

ASSUMPTION A5: For some set $\Omega \subset supp(V)$, for all $v \in \Omega$, $v_0 \in \Omega$, and $z \in supp(Z)$, we have $b_t(v, z) = b_t(v_0, z)$, $\tau^*(v, z) = \tau^*(v_0, z)$, and $r^*(v, z) \neq r^*(v_0, z)$.

In a small abuse of notation, let $b_t(z)$ and $\tau^*(z)$ denote $b_t(v, z)$ and $\tau^*(v, z)$, respectively, for $v \in \Omega$. The distribution of V can be discrete, e.g., V could be a scalar that only takes on a few different values. Assumption A5 says that there exists a variable V that affects r^*, and hence the true treatment probabilities, but after conditioning on other covariates Z does not affect either the measurement errors b_t or the conditional average treatment effect τ^* (at least for some values that V might take on).

Having a V that doesn't affect misclassification probabilities is sometimes used for identification in binomial response models with misclassification. See, e.g., Hausman, Abrevaya, and Scott-Morton (1998), Abrevaya and Hausman (1999), and Lewbel (2000). A typical assumption in misclassified binomial response is that b_0 and b_1 are constants, which would imply that any elements of X could serve as V for that part of Assumption A5.

Having V affect r^* but not τ^* is a weaker version of the type of exclusion of assumption that is commonly used in the identification of selection models. See e.g., Heckman (1990) for a discussion. Variants of this assumption are used by Manski (1990) to sharpen bounds on unidentified treatment effects, and by Imbens and Angrist (1994) to identify local average treatment effects. For a job training program, an example of V might be nonwage related income or benefits, or more generally any variable that, after conditioning on other covariates, does not affect the average effectiveness of the program but is correlated with eligibility or selection, such as distance to schools as employed by Card (1995) and others.

ASSUMPTION A6: There exists three elements $v_k \in \Omega$, $k = 0, 1, 2$, such that

$$\left(\frac{\tau(v_0,z)}{r(v_1,z)} - \frac{\tau(v_1,z)}{r(v_0,z)}\right)\left(\frac{\tau(v_0,z)}{1-r(v_2,z)} - \frac{\tau(v_2,z)}{1-r(v_0,z)}\right) \neq \left(\frac{\tau(v_0,z)}{r(v_2,z)} - \frac{\tau(v_2,z)}{r(v_0,z)}\right)\left(\frac{\tau(v_0,z)}{1-r(v_1,z)} - \frac{\tau(v_1,z)}{1-r(v_0,z)}\right)$$

The main content of Assumption A6 is that V can take on at least three values. Given Assumption A5, the required inequality in Assumption A6 will only fail to hold if $\tau(v_0, z) = 0$ or if a complicated equality relationship holds amongst the three conditional outcomes and conditional treatment probabilities, which would require a perfect coincidence between probabilities and outcomes. Assumption A6 can be empirically tested, because these $\tau(v_k, z)$ and $r(v_k, z)$ functions are conditional expectations of observable data, and so can be directly estimated (they are identified by Assumption A4). Finally, note that if V can take on more than three values, then Assumption A6 will hold as long as there exists any one triplet of V values that satisfies the necessary inequality.

THEOREM 2: Let Assumptions A1, A2, A3, A4, A5, and A6 hold. Then the conditional misclassification probabilities $b_0(x)$ and $b_1(x)$, the conditional probability of treatment $r^*(x)$, and the conditional average treatment effect $E[Y(1) - Y(0) \mid X = x]$ are all identified. Also, if the condition in Assumption A3 that $b_0(x) + b_1(x) < 1$ is replaced by $b_0(x) + b_1(x) \neq 1$, then the conditional average treatment effect is identified up to sign.

A key component of Theorem 2 is that data on outcomes helps to identify misclassification probabilities. In particular, it follows from Theorem 1 that $\tau(v_k, z)m[b_0(z), b_1(z), r(v_0, z)] = \tau(v_0, z)m[b_0(z), b_1(z), r(v_k, z)]$. Substituting equation (3) into this expression yields an equation that, for each value of z, depends only on the identified functions τ and r and on the two unknowns b_0 and b_1. Evaluating this expression for $k = 1$

and $k = 2$ gives two equations in the two unknowns. These equations are nonlinear, but the proof of Theorem 2 shows that these equations still uniquely define and thereby identify b_0 and b_1. Identification of true treatment effects and probabilities then follows from Theorem 1.

Each value that V can take on provides another equation that b_0 and b_1 must satisfy, so in general the larger is the set Ω of values that V can take on (which satisfy Assumption A5), the greater will be the number of overidentifying restrictions determining $b_0(z)$ and $b_1(z)$.

2.3 Identification by a Two Valued Instrument or a Second Treatment Measure

Identification based on Theorem 2 requires V to take on at least three different values. It is shown below that only a binary V is needed when some assumptions are strengthened. An example of a binary V is a second mismeasure of T^*. Kane, Rouse, and Staiger (1999) show that, in a linear model, treatment effects and treatment probabilities can be identified given two different mismeasures of T^*, and provide a number of returns to schooling examples where two such measures are available. Theorem 3 below generalizes their result by showing nonparametric identification given any binary instrument.

Define the functions

$$h_t^*(x) = E(Y \mid X = x, T^* = t)$$

$$h_t(x) = E(Y \mid X = x, T = t)$$

for $t = 0, 1$. Note that $\tau^*(x) = h_1^*(x) - h_0^*(x)$ and $\tau(x) = h_1(x) - h_0(x)$.

ASSUMPTION B4: Assume $r(x)$, $h_0(x)$ and $h_1(x)$ are identified.

Assumption B4 is a slight strengthening of Assumption A4. The functions $r(x)$, $h_0(x)$ and $h_1(x)$ are conditional expectations of observable data, so Assumption A4 will hold given any data set that permits consistent estimation of these conditional expectations.

ASSUMPTION B5: For some set $\Omega \subset supp(V)$, for all $v \in \Omega$, $v_0 \in \Omega$, and $z \in supp(Z)$, we have $b_t(v, z) = b_t(v_0, z)$, $h_t^*(v, z) = h_t^*(v_0, z)$, and $r^*(v, z) \neq r^*(v_0, z)$.

Once again let $b_t(z)$ and $\tau^*(z)$ denote $b_t(v, z)$ and $\tau^*(v, z)$, respectively, for $v \in \Omega$, and now also let $h_t^*(z)$ denote $h_t^*(v_0, z)$ for $t = 0, 1$. Assumption B5 differs from Assumption A5 only in that it requires $h_0^*(v, z)$ and $h_1^*(v, z)$ to not depend on v for all $v \in \Omega$, instead of only requiring that the difference

$\tau^*(v,z) = h_1^*(v,z) - h_0^*(v,z)$ not depend on v. Assumption B5 requires that the outcome Y itself be conditionally independent of V, while Assumption A5 only required that the conditional average treatment effect be independent of V.

Assumption B5 is particularly plausible when V is a second mismeasure of T^*. In that case a sensible extension of Assumption A2 that treats T and V equivalently is $E(Y \mid Z, T^*, T, V) = E(Y \mid Z, T^*)$, which if true would suffice to make Assumption A2 and $h_t^*(v,z) = h_t^*(v_0,z)$ hold. Also, when V is a second mismeasure then the assumption that $b_t(v,z) = b_t(v_0,z)$ is comparable to the Kane, Staiger and Rouse (1999) assumption that the two mismeasures of T^* be conditionally independent of each other, conditioning on T^* and Z.

