Does the Food Stamp Program Really Increase Obesity? The Importance of Accounting for Misclassification Errors

Vassilopoulos, Achilleas and Drichoutis, Andreas and Nayga, Rodolfo and Lazaridis, Panagiotis

Agricultural University of Athens, University of Ioannina, University of Arkansas

January 2011

Does the Food Stamp Program Really Increase Obesity?

The Importance of Accounting for Misclassification Errors

Achilleas Vassilopoulos[1], Andreas C. Drichoutis[2], Rodolfo M. Nayga, Jr.[3] and Panagiotis Lazaridis[4]

[1] Dept. of Agricultural Economics and Rural Development, Agricultural University of Athens, Greece, email: avassilopoulos.aua(at)gmail.com

[2] Dept. of Economics, University of Ioannina, Greece, email: adrihout(at)cc.uoi.gr

[3] Dept. of Agricultural Economics and Agribusiness, University of Arkansas, USA, email: rnayga(at)uark.edu

[4] Department of Agricultural Economics and Rural Development, Agricultural University of Athens, Greece, email: t.lazaridis(at)aua.gr

Abstract

Over the last few decades, the prevalence of obesity among US citizens has grown rapidly, especially among low-income individuals. This has led to questions about the effectiveness of nutritional assistance programs such as the Supplemental Nutrition Assistance Program (SNAP), formerly known as the Food Stamps Program (FSP). Results from previous studies generally suggest that FSP participation increases obesity. This finding is however based on analyses that assumed that participants do not misclassify their program participation. Significant misclassification errors have been reported in the literature. Using propensity score matching estimation and a new method to conduct extensive sensitivity analysis, we find that this finding is quite sensitive to misclassification errors above 10% and to functional form assumptions.

***JEL* codes:** C63, D12, I1

Key Words: matching estimators, sensitivity analysis, food stamps, obesity

Does the Food Stamp Program Really Increase Obesity?

The Importance of Accounting for Misclassification Errors

Abstract

Over the last few decades, the prevalence of obesity among US citizens has grown rapidly, especially among low-income individuals. This has led to questions about the effectiveness of nutritional assistance programs such as the Supplemental Nutrition Assistance Program (SNAP), formerly known as the Food Stamps Program (FSP). Results from previous studies generally suggest that FSP participation increases obesity. This finding is however based on analyses that assumed that participants do not misclassify their program participation. Significant misclassification errors have been reported in the literature. Using propensity score matching estimation and a new method to conduct extensive sensitivity analysis, we find that this finding is quite sensitive to misclassification errors above 10% and to functional form assumptions.

***JEL* codes:** C63, D12, I1

Key Words: matching estimators, sensitivity analysis, food stamps, obesity

1. Introduction

Obesity is increasing worldwide in dramatic rates. The World Health Organization indicated that there were 1.6 billion overweight adults and at least 400 million obese adults in the world in 2005 (WHO (2006) Obesity and Overweight. Fact Sheet No 311. World Health Organization). By 2015, these figures are expected to rise to 2.3 billion overweight and 700 million obese adults. Obesity effects on health are well supported

by the medical literature and include a long non-exhaustive list that includes osteoarthritis, sleep apnea, asthma, high blood pressure, gallbladder disease, cholesterol, type II diabetes, cardiovascular disease, stroke, renal and genitourinary diseases (Bray, 2004; Ejerblad et al., 2006; Esposito et al., 2004; Grundy, 2004; van der Steeg et al., 2007; Whitmer et al., 2005). Obesity may also inflict severe emotional harm, such as social stigmatization, depression, and poor body image.

Researchers have rightly responded to this unprecedented rise of obesity as evidenced by the exploding number of papers published in the nutrition/medical as well as the economics literature. The economic causes of obesity are nicely analyzed in Rosin (2008). Among the many factors linked to the high obesity prevalence are the increased opportunity cost of time for food preparation, along with the availability of "cheap" calories provided by fast-food restaurants, as well as the adoption of sedentary lifestyles (Cutler et al., 2003; Lakdawalla et al., 2005; Philipson and Posner, 2003). One interesting aspect of the obesity epidemic is that prevalence rates have been found to be higher and to increase more rapidly among lower income people, a group usually associated with fewer resources and poor diets.

To this respect, a number of nutrition assistance programs funded by the U.S. government target specific groups of low income people to address dietary and nutrition concerns. The Supplemental Nutrition Assistance Program (SNAP), formerly known as the Food Stamps Program (FSP)[1] is by far the largest nutrition assistance program in the US. The FSP as implemented in 1964 was designed to alleviate hunger by distributing coupons that could only be used to purchase food at grocery stores. FSP benefits are given to a single person or family who meets the program's requirements pertaining to income, assets, work and immigration status.