ASSUMPTION B6: There exists two elements $v_0 \in \Omega$ and $v_1 \in \Omega$ such that

$$\left(\frac{h_0(v_1,z) - h_0(v_0,z)}{[1-r(v_1,z)]^{-1} - [1-r(v_0,z)]^{-1}}\right)\left(\frac{\tau(v_0,z)}{1-r(v_1,z)} - \frac{\tau(v_1,z)}{1-r(v_0,z)}\right) \neq$$
$$\left(\frac{h_1(v_1,z) - h_1(v_0,z)}{[r(v_1,z)]^{-1} - [r(v_0,z)]^{-1}}\right)\left(\frac{\tau(v_k,z)}{r(v_0,z)} - \frac{\tau(v_0,z)}{r(v_k,z)}\right)$$

The main content of Assumption B6 is that V can take on two different values. Analogous to Assumption A6, the required inequality will only fail to hold given a perfect coincidence between magnitudes of probabilities and of outcomes. Assumption B6 is testable because its component functions are identified by Assumption B4. If V can take on more than two values, then Assumption B6 will hold as long as there exists any pair of V values that satisfy the inequality.

THEOREM 3: Let Assumptions A1, A2, A3, B4, B5, and B6 hold. Then the conditional misclassification probabilities $b_0(x)$ and $b_1(x)$, the conditional probability of treatment $r^*(x)$, and the conditional average treatment effect $E[Y(1) - Y(0) \mid X = x]$ are all identified. Also, if the condition in Assumption A3 that $b_0(x) + b_1(x) < 1$ is replaced by $b_0(x) + b_1(x) \neq 1$, then the conditional average treatment effect is identified up to sign.

The identification in Theorem 3 comes from expressing the observable $h_t(v,z)$ as a function of the observable $r(v,z)$ and the four unknown functions $h_0^*(z)$, $h_1^*(z)$, $b_0(z)$, $b_1(z)$. For each value of z, observing this relationship for $t = 0, 1$ and for $v = v_0, v_1$ gives four equations in these four unknowns. These

equations are nonlinear, but the proof of Theorem 3 shows that they uniquely define and thereby identify $b_0(z)$ and $b_1(z)$, and so then by Theorem 1 the true selection probabilities and treatment effects are identified. Each value that V can take on provides more equations that the unknown functions must satisfy, so in general the larger is the set Ω, the greater will be the number of overidentifying restrictions.

3 Conditional Moments

To construct estimators, the previous identification conditions will now be expressed in the form of conditional moments. Assume the distribution of V is discrete, define $\Omega = supp(V) = \{v_0, ..., v_K\}$ and let $r_k^*(z) = r^*(v_k, z)$. Let $W = (Y, T, V)$.

Define the vector valued function $q_0(z)$ as the vector of $K + 4$ elements

$$q_0(z) = \big(b_0(z), b_1(z), r_0^*(z), ...r_K^*(z), \tau^*(z)\big) \tag{6}$$

and define g as the vector valued function $g[q_0(z), w]$ consisting of the following $2K + 2$ elements

$$[b_0(z) + (1 - b_0(z) - b_1(z))r_k^*(z) - T]I(V = v_k), \qquad k = 0, ..., K \tag{7}$$

$$\tau^*(z)I(V = v_k) + \frac{YT - (1 - b_1(z))r_k^*(z)\tau^*(z)I(V = v_k)}{b_0(z) + (1 - b_0(z) - b_1(z))r_k^*(z)}$$

$$- \frac{Y(1 - T) + (1 - b_0(z))(1 - r_k^*(z))\tau^*(z)I(V = v_k)}{1 - [b_0(z) + (1 - b_0(z) - b_1(z))r_k^*(z)]}, \qquad k = 0, ..., K \tag{8}$$

COROLLARY 1: Define the function q_0 by equation (6) and the function g as the vector of functions (7) and (8). For any value of z in its support, if Assumptions A1, A2, A3, A4, A5, and A6 hold then the only function $q(z)$ that satisfies $E[g(q(Z), W) \mid Z = z)] = 0$ and has first two elements that are nonnegative and sum to less than one, is $q(z) = q_0(z)$.

Note in Corollary 1 that the first two elements of $q_0(z)$ are $b_0(z)$ and $b_1(z)$, and so are restricting these functions to be positive and sum to less than one.

The objects we wish to estimate are elements of $q_0(z)$. Corollary 1 shows that the identification based on Theorem 2 can be espressed as the statement that the unknown functions $q_0(z)$ are the solutions to the

vector of conditional moments $E[g(q(Z), W) \mid Z = z)] = 0$ (with the added inequality constraints on b_0 and b_1) Note that this requires V to take on at least three different values, so $K \geq 2$.

The identification based on Theorem 3 can also be expressed as conditional moment restrictions, as follows. Redefine $q_0(z)$ to be the vector valued function consisting of the $K + 5$ elements

$$q_0(z) = \left(b_0(z), b_1(z), r_0^*(z), ...r_K^*(z), \tau^*(z), h_0^*(z)\right)' \tag{9}$$

where $h_0^*(z) = h^*(\upsilon_0, z)$. Redefine $g[q_0(z), W]$ to be the vector of $3K + 3$ functions consisting of the same $2K + 2$ functions (7) and (8) as before, but also including the $K + 1$ additional functions

$$\left[YT - (1 - b_1(z))r_k^*(z)\tau^*(z) - (b_0(z) + (1 - b_0(z) - b_1(z))r_k^*(z))h_0^*(z)\right] I(V = \upsilon_k), \qquad k = 0, ..., K \tag{10}$$

COROLLARY 2: Define the function q_0 by equation (9) and the function g as the vector of functions (7), (8), and (10). For any value of z in its support, if Assumptions A1, A2, A3, B4, B5, and B6 hold then the only function $q(z)$ that satisfies the equation $E[g(q(Z), W) \mid Z = z)] = 0$ and has first two elements that are nonnegative and sum to less than one, is $q(z) = q_0(z)$.

Corollary 2 is based on Theorem 3. This was the case where the instrument V is assumed to be conditionally independent of the outcome Y. This only required V to take on as few as two different values, Corollary 2 only requires $K \geq 1$ and V could be, like T, another proxy for T^*.

For testing purposes, it may be of interest to estimate the misclassified treatment effect $\tau_k(z) = \tau(\upsilon_k, z)$ and misclassified treatment parameters $r_k(z) = r(\upsilon_k, z)$. To do so, redefine $q_0(z)$ as the $K + 1$ vector

$$q_0(z) = (r_0(z), ...r_K(z), \tau(z)) \tag{11}$$

and define $g[q_0(z), \omega]$ as the vector valued function consisting of the $2K + 2$ elements

$$[r_k(z) - T]I(V = \upsilon_k), \qquad k = 0, ..., K \tag{12}$$

$$\left(\frac{YT}{r_k(z)} - \frac{Y(1 - T)}{1 - r_k(z)} - \tau(z)\right) I(V = \upsilon_k), \qquad k = 0, ..., K \tag{13}$$

COROLLARY 3: Define the function q_0 by equation (11) and the function g as the vector of functions (12) and (13). For any value of z in its support, if Assumptions A1 and A4 hold then the only function $q(z)$ that satisfies the equations $E[g(q(Z), W) \mid Z = z)] = 0$ is $q(z) = q_0(z)$.

As defined for Corollaries 1 and 2, the vector $q_0(z)$ contains the treatment effects, true treatment probabilities, and misclassification probabilities, while in Corollary 3 the vector $q_0(z)$ contains the misclassified treatment effects and probabilities. In each case a corresponding function g is provided such that $q_0(z)$ uniquely satisfies the conditional moment restriction $E[g(q_0(Z), W) \mid Z = z)] = 0$.

4 Estimation

This section describes estimators for $q_0(z)$ given the conditional moment restriction $E[g(q_0(Z), W) \mid Z = z)] = 0$. While the application here will be to treatment effects as described in the previous section, other econometric models can also be cast in this conditional moment restriction form, and so could be estimated using the estimators described here. For example, consider the nonparametric probit: model $W = I[q_0(z) + e \geq 0]$, where $q_0(z)$ is an unknown function and e is a standard normal independent of Z, or has some other known distribution. Then $g(q_0(Z), W) = W - F_e[q_0(z)]$ where F_e is the CDF of $-e$. Nonparametric censored or truncated regression would have a similar form. Another class of examples could be Euler equations, which are mean zero conditional on information in a given time period, and could have parameters $q_0(z)$ (such as preference parameters) that are unknown functions of observables.

Three estimators are provided. The first estimator and is for use when $q_0(z)$ can be finitely parameterized. The second and third are, respectively, for nonparametric estimation of $q_0(z)$ when Z is discretely or continuously distributed. For these estimators it is assumed that we have data consisting of Z_i, W_i for $i = 1, ..., n$. Limiting distributions are provided assuming these observations are independent and identically distributed.

4.1 Parameterized Estimation

Suppose we can write $q_0(z) = s(z, \beta_0)$ where s is a known function and β_0 is finite vector of unknown parameters. Only the dependence of probabilities and of treatment effects on z is parameterized here;.the dependence of probabilities and treatment effects on v and on unobservables is still left unspecified. In this

case the conditional moments $E[g(q(Z), W) \mid Z = z)] = 0$ imply unconditional moments

$$E[\eta_j(Z)g(s(Z, \beta_0), W)] = 0, \qquad j = 1, ..., J \tag{14}$$

for any J bounded functions $\eta_j(Z)$ chosen by the econometrician. Given the unconditional moments of equation (14) for $j = 1, ..., J$, we may apply Hansen's (1982) Generalized Method of Moments (GMM) to obtain a consistent asymptotically normal estimate of β_0. Identification of β_0 will depend on the specification of the function s and $\eta_j(Z)$, but Corollaries 1 and 2 imply that as long as β_0 is identified from $q_0(z) = s(z, \beta_0)$, it should be possible to choose $\eta_j(Z)$ functions to identify β_0. Identification requires that the dimension of s (either $2K + 2$ using Corollary 1 or $3K + 3$ using Corollary 2) times J be greater or equal to the dimension of β_0.

Let $G(\beta, W, Z)$ be the vector consisting of all the elements of $\eta_j(Z)g(s(Z, \beta), W)$ for $j = 1, ..., J$. Given n independently, identically distributed draws W_i, Z_i, the standard GMM estimator is

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} G(\beta, W_i, Z_i)' \Omega_n \sum_{i=1}^{n} G(\beta, W_i, Z_i)$$

for some sequence of positive definite Ω_n. If Ω_n is a consistent estimator of $E\left[G(\beta_0, W, Z)G(\beta_0, W, Z)'\right]$, then efficient GMM has

$$\sqrt{n}(\widehat{\beta} - \beta_0) \to^d N\left(0, E\left(\frac{\partial G(\beta_0, W, Z)}{\partial \beta'}\right) E\left[G(\beta_0, W, Z)G(\beta_0, W, Z)'\right]^{-1} E\left(\frac{\partial G(\beta_0, W, Z)}{\partial \beta'}\right)'\right)$$

4.2 Estimation With Discrete Covariates

Now assume we do not have a parameterization for $q_0(z)$, but Z is discretely distributed, or more specifically, that Z has one or more mass points and we only wish to estimate $q_0(z)$ at those points.

Let $\theta_{z0} = q_0(z)$. If the distribution of Z has a mass point with positive probability at z, then

$$E[g(\theta_z, W) \mid Z = z] = \frac{E[g(\theta_z, W)I(Z = z)]}{E[I(Z = z)]}$$

so $E[g(q_0(z), W) \mid Z = z] = 0$ if and only if $E[g(\theta_{z0}, W)I(Z = z)] = 0$. It therefore follows from Corollaries 1 and 2 that θ_{z0} is identified from these moment conditions, and we may estimate parameters θ_{z0} by the ordinary GMM estimator

$$\widehat{\theta}_z = \arg\min_{\theta_z} \sum_{i=1}^{n} g(\theta_z, W_i)' I(Z_i = z) \Omega_n \sum_{i=1}^{n} g(\theta_z, W_i)' I(Z_i = z) \tag{15}$$

for some sequence of positive definite Ω_n. If Ω_n is a consistent estimator of $E[g(\theta_{z0}, W)g(\theta_{z0}, W)'I(Z = z)]$, then efficient GMM gives

$$\sqrt{n}(\widehat{\theta}_z - \theta_{z0}) \to^d N\left(0, E\left(\frac{\partial g(\theta_{z0}, W)I(Z=z)}{\partial \theta_z'}\right) E\left[g(\theta_{z0}, W)g(\theta_{z0}, W)'I(Z=z)\right]^{-1} E\left(\frac{\partial g(\theta_{z0}, W)I(Z=z)}{\partial \theta_z'}\right)'\right)$$

If $K > 3$ using Corollary 2 or if $K > 2$ using Corollary 3 then then θ_{z0} is overidentified and standard tests of moment validity may be applied.

GMM assumes parameters have compact support. This could be imposed, consistent with Assumptions A3 and A6 by assuming that $\delta \leq r_k^*(z) \leq 1 - \delta$, $0 \leq b_t(z)$, $b_0(z) + b_1(z) \leq 1 - \delta$, and $\delta \leq |\tau^*(z)| \leq 1/\delta$ for some small $\delta > 0$.

Let $\widehat{\tau}_z^*$ denote the element of $\widehat{\theta}_z$ that corresponds to the conditional average treatment effect. This $\widehat{\tau}_z^*$ is a consistent estimator of the treatment effect $\tau^*(z)$ provided that this effect is nonzero. However, $\tau^*(z) = 0$ violates Assumption A6, so one cannot use an ordinary Wald t-statistic to test for a zero treatment effect (the ordinary t statistic will be valid for testing other values, such as whether $\tau^*(z)$ equals a given small, nonzero value). By Theorem 1, $\tau^*(z) = 0$ if and only if $\tau(\upsilon_k, z) = 0$, so if the existence of a nonzero treatment effect is in doubt, one may test for it by estimating the treatment effect in the usual way as if T equaled T^*. In short, mismeasurement in T can be ignored for testing $\tau^*(z) = 0$. Specifically, we may estimate $\tau(\upsilon_k, z)$ by applying the above GMM estimator with q and g defined by Corollary 3, and then apply an ordinary Wald test of the hypothesis that $\tau_k(z)$ is zero for $k = 0, ..., K$.