[1] For the rest of the paper we use the term FSP rather than SNAP since this program is still more popularly known as the food stamps program.

Most benefit periods last for 6 months but some can be as short as 1 month or as long as 3 years. Currently, electronic benefit transfers that operate essentially as debit cards have replaced food stamp coupons. According to USDA data, about 40 million individuals and 18 million households nation-wide participate in this program, with total amount of benefits reaching 65 billion USD in 2010. Eligibility and benefits are based on household size, household assets, and income. Other food assistance programs in the US include the School Breakfast Program (SBP), the National School Lunch Program (NSLP) and the Women, Infant and Children Program (WIC).

Due to increasing obesity rates in the US, particularly among low-income individuals, we focus our study on assessing the effect of FSP on obesity. There are two main theories on how food stamp benefits could contribute to weight gain: (1) food stamps encourage beneficiaries to spend more money on food than they otherwise would (and presumably, to eat more); and (2) food stamp participation is linked to a cycle of deprivation followed by abundance and binge eating, which results in weight gain over time (Ver Ploeg et al. 2007).

Several studies have examined the effect of FSP participation on various outcomes. These studies differ in terms of the targeted groups (e.g., children, adult women/men and the elderly), the outcomes of interest (e.g., Body Mass Index, food security index, probability of being overweight/obese), the nature of the data (e.g., cross-sectional, longitudinal), the sources of the data[2] and the methodology they employ[3]. Results from a number of past studies suggest a positive effect of FSP participation on adult obesity. For example, Baum (2007) found that FSP participation

[2] The Panel Study of Income Dynamics (PSID) along with the Child Development Supplement (CDS), the National Health and Nutrition Examination Survey (NHANES), the National Longitudinal Survey of Youth (NLSY79), the Health and Retirement Study (HRS) and the Asset and Health Dynamics Among the Old (AHEAD) are some of them.

[3] Descriptive statistics, OLS and Logistic Regressions, IV estimators (with and without fixed effects), Bivariate Probit, Dynamic and Lagged Models, Hazard Models and Propensity Score Matching.

increases the probability of being obese in females aged 20-28 while the amount of food stamps benefit was positively related to BMI in males of the same age. Gibson (2003) concluded that FSP participation is responsible for a 2 percentage point increase in the BMI of adult women. This effect was even greater in the case of long-term participation of women. Chen et al. (2005) also showed that women FSP recipients have an obesity rate that is 6.7 percent higher than that of women non-recipients. On the other hand, Kaushal (2007) found no significant effect of FSP participation on obesity of both men and women.

These past studies, however, did not take into account the misclassification errors associated with self-reported FSP participation status. This issue is important since results could be sensitive to these misclassification errors that have been reported in the literature to be non-trivial. For instance, Bollinger & David (1997) and Bitler et al. (2003) suggest that about 10%-15% of FSP recipients do not report FSP participation when asked by the interviewer. Meyer et al. (2010) report an even higher 35%-50% misclassification error. If this is the case, then the findings from past studies that associated FSP participation with increasing obesity could be biased and misleading. Our objective in this study is to assess the effect of misclassification errors on the effects of FSP participation on obesity. We also take into account the complex endogeneity issues inherent in these types of analysis and extensively test the robustness of our results to deviations from functional form assumptions. Specifically, we simulate various scenarios where we vary percentages of misreported participation in the FSP.

Also, since FSP participation is not randomly but rather endogenously assigned to subjects according to some observable (e.g., eligibility criteria) and unobservable (e.g., information acquisition, attitudes etc.) factors, there is always the

chance that some of these factors are highly correlated with the outcome of interest, making it hard to identify the causal effect of FSP participation without a proper identification strategy. Hence, if one or more variables influencing FSP participation also affect BMI, then simple descriptive statistics or regression analysis of BMI or the probability of being overweight/obese on a binary variable indicating FSP participation, would be biased due to the well known problem of self-selection. On the other hand, if one employs techniques which are designed to circumvent self-selection bias (e.g., selection models, IV estimators), obtaining a consistent point estimate requires a valid exclusion restriction which is not always plausible with typical data limitations. Lastly, estimators that do not rely on such restrictions but account for selection bias (e.g Heckman's bivariate normal selection, propensity score matching) assume specific functional forms for identification, thus failing to reveal the true underlying patterns when these functional forms are not known. In this study, we employ propensity score matching estimators and perform an extensive sensitivity analysis to test the robustness of restrictive assumptions. Quoting Angrist & Pischke (2010), scrutinizing our results through a sensitivity analysis process is what takes the con out of the econometrics.