More generally, the moments in Corollary 3 may be estimated either separately, or (if $\tau^*(z)$ is nonzero) together with those of Corollary 1 or 2 to test differences between true and misclassified treatment probabilities or effects.

Additional moments for estimating $\widehat{\theta}_z$ may be constructed given more information about the misclassification probabilities $b_0(z)$ and $b_1(z)$. For example, in some applications it may be known that one or the other of these probabilities is zero, or that these probabilities are equal to each other. Given either of these constraints, only a binary V would be required for identification and estimation based on Theorem 2 without the added assumptions required for Theorem 3.

GMM based on Corollary 1 or 2 provides estimates conditional on a given $z \in supp(Z)$. To estimate an unconditional average treatment effect, observe that this effect is

$$E(Y \mid T^* = 1) - E(Y \mid T^* = 0) = \sum_{z \in supp(Z)} \tau_z^* E[I(Z = z)]$$

which would be estimated by the sample average $\sum_{i=1}^{n}\sum_{z\in supp(Z)}\widehat{\tau}_z^* I(Z_i = z)/n$. Note that $cov(\widehat{\tau}_z^*, \widehat{\tau}_{\widetilde{z}}^*) = 0$ for $z \neq \widetilde{z}$ because $\widehat{\tau}_z^*$ and $\widehat{\tau}_{\widetilde{z}}^*$ are estimated using different subsets of data.

If constraints are known to exist on the parameters across values of z, then the GMM estimates for each z can be stacked into one large GMM to improve efficiency. For example, if misclassification probabilities b_t are known to be constant, or more generally independent of some elements of z, then that restriction could be imposed in the collection of moments $E[g(\theta_{z0}, W)I(Z = z)] = 0$ for all $z \in supp(Z)$.

If the Corollary 2 definitions are being used and if V is, like T, a mismeasure of T^*, one could switch the roles of T and V to obtain additional moment conditions for estimating τ^*. Alternatively, it may preferable to use whichever measure is likely to be the more accurate one as T, to ensure that assumption A3 is satisfied.

4.3 Local GMM Estimation For Continuous Covariates

Continue to assume that $E[g(q_0(Z), W) \mid Z = z] = 0$, where g is known, q_0 is unknown and not parameterized, and now Z is continuously distributed. A local GMM estimator is proposed. The idea is to apply equation (15) to the case of continuous Z by replacing averaging over just observations $Z_i = z$ with local averaging over observations Z_i in the neighborhood of z.

Assumption C1. Let $Z_i, W_i, i = 1, ..., n$, be an iid random sample of observations of random vectors Z, W. The d vector Z is continuously distributed with density function $f(Z)$. For given point z in the interior of $supp(Z)$ having $f(z) > 0$, a compact set $\Theta(z)$, and a given vector valued function $g(q, w)$ where $g(q(z), w)$ is twice differentiable in the vector $q(z)$ for all $q(z) \in \Theta(z)$, there exists a unique $q_0(z) \in \Theta(z)$ such that $E[g(q_0(Z), W) \mid Z = z] = 0$. The dimension of the vector $g(q, w)$ is greater than or equal to the dimension of the vector $q(z)$. Let Ω_n be a finite positive definite matrix for all n, as is $\Omega_0 = plim_{n\to\infty}\Omega_n$.

Assumption C1 provides the required moment condition structure of the model, and Assumption C2 below provides conditions for local averaging. Define $\varepsilon[q(z), W]$ and $V[q(z)]$ by

$$\begin{aligned}\varepsilon[q(z), W] &= g(q(z), W)f(z) - E[g(q(z), W)f(z) \mid Z = z] \\ V[q(z)] &= E\left[\varepsilon(q(z), W)\varepsilon(q(z), W)^T \mid Z = z\right]\end{aligned}$$

Assumption C2. Let η be some constant greater than 2. Let K be a nonnegative symmetric kernel function satisfying $\int K(u)du = 1$ and $\int ||K(u)||^{\eta}du$ is finite. For all $q(z) \in \Theta(z)$, $E[||g(q(z), W)f(z)||^{\eta} \mid Z = z]$, $V[q(z)]$, $E[[\partial g(q(z), W)/\partial q(z)]f(z) \mid Z = z]$, and $Var[[\partial g(q(z), W)/\partial q(z)]f(z) \mid Z = z]$ are finite and continuous at z and $E[g(q(z), W)f(z) \mid Z = z]$ is finite and twice continuously differentiable in z.

Define

$$R(z) = E\left(\frac{\partial g[q(Z), W]}{\partial q(Z)^T} f(Z) \mid Z = z\right)$$

$$S_n(q(z)) = \frac{1}{nb^d}\sum_{i=1}^{n} g[q(z), W_i]K\left(\frac{z - Z_i}{b}\right)$$

where $b = b(n)$ is a bandwidth parameter. The proposed local GMM estimator is

$$\widehat{q}(z) = \arg \inf_{q(z)\in\Theta(z)} S_n(q(z))^T \Omega_n S_n(q(z))$$

THEOREM 4: Given Assumptions C1 and C2, if the bandwidth b satisfies $nb^{d+4} \to 0$ and $nb^d \to \infty$, then $\widehat{q}(z)$ is a consistent estimator of $q_0(z)$ with limiting distribution

$$(nb)^{1/2}[\widehat{q}(z) - q_0(z)] \to^d N\left[0, (R(z)^T \Omega R(z))^{-1} R(z)^T \Omega V(q_0(Z))\Omega R(z)(R(z)^T \Omega R(z))^{-1} \int K(u)^2 du\right]$$

Theorem 4 assumes a bandwidth rate that makes bias shrink faster variance, and so is not mean square optimal. One could instead choose the mean square optimal rate where nb^{d+4} goes to a constant, but the resulting bias term would have a complicated form that depends on, among other terms, the kernel regression biases in both $S_n(q_0(z))$ and $S_n'(q_0(z))$.

5 Monte Carlo

For estimation based on Corollary 1, the first simulation design has Z empty and $\Omega = \{0, 1, 2\}$. For each observation, a V is drawn with $prob(V = v_k) = 1/3$. Next a T^* is drawn from $\{0, 1\}$ with $prob(T^* =$

$1 \mid V = v_k) = r^*(v_k)$ with $r^*(0) = 3/4$, $r^*(1) = 1/2$, and $r^*(2) = 1/4$. Next an outcome Y is drawn from from a normal $N(T^*, 1)$ distribution, which makes the average treatment effect be $\tau^* = 1$, and finally T is randomly drawn from $\{0, 1\}$ with $prob(T = T^*) = .8$, so the misclassification probabilities are $b_0 = b_1 = .2$. This model is exactly identified using Theorem 2 and Corollary 1, so the GMM weighting matrix W is taken to be the identity matrix. The sample size is $n = 1000$ and the number of simulations is $10,000$.

This design makes $\tau(0) = \tau(2) = .4945$ and $\tau(1) = .6$ so the limiting value of the estimated average treatment that would be obtained if one did not correct for misclassification error is $E[\tau(V)] = .530$, that is, a bias of almost 47% of the true effect $\tau^* = 1$.