We build on the work of Ichino et al. (2008) who proposed an excellent way of testing the robustness of matching estimators while avoiding parametric assumptions. We extend this method to account for the misclassification errors which are common in some areas of interest, such as the one involving FSP participation. Our data come from the 2005–2006 National Health and Nutrition Examination Survey (NHANES). NHANES is designed to assess the health and nutritional status of adults and children in the US and is unique in that it combines interviews and physical examinations.

2. Methods

2.1 Propensity score matching

The research question of interest to us is whether participating in the FSP increases the probability of being obese. Formally, assume that there is a binary variable Y_i that takes the value of 1 if respondent *i* has both a BMI[4] greater than 30 kg/m^2 and a waste circumference (WC) greater than 100 cm, and 0 otherwise. The second condition (WC > 100 cm) is usually added in order to account for the misleading classification of BMI when it comes to athletes or elder people. Define a second binary variable T_i, which equals 1 for participants and 0 for the non-participants. In notation form:

$$Y_i = \begin{cases} 1, \text{ if } \mathrm{BMI}_i \geq 30 \text{ and } \mathrm{WC}_i \geq 100 \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

and

$$T_i = \begin{cases} 1, \text{ if individual is receiving FSP benefits} \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

A mere comparison of the mean Y_i, namely the obesity rate, between the treated and the control group does not reveal a causal relationship between the FSP participation and the outcome of interest. It is likely that the two groups differ in many other characteristics that could lead to differences in the mean Y_i even if food stamps were not received by either group. If we denote by $Y(1)$ the potential

[4] The BMI (Body Mass Index) is used to define nutritional status and is derived from the division of Weight in kilograms by the square of height in meters. The acceptable range is the same for men and women and lies between 20 and 25. Obesity is taken to start at a BMI of 30 and gross obesity at 40. A BMI of 18-20 is defined as mild starvation and severe starvation begins when BMI falls below 16.

outcome for the treated population and by $Y(0)$ the potential outcome for the same individuals, have they not been treated, we can define the effect of the treatment on each treated individual as:

$$t_i = Y_i(1) - Y_i(0) \tag{3}$$

which averaged over the population gives us the average treatment effect on the treated (ATT), namely:

$$t_{ATT} = E\left[Y(1) \mid T = 1\right] - E\left[Y(0) \mid T = 1\right] \tag{4}$$

Since $E[Y(0) \mid T = 1]$ is not observed, one needs to make some additional assumptions in order to estimate the ATT. The first is that Y, conditional on a set of observable covariates C, does not influence participation in the program:

$$Y_0 \perp T \mid C, \ \forall C \tag{5}$$

This assumption is widely known as the Conditional Independence Assumption (CIA), the restrictive nature of which seems unappealing to many researchers and decreases the popularity of matching estimators. A second assumption is the common support or overlap condition, which ensures that for every FSP participant, there are non-participants with the same observable covariates, that is:

$$Pr\left(T = 1 \mid C\right) < 1 \tag{6}$$

In all our estimations, we ensure that observations falling out of the common support region are excluded. If assumptions (5) and (6) hold, then after conditioning on a vector of *C* covariates, ATT becomes estimable through (4) by substituting the unobservable part $E[Y(0) \mid T = 1, C]$, with its observable counterpart, $E[Y(0) \mid T = 0, C]$. To solve the dimensionality problem arising when *C* includes many covariates,

Rosenbaum and Rubin (1983) suggested the use of the propensity score $P(C_i) = \Pr(T_i = 1|C_i)$, instead of C as the conditioning variable. The propensity score matching estimator of ATT is then given by:

$$ATT_{PSM} = E\left[Y(1)|T=1, P(C)\right] - E\left[Y(0)|T=0, P(C)\right] \qquad (7)$$

In our case and inasmuch as the FSP is designed to help low-income groups, we used as a control group the eligible non-recipients that are classified using the most important eligibility criterion of FSP participation, the Poverty Income Ratio (PIR). The PIR is also the only available eligibility criterion in our dataset. Other unobservable characteristics that cannot be controlled for (and which could render CIA implausible) are less likely to differ among individuals of these two classes.

ATT can be estimated using several matching algorithms such as nearest neighbor, kernel, stratification, radius and spline smoothing. We use nearest neighbour propensity score matching, using the four nearest neighbours[5] and report analytical standard errors since the bootstrap variance estimator is invalid for nearest neighbor matching (Abadie & Imbens 2006). The variables used to estimate the propensity score are shown in Table 1. The first two variables are the outcome and the control variable respectively. In selecting the variables to be included in the propensity score estimation we rely on the advice of Rubin & Thomas (1996) and the evidence provided by Brookhart et al. (2006) that one should include in the estimation of propensity scores all variables that are thought to be related to the outcome, regardless of whether they are related to the exposure. Household's FSP participation (*FS_hh*) was used instead of the individual's participation status since FSP benefits