Estimation was done using constrained GMM, with the constraints being that the estimated probabilities of treatment r^* lie between .01 and .99, and that the misclassification probabilities b_0 and b_1 be nonnegative with $b_0 + b_1 \leq .99$. These constraints were imposed on estimation using the CML estimation procedure in GAUSS.

For evaluating estimation based on Corollary 2, the simulation design has Z empty. For each observation, a V is drawn from $\Omega = \{0, 1\}$ with $prob(V = v_k) = 1/2$, T^* is drawn from $\{0, 1\}$ with probabilities $r^*(0) = 3/4$ and $r^*(1) = 1/4$, and then Y and T are drawn the same as before. Estimation is again based on constrained GMM. This design is again exactly identified, and implemented with $n = 1000$ and $10,000$ simulations.

The third simulation design uses the local GMM estimator, with a normal kernel, bandwidth given by Silverman's rule, and the g function defined in Corollary 2. This simulation design has each observation Z drawn from a uniform on the interval $[-.2, .2]$ and V drawn from $\Omega = \{0, 1\}$ with $prob(V = v_k) = 1/2$. Next, each T^* is drawn from $\{0, 1\}$ with probabilities $r^*(V, Z)$ defined by $r^*(0, z) = 3/4 - z$ and $r^*(1, z) = 1/4 + z$. Then each outcome Y is drawn from from a normal $N((1 + Z)T^*, 1)$ distribution, which makes the conditional average treatment effect be $\tau^*(z) = 1 + z$, and finally T is randomly drawn from $\{0, 1\}$ with $prob(T = T^*) = .8$ as before. This design is constructed to have treatment probabilities treatment effects that vary with Z, but are the same as the second design at $Z = 0$. The local GMM estimator is applied local to $Z = 0$, and so provides estimates of the probabilities and treatment effects at (that is, conditional upon) $Z = 0$.

Table 1 reports the results. Estimation based on GMM with the two valued instrument is especially precise, with unbiased mean and median for τ^*, root mean squared error of .10 and median absolue error of .07. Estimation of the treatment effect τ^* based on the three valued instrument shows a 5% mean bias and

1% median bias, compared to the 47% bias that would result from failure to account for misclassification. The estimates of treatment and misclassification probabilities display similar levels of accuracy. The local GMM estimator also performs quite well, with only modest biases. The local GMM of course has larger errors (approximately double those of the second simulation), because it only makes significant use of the fraction of observations that are in the neighborhood of $Z = 0$, while the second simulation is equivalent to having $Z = 0$ for all observations.

6 Conclusions

This paper provides conditions for identification of treatment probabilities, misclassification probabilities, and average treatment effects when treatment may be mismeasured. Estimators that employ these identification conditions are provided, based on direct estimation of relevant conditional expections. It would useful to explore how other treatment effect estimators such as matching and propensity score based methods might be adapted to the present application where treatment is mismeasured.

References

[1] ABADIE, A. (2003), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," *Journal of Econometrics*, 113, 231-263.

[2] ABADIE, A. AND G. IMBENS, (2002), "Simple and Bias Corrected Matching Estimators for Average Treatment Effects," NBER working paper.

[3] ABREVAYA, J. AND J. A. HAUSMAN, (1999), ""Semiparametric Estimation With Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells", *Annales d'Economie et de Statistique*, 55/56, 243-275.

[4] AIGNER, D. J. (1973), "Regression With a Binary Independent Variable Subject to Errors of Obervation," *Journal of Econometrics*, 1, 249-60.

[5] ANGRIST, J., G. IMBENS, AND D. B. RUBIN, (1996), "Identificaiton of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-455.

[6] BALKE, A. AND J. PEARL, (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171-1176.

[7] BOLLINGER, C. R. (1996), "Bounding Mean Regressions When a Binary Regressor Is Mismeasured," *Journal of Econometrics,* 73, 387-399.

[8] BROWN, J. N., AND A. LIGHT, (1992), "Interpreting Panel Data on Job Tenure," *Journal of Labor Economics,"* 10, 219-257.

[9] CARD, D. (1995), "Using Geographic Variations in College Proximity to Estimate the Returns to Schooling," in *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp*," L. N. Christofides, E. K. Grand, and R. Swidinsky, eds., Toronto: University of Toronto Press.

[10] CARD, D. (1996), "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica,* 64, 957-979.

[11] CHUA, T. C. AND W. A FULLER, (1987), "A Model For Multinomial Response Error Applied to Labor Flows," *Journal of the American Statistical Association*, 82, 46-51.

[12] FINNEY, D. J. (1964) Statistical Method in Biological Assay. Havner: New York.

[13] GOZALO, P. AND O. LINTON, (2000), "Local Nonlinear Least Squares Estimation: Using Parametric Information Nonparametrically," Journal of Econometrics 99, .

[14] GUSTMAN, A. L. AND T. L. STEINMEIER, (2001), "What People Don't Know About Their Pension and Social Security," in Gale, Shoven, and Warshawsky, eds., *The Evolving Pension System: Trends, Effects, and Proposals for Reform,* Washington: Brookings Institution Press.

[15] HAHN, J., (1998), "On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315-331.

[16] HANSEN, L., (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

[17] HAUSMAN, J. A., J. ABREVAYA, AND F. M. SCOTT-MORTON (1998), "Misclassification of the Dependent Variable in a Discrete-Response Setting," *Journal of Econometrics,* 87, 239-269.

[18] HECKMAN, J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679-693.

[19] HECKMAN, J. (1976), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153-161.

[20] HECKMAN, J. (1990), "Varieties of Selection Bias," *American Economic Review, Papers and Proceedings*, 80, 313-338.

[21] HECKMAN ,J. H. ICHIMURA AND P. TODD, (1998), "Matching as an Econometric Evaluations Estimator, *Review of Economic Studies*, 65, 261-294.

[22] HECKMAN ,J. AND R. ROBB, (1985), "Alternate Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer, eds., *Longitudinal Analysis of Labor Market Data*, New York: Cambridge University Press.

[23] HECKMAN ,J. AND E. VYTLACIL, (2001), "Structural Equations, Treatment Effects and Econometric Policy Evaluation," unpublished manuscript.

[24] HIRANO, K., G. IMBENS, AND G. RIDDER, (2002), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," Unpublished manuscript.

[25] HIRSCH, B. T., (2003), "Reconsidering Union Wage Effects: Surveying New Evidence on an Old Topic," Unpublished manuscript, Trinity University.

[26] HOROWITZ, J. L. AND C. F. MANSKI, (2000), "Nonparametric Analysis of Randomized Experiments With Missing Covariate and Outcome Data, *Journal of the American Statistical Association*, 95, 77-84.

[27] IMBENS, G. W. AND J. D. ANGRIST, (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

[28] KANE, T. J., C. E. ROUSE, AND D. STAIGER, (1999), "Estimating Returns to Schooling When Schooling is Misreported," NBER working paper #7235.

[29] KLEPPER, S., (1988), "Bounding the Effects of Measurement Error in Regressions Involving Dichotomous Variables," *Journal of Econometrics,* 37, 343-359.

[30] LEWBEL, A., (2000), "Identification of the Binary Choice Model With Misclassification," *Econometric Theory*, 16, 603-609.

[31] LEWBEL, A., (2002), ""Endogeneous Selection or Treatment Model Estimation," Unpublished manuscript, Boston College.

[32] MANSKI, C. F. (1990) "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

[33] MANSKI, C. F. (1997) "Monotone Treatment Response," *Econometrica*, 65, 1311-1334.

[34] MCFADDEN, D., (1984), "Econometric Analysis of Qualitative Response Models," In: Griliches, Z., Intriligator, M.D.(Eds.), Handbook of Econometrics, vol. 2. North-Holland, Amsterdam.

[35] MOLINARI, F. (2001) "Identification of Probability Distributions With Misclassified Data" Unpublished manuscript, Northwestern University.

[36] MOLINARI, F. (2002) "Missing Treatments" Unpublished manuscript, Northwestern University.

[37] NEWEY, W. K. AND J. L. POWELL, (2003), "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71 1565-1578.

[38] POTERBA, J. M. AND L. H. SUMMERS (1995) "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model With Errors in Classification," *Review of Economics and Statistics*, 77, 207-216.

[39] POWELL, J. L., (1994), "Estimation of Semiparametric Models," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2444-2521, Amsterdam: Elsevier.

[40] ROBINS, J., (1997), "Non-Response Models for the Analysis of Non-Monotone Non-Ignorable Missing Data," *Statistics in Medicine*, 16, 21-37.

[41] ROBINS, J., S. MARK AND W. NEWEY, (1992), "Estimating Exposure Effects by Modeling the Expectations of Exposure Conditional on Confounders," *Biometrics*, 48, 479-495.

[42] ROSENBAUM, P. AND D. RUBIN, (1985), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516-524.

[43] RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies, " *Journal of Educational Psychology*, 76, 688-701.

[44] VYTLACIL, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*.

7 Appendix

PROOF OF THEOREM 1: Recall $r^*(x) = E(T^* \mid X = x)$.

By the definition of $r(x)$,

$$\begin{aligned} r(x) &= \Pr(T=1 \mid X=x, T^*=1)\Pr(T^*=1 \mid X=x) + \Pr(T=1 \mid X=x, T^*=0)\Pr(T^*=0 \mid X=x) \\ &= [1-b_1(x)]r^*(x) + b_0(x)[1-r^*(x)] \end{aligned}$$

So $r^*(x) = [r(x) - b_0(x)]/[1 - b_0(x) - b_1(x)]$.

Next, define

$$\begin{aligned} p_t(x) &= E[I(T^*=t) \mid X=x, T=t] \\ r_t^*(x) &= E[I(T^*=t) \mid X=x] \end{aligned}$$

so $r_1^*(x) = r^*(x)$ and $r_0^*(x) = 1 - r^*(x)$. By Bayes rule

$$\begin{aligned} p_t(x) &= \frac{\Pr(T=t \mid X=x, T^*=t)\Pr(T^*=t \mid X=x)}{\Pr(T=t \mid X=x)} \\ &= \frac{\Pr(T=t \mid X=x, T^*=t)\Pr(T^*=t \mid X=x)}{\Pr(T=t \mid X=x, T^*=t)\Pr(T^*=t \mid X=x) + \Pr(T=t \mid X=x, T^*=1-t)\Pr(T^*=1-t \mid X=x)} \\ &= \frac{[1-b_t(x)]r_t^*(x)}{[1-b_t(x)]r_t^*(x) + b_{1-t}(x)[1-r_t^*(x)]} = \frac{[1-b_t(x)]r_t^*(x)}{[1-b_t(x)-b_{1-t}(x)]r_t^*(x) + b_{1-t}(x)} \end{aligned}$$

So

$$\begin{aligned} p_1(x) &= \frac{[1-b_1(x)]r^*(x)}{[1-b_1(x)-b_0(x)]r^*(x) + b_0(x)} \\ &= \frac{[1-b_1(x)][r(x)-b_0(x)]}{[1-b_1(x)-b_0(x)]r(x)} \\ p_0(x) &= \frac{[1-b_0(x)][1-r^*(x)]}{[1-b_0(x)-b_1(x)][1-r^*(x)] + b_1(x)} \end{aligned} \tag{16}$$

$$= \frac{[1 - b_0(x)][1 - r(x) - b_1(x)]}{[1 - b_1(x) - b_0(x)][1 - r(x)]} \tag{17}$$

Now define $h_t^*(x) = E(Y \mid X = x, T^* = t)$. By Assumption A2, $h_t^*(x) = E(Y \mid X = x, T^* = t, T)$ and so

$$\begin{aligned} E(Y \mid X = x, T = t) &= \Sigma_{j=0}^{1} E(Y \mid X = x, T = t, T^* = j) \Pr(T^* = t \mid X = x, T = j) \\ &= h_t^*(x) p_t(x) + h_{1-t}^*(x)[1 - p_t(x)] \end{aligned} \tag{18}$$

Substituting this expression into the definition of $\tau(x)$ gives

$$\begin{aligned} \tau(x) &= h_1^*(x) p_1(x) + h_0^*(x)(1 - p_1(x)) - [h_0^*(x) p_0(x) + h_1^*(x)(1 - p_0(x))] \\ &= [h_1^*(x) - h_0^*(x)][p_1(x) + p_0(x) - 1] \\ &= \tau^*(x)[p_1(x) + p_0(x) - 1] \end{aligned}$$

so $\tau(x) = \tau^*(x) m[b_0(x), b_1(x), r(x)]$ where

$$\begin{aligned} m[b_0(x), b_1(x), r(x)] &= [p_1(x) + p_0(x) - 1] \\ &= \frac{[1 - b_1(x)][r(x) - b_0(x)]}{[1 - b_1(x) - b_0(x)] r(x)} + \frac{[1 - b_0(x)][1 - r(x) - b_1(x)]}{[1 - b_1(x) - b_0(x)][1 - r(x)]} - 1 \\ &= \left(\frac{1}{1 - b_1(x) - b_0(x)}\right)\left(1 - \frac{[1 - b_1(x)] b_0(x)}{r(x)} - \frac{[1 - b_0(x)] b_1(x)}{1 - r(x)}\right). \end{aligned}$$

For identification, $E[Y(1) - Y(0) \mid X = x] = \tau^*(x)$ by Assumption A1, and $\tau^*(x) = \tau(x)/m[b_0(x), b_1(x), r(x)]$, where $m[b_0(x), b_1(x), r(x)]$ is identified by identification of $b_0(x)$, $b_1(x)$, and $r(x)$. Assumption A3 ensures that $m[b_0(x), b_1(x), r(x)]$ is finite and nonzero.