[5] When selecting the number of matches one has to consider the bias-variance trade-off, since utilizing multiple matches for each treated individual will generally increase bias (2nd, 3rd, and 4th closest matches are, by definition, farther away from the treated individual than is the 1st closest match) while on the other hand, it can decrease variance due to the larger matched sample size (Stuart 2010). We use four matches in order not to rely on too little information but to also avoid incorporating observations that are not sufficiently similar. Like all smoothing parameters, the final inference can depend on the choice of the number of matches (Abadie et al. 2004).

are most certainly shared among the members of the household. For the same reason, WIC participation (*WIC_hh*) was included in the set of covariates. Other factors such as *Alcohol* and *Smoker* were included to account for the non-food expenses of the groups which could reduce available resources for food and decrease or increase the probability of being obese. *Chronic* and *DocDiab* are used to account for the possible links between these different conditions and obesity. To account for the absence of spatial information, we use pseudo-strata and pseudo-primary sampling unit dummies ($PSU_{1\text{-}2,}$ $Strata_{1\text{-}15}$) that are available in NHANES[6]. Square, cubic and interaction terms for all continuous variables and their transformations were also included in the model. Millimet and Tchernis (2009) showed that over-specifying the model used to estimate the propensity score is always the best strategy, considering the penalty associated with the under-specification of the model. Finally, we include demographic variables such as age, ethnic characteristics, educational level, income, marital status, and household size to capture the biological differences affecting BMI, the awareness about nutrition issues as well as the within-household consumption dynamics in the allocation of resources. The results of the above procedure will be referred to as the Unconfounded Baseline Estimates (UBE*s*) and are then tested in ways described below.

2.2. Misclassification errors

Due to the self-reported nature of the FSP participation data at hand, (7) is not estimable, since what we observe is not T_i but $T_{i,obs}$. The difference between T_i and $T_{i,obs}$ depends on whether the individuals that stated non-participation in FSP were

[6] Actual strata and primary sampling units are not to be disclosed due to the risk of identification of the respondents. However, their 'pseudo' counterparts are designed to give the exact same results.

actually non-recipients or not[7]. Given that there is no way to identify the subjects that made a false-statement, it might seem tempting to use the probability $\Pr\left(T_i = 1 \mid T_{i,obs} = 0\right)$ as a weight in order to derive the two right-hand side parameters of (7) by their $T_{i_{obs}}$'s counterparts. Nevertheless, even if such weights were available, the misclassification of subjects in the treated and the control groups would have caused a severe bias to these observable counterparts through $P(C)$, thereby making the results completely uninformative. In addition, in the estimation with a confounder (analyzed in the next section) if the researcher assumes that its distribution follows that of a known variable, there would be no possible way to define the parameters $\Pr_{ij}$ that characterize its distribution, since actual *i's* are not known.

We circumvent this problem by simulating different scenarios where a respondent that reported not to have received food stamps misrepresents her true state of participation by some probability $\Pr\left(T_i = 1 \mid T_{i,obs} = 0\right)$. To avoid further functional form assumptions about the probability distribution, we test different misclassification values[8] in an attempt to discover a cut-off point, beyond which our results fall flat. Specifically, we test misclassification error percentages of 5, 10, 15, 20, and 25 to cover the possible misclassification errors suggested by Bollinger & David (1997), Bitler et al. (2003), and Meyer et al. (2010). To accomplish this, we created *m* new databases[9] for each level of misclassification errors (i.e. 5x*m* in total), with each of these datasets containing all the variables that are exactly as in the original database

[7] Although we ignore the portion of individuals acting the other way around (i.e., reporting being FSP participants while they are not) due to the fact that it is usually a small group, the methodology proposed can be easily extended to include this option as well.

[8] Note that these values can be further decomposed into $\Pr\left(T_i = 1 \mid T_{i,obs} = 0, Y = 1\right)$ and $\Pr\left(T_i = 1 \mid T_{i,obs} = 0, Y = 0\right)$, if the researcher has strong evidence or a meaningful explanation on why the probability of misclassification can be related to the outcome of interest.

[9] We used m=1000 but we keep this notation for demonstration simplicity.

and a new participation indicator (*FS_hh_new*). The values of this dummy are similar to *FS_hh,* with the only difference being that a random[10] percentage of zeroes (5%-25% depending on the level of misclassification error examined) in the latter (*FS_hh*) are transformed into ones in the former (*FS_hh_new*). Hence, we consider 5%-25% of non-participation reports to be false-statements.

It should be mentioned that during this database generation process, a Hotelling's test was performed in each of the *m* databases to make sure that the averages of the covariates in the treated group and the weighted averages of the same covariates in the control group were not significantly different.[11] The procedure was carried out in an iterative way and if the test revealed that the two groups (participants and eligible nonparticipants) could be distinguished on at least one of the covariates at the 10% confidence level, the database was discarded and the procedure was continued until *m* databases were created. This way, the danger of violating the balancing property (i.e. having the treated and control units have the same distribution of observable covariates) is mitigated. This precaution however proved unnecessary since no single database was discarded for this reason, which shows that the balancing property was not at all threatened by the misclassification errors. After this, we then proceeded with the estimation of the ATT_{PSM}^{mis} (i.e., the average of the ATT*s* of each database). Since our previous step led to the generation of *m* different participation dummies, for each level of misclassification errors we get *m ATTs*. Although it is relatively easy to obtain a point estimate by averaging over all point estimates, the calculation of the standard errors is less straightforward. Using Rubin's (1987a) combination of repeated complete-data variances, we derive the standard errors as:

[10] Determined by a pseudo-random number generator of Stata/SE 11.
[11] We use weighted averages for the control group since we employ a 4 to 1 nearest neighbor matching. The weights are the common normalized weights that were then used in the estimation of the 4 to 1 nearest neighbor matching.

$$se_{ATT_{PSM}^{mis}} = \sqrt{\bar{V} + \left(1 + \frac{1}{m}\right) B_m} \qquad (8)$$

where $\left(1 + \frac{1}{m}\right)$ is the correction factor (correcting for the fact that m is finite), $\bar{V}$ is the average of the variances associated with each of the *m* estimated *ATTs* :

$$\bar{V} = \frac{1}{m} \sum_{k=1}^{m} V_k \qquad (9)$$

and B_m is the variance among the *m* estimated *ATTs* :

$$B_m = \frac{1}{m-1} \sum_{k=1}^{m} (ATT_k - \overline{ATT})^2 \qquad (10)$$

For a large number of replications the statistic $(\hat{ATT} - \overline{ATT}) / se_{ATT_{PSM}^{mis}}$ is approximately normal.

2.3 Confounders

Another possible pitfall of the methodology is the bias that could arise from the subtraction of (7) from (4), i.e., the possible bias of the estimated ATT_{PSM} estimator in the case of a failure of the CIA:

$$BIAS = E\left[Y(0) \mid T = 1, P(C)\right] - E\left[Y(0) \mid T = 0, P(C)\right] \qquad (11)$$

This bias is minimized when $P(C) = 0.5$ (Black and Smith, 2004; Heckman and Navarro-Lozano, 2004). Hence, Black and Smith (2004) suggested to estimate the ATT_{PSM} , within a 'thick support' region of the propensity score (i.e., $0.33 < P(C) < 0.66$). However, if treatment effect varies with C , the estimated parameter would deviate from the corresponding population parameter and thus might not be very informative. We follow a different strategy instead: we assume that CIA

holds and scrutinize our results by simulating several binary 'confounders' U.[12] In case of CIA failure, such confounders once added would impose CIA to the model and consequently would transform (7) into:

$$ATT_{PSM}^{conf} = E\left[Y(1)|T=1, P(C,U)\right] - E\left[Y(1)|T=0, P(C,U)\right] \qquad (12)$$

We are particularly interested in assessing how the baseline estimates (7) would change with the addition of possible confounders *U*, in order to perform a robustness check of our estimates. If our findings suggest that ATT_{PSM} estimates are robust to such confounders *U*, then one can be more confident of the interpretation of the results. According to Ichino et al. (2008), it is preferable to avoid parametric assumptions about the simulated confounder. Different hypotheses about the distribution of the confounding factor could be tested by imposing the values of the parameters characterizing the distribution of U ($Pr_{ij} = \Pr(U = 1|T = i, y = j, C)$) $\forall\, i, j \in \{0,1\}$) then predicting a value for each subject according to these parameters, and finally estimating ATT_{PSM}^{conf} and repeating the same process *n* times for the same distribution parameters.[13] The formulas for the calculation of the standard errors of the ATT_{PSM}^{conf} are also those shown in (9)-(11) by replacing the subscript *m* with *n*.

In each of the *n* iterations, two logit models (two odds ratios) are fitted (calculated) to assess the plausibility of the existence of such a confounder. The first ($\Pr(Y = 1|T = 0, \mathrm{U}, C)$) is estimated to show the effect that such a confounder would have on the odds of being obese in the case of no treatment (outcome effect), while the second ($\Pr(T = 1|\,\mathrm{U}, C)$) is employed to highlight the relative importance of the hypothesized confounder on the participation probability (selection effect). The

[12] If *UBEs* are robust in the presence of binary confounders, then this result holds even if the true ones are continuous (see Ichino et al. (2006) for a proof via Monte Carlo simulations).
[13] We used n=1000 but keep this notation for its demonstration simplicity.

results of these two odd ratios are referred to as α and ε respectively. The mean ATT_{PSM}^{conf} are the Confounded Baseline Estimates (CBE*s*) that should also be compared with the UBEs.