PROOF OF THEOREM 2: For a given z, we have for all $v \in \Omega$, using Theorem 1,

$$\frac{\tau(v, z)}{\tau(v_0, z)} = \frac{m[b_0(z), b_1(z), r(v, z)]}{m[b_0(z), b_1(z), r(v_0, z)]}$$

To ease notation further, drop z. Then $m[b_0, b_1, r(v)]\tau(v_0) - m[b_0, b_1, r(v_0)]\tau(v) = 0$ and substituting in for m gives

$$0 = \left(1 + \frac{(b_1 - 1) b_0}{r(v)} + \frac{(b_0 - 1) b_1}{1 - r(v)}\right) \tau(v_0) - \left(1 + \frac{(b_1 - 1) b_0}{r(v_0)} + \frac{(b_0 - 1) b_1}{1 - r(v_0)}\right) \tau(v) \tag{19}$$

$$0 = (1-b_1)b_0\left(\frac{\tau(v_0)}{r(v)} - \frac{\tau(v)}{r(v_0)}\right) + (1-b_0)b_1\left(\frac{\tau(v_0)}{1-r(v)} - \frac{\tau(v)}{1-r(v_0)}\right) + \tau(v) - \tau(v_0) \tag{20}$$

Evaluate this equation at $v = v_k$, and rewrite it as

$$0 = B_0 w_{0k} + B_1 w_{1k} + w_{2k}$$

where $B_t = (1-b_{1-t})b_t$ and each w_{jk} is a function of $r(v_0)$, $r(v_k)$, $\tau(v_0)$, and $\tau(v_k)$. Given that Ω contains three elements v_0, v_1, and v_2, we have two equations $0 = B_0 w_{0k} + B_1 w_{1k} + w_{2k}$ for $k = 1, 2$ that are linear in the two unknowns B_0 and B_1, and so can be uniquely solved as long as the matrix of elements w_{jk}, $j = 0, 1$, $k = 1, 2$, is nonsingular. The inequality in Assumption A6 makes the determinant of this matrix nonzero, as required.

Now let $s = 1-b_1-b_0$. It follows from $B_t = (1-b_{1-t})b_t$ that $(s+b_0)b_0 = B_0$ and $2b_0 = B_0 - B_1 + 1 - s$. Substituting the second of these equations into the first and solving for s gives

$$1 - b_1 - b_0 = s = \pm\left[(B_0 - B_1 + 1)^2 - 4B_0\right]^{1/2}$$

if the assumption regarding s is $s \neq 0$, then we have that s is identified up to sign. By Theorem 1 $\tau^* = \tau(v)/m[b_0, b_1, r(v)]$ and

$$m[b_0, b_1, r(v)] = \left(\frac{1}{s}\right)\left(1 - \frac{B_0}{r(v)} - \frac{B_1}{1-r(v)}\right)$$

so it follows that $\tau^*(x)$ is identified up to sign. Making the stronger assumption that $s > 0$, we have s is identified, so b_0 and b_1 are now identified by $b_0 = (B_0 - B_1 + 1 - s)/2$ and $b_1 = -(B_0 - B_1 + 1 + s)/2$, and by Theorem 1, identification of these misclassification probabilities means that $r^*(x)$ and $\tau^*(x)$ are also identified.

PROOF OF THEOREM 3: Assumption B4 implies Assumption A4, so Theorem 1 holds. By equations (17), (16), (18), and Assumption B5,

$$h_1(v,z) = h_0^*(z) + [h_1^*(z) - h_0^*(z)]\frac{[1-b_1(z)][r(v,z) - b_0(z)]}{[1-b_1(z)-b_0(z)]r(v,z)} \tag{21}$$

$$h_0(v,z) = h_1^*(z) + [h_0^*(z) - h_1^*(z)]\frac{[1-b_0(z)][1-r(v,z) - b_1(z)]}{[1-b_1(z)-b_0(z)][1-r(v,z)]} \tag{22}$$

Dropping z again to ease notation,

$$h_1(v) - h_1(v_0) = \frac{(h_1^* - h_0^*)(1-b_1)b_0}{1-b_1-b_0}\left(\frac{1}{r(v)} - \frac{1}{r(v_0)}\right)$$

$$h_0(v) - h_0(v_0) = -\frac{(h_1^* - h_0^*)(1-b_0)b_1}{1-b_1-b_0}\left(\frac{1}{1-r(v)} - \frac{1}{1-r(v_0)}\right)$$

$$0 = \left(\frac{h_1(v) - h_1(v_0)}{[r(v_1)]^{-1} - [r(v_0)]^{-1}}\right)(1-b_0)b_1 + \left(\frac{h_0(v) - h_0(v_0)}{[1-r(v_1)]^{-1} - [1-r(v_0)]^{-1}}\right)(1-b_0)b_1$$

Evaluate this equation and equation (20) at $v = v_1$, and rewrite the pair as

$$0 = B_0 u_{0s} + B_1 u_{1s} + u_{2s}$$

where $B_t = (1-b_{1-t})b_t$ and each u_{js} is a function of $h_1(v_k)$, $h_0(v_k)$, $r(v_k)$, and $\tau(v_k)$ for $k = 0, 1$. Given that Ω contains three elements v_0 and v_1, we have two equations $0 = B_0 u_{0s} + B_1 u_{1s} + u_{2s}$ for $s = 1, 2$ that are linear in the two unknowns B_0 and B_1, and so can be uniquely solved as long as the matrix of elements u_{js}, $j = 0, 1$, $s = 1, 2$, is nonsingular. The inequality in Assumption B6 makes the determinant of this matrix nonzero, as required. Given identification of B_0 and B_1, the remainder of the proof is then identical to last part of the proof of Theorem 2.

PROOF OF COROLLARY 1: To ease notation, drop the argument z everywhere and let all expectations below be conditional on $Z = z$. Let $I_k = I(V = v_k)$. Having the mean of equation (7) equal zero makes $b_0 + (1-b_0-b_1)r_k^* = E(I_k T)/E(I_k)$, which equals the true r_k.by definition of r_k. Solving the resulting equation $b_0 + (1-b_0-b_1)r_k^* = r_k$ for r_k^* and substituting the result into equation (8) gives

$$\left(\frac{YT}{r_k} - \frac{(1-b_1)\tau^*}{r_k}\frac{r_k - b_0}{(1-b_0-b_1)} - \frac{Y(1-T)}{1-r_k} - \frac{(1-b_0)\tau^*}{1-r_k}\frac{1-b_1-r_k}{(1-b_0-b_1)} + \tau^*\right) I_k$$

Setting the mean of this result to zero and dividing by $E(I_k)$ gives

$$\frac{E(YTI_k)}{r_k E(I_k)} - \frac{(1-b_1)\tau^*}{r_k}\frac{r_k - b_0}{(1-b_0-b_1)} - \frac{E[Y(1-T)I_k]}{(1-r_k)E(I_k)} - \frac{(1-b_0)\tau^*}{1-r_k}\frac{1-b_1-r_k}{(1-b_0-b_1)} + \tau^* = 0$$

which, using $r_k = E(TI_k)/E(I_k)$ simplifies to

$$\frac{E(YTI_k)}{E(TI_k)} - \frac{(1-b_1)\tau^*}{r_k}\frac{r_k - b_0}{(1-b_0-b_1)} - \frac{E[Y(1-T)I_k]}{E[(1-T)I_k]} - \frac{(1-b_0)\tau^*}{1-r_k}\frac{1-b_1-r_k}{(1-b_0-b_1)} + \tau^* = 0.$$

which, after rearranging terms and using $E(TI_k) = prob(T = 1, V = v_k)$ gives

$$\begin{aligned} E(Y \mid T = 1, V = v_k) - E(Y \mid T = 0, V = v_k) \\ = \left(\frac{(1-b_1)}{r_k} \frac{r_k - b_0}{(1-b_0-b_1)} + \frac{(1-b_0)}{1-r_k} \frac{1-b_1-r_k}{(1-b_0-b_1)} - 1 \right) \tau^* \end{aligned}$$

which, by the definitions of the function τ and m reduces to $\tau(v_k) = m(b_0, b_1, r_k)$. It has now been shown that the conditional mean of g equalling zero equivalent to $r(v_k) = b_0 + (1 - b_0 - b_1)r_k^*$ and $\tau(v_k) = m[b_0, b_1, r(v_k)]\tau^*$ with the true functions $r(v_k)$ and $\tau(v_k)$, and by Theorem 2 the only solutions to these equations for $k = 0, ..., K$ that also satisfy $b_0 \geq 0$, $b_1 \geq 0$, and $b_0 + b_1 < 1$ are the true values of $r_0^*, ..., r_K^*, b_0, b_1$, and τ^*.