As a first simulation practice, we simulate a neutral confounder (i.e., a confounder which has exactly a 50% chance to be 1 in all possible treatment/outcome combinations) and the confounders that mimic the distribution of some of the demographic variables (*Male, Chronic, Educ*$_1$*, Educ*$_2$*, Educ*$_3$*, MarStat*$_1$*, MarStat*$_2$). The results are shown in Tables 5-7. Since these results are highly dependent upon the selection of the covariates, we then search for the existence of a set of parameters pr_{ij} that could drive ATT to zero. As shown in Ichino et al. (2006), such 'dangerous' confounders can be simulated by fixing the probability Pr(U = 1) and the difference $pr_{11} - pr_{10}$ at some predetermined values[14] and then assigning positive values[15] to $d = p_{01} - p_{00}$ and $s = pr_{1.} - pr_{0.}$ where $pr_{i.} = \Pr(U = 1|T = i, C)$. We are then able to assess the plausibility of this particular configuration of parameters through α and ε. If only highly implausible confounders are driving ATT_{PSM}^{conf} to zero, our findings are considered robust to functional form assumptions.

2.4 Misclassification errors and Confounders

Up to this point, we have managed to test the robustness of our UBE*s* by assuming misclassification errors and confounding variables separately. However, we do not know how the results would look like if these two deviations coexisted in our

[14] We have set the value of Pr(U = 1) to 0.3 and that of $p_{11} - p_{10}$ to 0. Since these quantities are not expected to represent a real threat to the baseline estimate, the results remain qualitatively intact when considering different values.

[15] To do so, we used the Matlab code available on the website http://www.tommasonannicini.eu , which returns all the pr_{ij} parameters that simulate U with *d* and *s* varying from 0.1 to 0.6, given the fixed parameters.

settings. Hence, we need a combination of the two procedures described above to further test the validity of our results. In this paper, the combining rule is a nested imputation approach as described in Shen (2000) and employed in D. B. Rubin (2003) and Harel (2007). According to this procedure, in each of the *m* databases created for each level of misclassification errors (as previously demonstrated under the subtitle *"Misclassification errors"*), we construct *n* confounders for each set of parameters pr_{ij} as described above in the calculation of ATT_{PSM}^{conf} . As a result we end up with *m*×*n* (i.e. 1 million) *ATTs* for each level of misclassification errors and for each confounder examined, the average of which provides the Misclassified Confounded Estimates (MCE*s*). The calculation of the standard errors of the $ATT_{PSM}^{mis,conf}$ *s* is now more tedious since the variability comes from multiple sources and is represented as:

$$se_{ATT_{PSM}^{mis,conf}} = \sqrt{\bar{V} + \left(1 - \frac{1}{n}\right) B_n + (1 + \frac{1}{m}) B_m} \qquad (13)$$

where $(1 - \frac{1}{n})$ and $(1 + \frac{1}{m})$ are the correction factors (correcting for the fact that *m* and *n* are finite), $\bar{V}$ is the average of the variances associated with each of the *m*×*n* estimated *ATTs* :

$$\bar{V} = \frac{1}{m \times n} \sum_{k=1}^{m} \sum_{j=1}^{n} V_{kj} \quad , \qquad (14)$$

B_m is the between database variance:

$$B_m = \frac{1}{m-1} \sum_{k=1}^{m} (ATT_k - \overline{ATT})^2 \qquad (15)$$

and B_n is the average between imputation variance:

$$B_n = \frac{1}{m} \sum_{k=1}^{m} \frac{1}{n-1} \sum_{j=1}^{n} (ATT_{kj} - \overline{ATT_k})^2 \qquad (16)$$

For a large number of replications the statistic $(\hat{ATT} - \overline{ATT}) / se_{ATT_{PSM}^{mis,conf}}$, approximates normal.

To sum up, the methodology we employ in this paper consists of four steps. First, we estimate ATT_{PSM} using nearest neighbour propensity score matching, using the four nearest neighbours while assuming that CIA holds and that $\Pr\left(T_i = 1 \mid T_{i,obs} = 0\right) = 0$. The results of this procedure are the Unconfounded Baseline Estimates (UBE*s*). Second, we test the robustness of the UBE*s*, by simulating different misclassification scenarios; these are the Misclassified Baseline Estimates (MBEs) that are compared with the UBEs to check their robustness against different levels of misreported participation. In the next stage, we derive the CBEs and MCEs by augmenting the functional form used to estimate the UBEs and MBEs separately with a neutral confounder (i.e., a confounder which has exactly 50% chance to be 1 in all possible treatment/outcome combinations) and with confounders which mimic the distribution of known demographic variables (*Male, Chronic,* $Educ_1$*,* $Educ_2$*,* $Educ_3$*,* $MarStat_1$*,* $MarStat_2$). In the final step, we generate confounders that could drive the CBEs and MCEs to zero and test the plausibility of their existence through α and ε. This final step is non-trivial and took an enormous amount of computer time to run.

3. Data and Results

Researchers face additional problems when they have to deal with data on FSP participation and weight outcomes. The first is that many US national surveys collect self-reported data for weight and height which can render biased BMI values (Hill and Roberts, 1998; Roberts, 1995). We circumvent this problem by using the 2005-06 National Health and Nutrition Examination Survey (NHANES) which measures the weight and height of individuals, thus reducing intentional and unintentional deviations from the true values (i.e., measurement errors).

The 2005-2006 National Health and Nutrition Examination Survey (NHANES) is designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations and includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. The dataset includes 10,348 respondents in its fullest module. Observations for individuals younger than 18 years old were dropped from all subsequent analysis. Thus, 4267 observations were used in our estimations, of which 474 self-reported to be food stamp recipients.

Results from the probit model used to estimate the propensity scores are given in table 2, which also contains the percentage reduction of bias[16] due to the matching procedure as well as the probability of a type I error if we reject the null hypothesis of no remaining bias after the matching. There are only few remarks to be made on these results. First, we need to mention that we have not excluded the statistically insignificant covariates in the construction of the propensity scores since our aim is to get the most accurate estimation of that score and not of the model. Second, since no figure in the last column is smaller than 0.10, we can accept the hypothesis of no remaining bias for all covariates at the 10% confidence level. Finally, looking at the third column, we notice that there are only two cases (Age^2*Alcohol, $Hsize_2$) where the matching procedure increased the bias between the treated and the control group, but not severely since we still accept the null of no remaining bias.

At a first glance in Table 3, one can notice that the UBE*s* show that FSP participation increases the likelihood of being obese by 10.5% for individuals who

[16] We refer to bias as the standardized percentage difference in covariates means (Rosenbaum & Rubin, 1985).

assign themselves in the program. As we move on to Table 4 where the MBE*s* are presented for the five levels of misclassification errors[17], we can see that when 5% and 10% misclassification errors are assumed, the effect of the FSP on the likelihood of being obese of the participants slightly decreases but still remains statistically significant at the 10% level. However, if 15% or more of the participants have made a false-statement about their participation status, the ATT_{PSM}^{mis} suggests that there are no significant FSP effects on likelihood of being obese. Hence, the positive ATT is only robust to misclassification errors of 10% or less.

In Tables 5-7, the CBE*s* and MCE*s* for potential confounders that mimic the distribution of known covariates for 5% and 10% misclassification errors are presented. We only test the results on these two levels of misclassification errors since in our previous step, we found that for higher misclassification levels, the ATT_{PSM}^{mis} is insignificant even if the assumed functional form is correct. The CBE*s* show that when there are no misclassification errors, the results are robust to the existence of possible confounders that mimic the distribution of all selected covariates. From the values of α and ε, we conclude that such confounders are quite plausible. This is also true for the MCE*s* under 5% misclassification errors. However, when 10% misclassification errors are assumed, the $ATT_{PSM}^{mis,conf}$ becomes more sensitive to the existence of unmeasured variables, since for some confounders (those that mimic the distribution of *Male,* $Educ_2$, $MarStat_2$) it is statistically insignificant at the 10% significance level.

In Tables 8-10, the CBE*s* and MCE*s* are also presented but this time with confounders that are designed to carry all the properties of a 'dangerous' confounder.

[17] In the first line we also include the UBE*s* to facilitate comparisons. As a matter of fact, UBEs can be considered a special case of MBE where the level of misclassification errors is 0%.

As we move down each row in these tables, the selection effect (ε) is held constant and the outcome effect (α) of the hypothesized unmeasured variable increases whilst the exact opposite is true when moving along each line. Moving down the first column of Table 8, we find that for 0% misclassification errors the result of a positive effect of the FSP on the likelihood of being obese of the participants is very robust to unobservable confounders. If a variable that we could not capture was present with a selection effect of 1.58 to 1.63[18], it should also have an outcome effect of more than 7.33[19] to drive the $ATT_{PSM}^{mis,conf}$ to being statistically indistinguishable from zero. However, for unobservable confounders with higher selection effects, the non-significance of the point estimates indicates that the positive effect of the FSP could prove to be an artifact of the omitted variable(s). The same pattern is also observed for the 5% misclassification errors (Table 9) but in this case the estimator appears to be also more sensitive to the outcome effect of the possible confounders (although confounders with such high outcome effects are highly implausible). Finally, for 10% misclassification errors (Table 10), the MCE*s* indicate that $ATT_{PSM}^{mis,conf}$ is very sensitive to additional confounders since it becomes statistically insignificant even for small outcome/selection effects. Overall, the positive effect of FSP participation on obesity for participants is very (quite) robust to the existence of plausible confounders for 0% (5%) misclassification errors and therefore we can claim that the presence of an effect is not an artifact of our assumptions. Nevertheless, in the case of the coexistence of a 10% misclassification error with an unobservable confounder, the MCE*s* indicate a high probability of our results falling flat.

[18] Which in turn means that individuals for whom this confounder is equal to 1 are 58%-63% more likely to participate in the FSP that the others.

[19] Meaning that individuals for whom this confounder is equal to 1 are 633% more likely to be obese than others.

4. Concluding Remarks

The Food Stamp Program is one of the few nutritional assistance programs in the history of the US that has drawn so much attention. Due to high prevalence of obesity among low-income individuals, a number of papers have examined the effect of FSP participation on obesity. Most of these studies have suggested a positive effect. However, none of these studies has evaluated the potential effect of misclassification errors (i.e., misreporting of actual participation status) in the analysis. We feel that this is a very important issue since results could be sensitive to these misclassification errors up to a point where findings can be considered no longer valid. In this study, we examined the complex interrelationship of FSP participation and the likelihood of being obese of participants using propensity score matching and data from the 2005-2006 NHANES. We then tested the robustness of our results under different misclassification errors in the treatment variable as well as the extent of the presence of additional confounders that would be needed for the Conditional Independence Assumption to hold.

Our results suggest that participation in FSP is linked to a 10.5% higher likelihood of being obese for adult participants. This result is consistent with a number of previous studies previously discussed (e.g., Baum, 2007; Chen et al., 2005; Gibson, 2003; Ver Ploeg et al., 2007). This result is also robust to functional form assumptions but only when misclassification errors are 10% or less. Hence, if the predictions of Bollinger & David (1997) and Bitler et al. (2003) are accurate that about 10% to 15% of the participants are misreporting their FSP participation status, then one should be more cautious to conclude about the positive effect of FSP participation on obesity. Specifically, our results indicate that if the level of misclassification error is above 10%, the *ATT* becomes extremely sensitive to

plausible confounders. This issue is important since it can even be possible that misclassification errors are significantly greater than 15% according to Meyer et al. (2010). With misclassification errors of 15% or more, our results reveal no statistically significant effect even if the assumed functional form is the correct one.

Our findings have significant implications for future analyses of FSP participation effects since we provide credible evidence that questions the positive correlation between FSP and obesity suggested in previous studies that failed to address misreporting of participation status and functional form assumptions. Based on our findings, failure to account for these potential sources of biases can render results inaccurate and misinform policy makers. Similar to the majority of previous papers, a weakness of our study is the lack of information in our data about the duration of participation in the program. Nevertheless, an implication of our findings is that misreporting of self-reported participation information should also be taken into account when analyzing the effect of duration of FSP participation on health related outcomes. This would not be an issue with revealed or measured participation data but researchers tend to currently have limited access to these data.

Table 1. Names and descriptions of the variables

Variables	Description
Obese	Dummy, respondent's BMI≥30 kg/m^2 & WC≥100 cm
FS_hh	Dummy, household received food stamps last year
Age	Age of respondent
Alcohol	Average glasses (250 ml) of alcohol consumed by respondent the last 2 days
Chronic	Dummy, Respondent suffers from coronary heart disease, heart attack, stroke or liver condition
DocDiab	Dummy, Respondent has been diagnosed for diabetes/prodiabetes or at risk of diabetes
$Educ_1$	Dummy, up to 9th grade
$Educ_2$	Dummy, 9th-11th grade/High school grad/GED or equivalent
$Educ_3$	Dummy, Some College or Associate of Arts degree
$Educ_4$*	Dummy, College graduate or above
WIC_hh	Dummy, household received Women, Infants and Children benefits last year
$Hsize_1$	Dummy, Household size<2
$Hsize_2$	Dummy, 2 ≤ Household size<5
$Hsize_3$	Dummy, 5 ≤ Household size<7
$Hsize_4$*	Dummy, Household size ≥ 7
Inc_1	Dummy, Annual household income < $24,999
Inc_2	Dummy, $25,000<Annual household Income<$54,999
Inc_3*	Dummy, Annual household income > $55,999
Male	Dummy, Respondent male
$MarStat_1$	Dummy, Respondent married
$MarStat_2$	Dummy, Respondent divorced/separated/widowed
$MarStat_