PROOF OF COROLLARY 2: Again drop z as in the proof of Corollary 1. By that proof 1 we have that the conditional mean of g equalling zero is equivalent to $r(v_k) = b_0 + (1 - b_0 - b_1)r_k^*$, $\tau(v_k) = m[b_0, b_1, r(v_k)]\tau^*$, and the mean of equation (10) equaling zero, using the true functions $r(v_k)$ and $\tau(v_k)$. Setting the mean of (10) equal to zero and dividing the result by $E(I_k)[b_0 + (1 - b_0 - b_1)r_k^*] = E(I_k)r_k$ gives

$$\frac{E(YTI_k)}{r_k E(I_k)} - \frac{(1-b_1)r_k^*}{(1-b_1-b_0)r_k^* + b_0}\tau^* - h_0^* = 0$$

which is equivalent to equation (21). This can be subtracted from $\tau(v_k) - m[b_0, b_1, r(v_k)]\tau^* = 0$ to obtain (22), and by the proof of Theorem 3, equations (22), (21) and $r(v_k) = b_0 + (1 - b_0 - b_1)r_k^*$ along with $b_0 \geq 0$, $b_1 \geq 0$, and $b_0 + b_1 < 1$ uniquely identify the true values of $r_0^*, ..., r_K^*, b_0, b_1$, and τ^*.

PROOF OF COROLLARY 3: Setting the conditional mean of equation (12) equal to zero and solving for $r_k(z)$ yields the definition of $r_k(z)$, and setting the conditional mean of equation (13) equal to zero and solving for $\tau_k(z)$ yields the definition of $\tau_k(z)$.

PROOF OF THEOREM 4: Define

$$\begin{aligned} S_n'(q(z)) &= \frac{\partial S_n(q(z))}{\partial q(z)^T} = \frac{1}{nb^d} \sum_{i=1}^{n} \frac{\partial g[q(z), W_i]}{\partial q(z)^T} K\left(\frac{z - Z_i}{b}\right) \\ Q_n(q(z)) &= S_n(q(z))^T \Omega_n S_n(q(z)) \end{aligned}$$

Let $S_0(q(z)) = plim_{n\to\infty} S_n(q(z))$ and similarly for S_n' and Q_n. Assumptions C1 and C2 give sufficient conditions for consistency of these kernel estimators, so these probability limits exist and

$$\begin{aligned} S_0(q(z)) &= E[g(q(z), W) f(z) \mid Z = z] \\ Q_0(q(z)) &= S_0(q(z))^T \Omega_0 S_0(q(z)). \end{aligned}$$

Now consider consistency. We have pointwise convergence of $S_n(q(z))$ to $S_0(q(z))$and compactness of $\Theta(z)$. It is also the case that $|S_n'(q(z))| = O_p(1)$, since $|S_n'(q(z))|$ is a kernel estimator, and standard conditions have been provided for its consistency, that is, $plim|S_n'(q(z))| = E[|\partial g(W, q(z))/\partial q(z)| f(z) \mid Z = z]$. This suffices for stochastic equicontinuity, and therefore we have the uniform convergence

$$plim \sup_{q(z)\in\Theta(z)} |S_n(q(z)) - S_0(q(z))| = 0.$$

It follows that $Q_n(q(z))$ also converges uniformly to $Q_0(q(z))$. The assumptions provide compactness of $\Theta(z)$ and imply continuity of $Q_0(q)$. The quadratic form of Q_0 is uniquely maximized at $S_0(q_0(z)) = 0$ and hence at $q(z) = q_0(z)$, so the standard conditions for consistency $plim\widehat{q}(z) = q_0(z)$.are satisfied.

For the limiting distribution, Taylor expanding the first order conditions as in standard GMM gives

$$S_n'(\widehat{q}(z))^T \Omega_n \left[S_n(q_0(z)) + S_n'(\widetilde{q}(z))(\widehat{q}(z) - q_0(z)) \right] = 0$$

where $\widetilde{q}(z)$ lies between $\widehat{q}(z)$ and $q_0(z)$. By consistency of $\widehat{q}$, the uniform convergence of S_n, and using $R(z) = S_0'(q_0(z))$, this simplifies to

$$R(z)^T \Omega \left[S_n(q_0(z)) + R(z)(\widehat{q}(z) - q_0(z)) \right] = o_p(1)$$

Solving for $\widehat{q}(z) - q_0(z)$ and multiplying by $(nb)^{1/2}$ gives

$$(nb)^{1/2}(\widehat{q}(z) - q_0(z)) = (R(z)^T \Omega R(z))^{-1} R(z)^T \Omega (nb)^{1/2} S_n(q_0(z)) + o_p((nb)^{1/2}).$$

Now $S_0(q_0(z)) = 0$ and standard kernel regression limiting distribution theory gives

$$(nb)^{1/2} S_n(q_0(z)) \to^d N[0, V(q_0(z)) \int K(u)^2 du]$$

and the theorem follows.

Table 1. Simulation Results

Three Valued Instrument GMM

	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
$r^*(0)$	.750	.722	.105	.662	.734	.797	.109	.083	.067
$r^*(1)$	.500	.481	.100	.414	.484	.553	.102	.083	.071
$r^*(2)$	.250	.243	.115	.157	.231	.323	.115	.095	.085
b_0	.200	.193	.097	.133	.213	.268	.097	.077	.068
b_1	.200	.180	.074	.124	.190	.237	.076	.061	.053
τ^*	1.00	1.05	.298	.825	1.01	1.24	.302	.239	.203

Two Valued Instrument GMM

	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
$r^*(0)$	.750	.749	.061	.710	.751	.790	.061	.048	.040
$r^*(1)$	.250	.251	.061	.208	.248	.289	.061	.048	.040
b_0	.200	.197	.045	.170	.200	.228	.045	.035	.029
b_1	.200	.197	.045	.169	.199	.228	.045	.035	.030
τ^*	1.00	1.00	.101	.937	1.00	1.07	.101	.081	.068

Two Valued Instrument Local GMM

	TRUE	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE
$r^*(0)$	.750	.745	.113	.672	.753	.826	.113	.091	.077
$r^*(1)$	.250	.253	.114	.171	.245	.326	.114	.091	.078
b_0	.200	.189	.088	.134	.197	.252	.088	.070	.058
b_1	.200	.192	.083	.137	.196	.250	.083	.067	.056
τ^*	1.00	1.03	.199	.890	1.02	1.16	.201	.159	.133

Notes: The reported statistics are as follows. TRUE is the true value of the parameter, MEAN and SD are the mean and standard deviation of the estimates across the simulations. LQ, MED, and UQ are the 25% (lower) 50% (median) and 75% (upper) quartiles. RMSE, MAE, and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates.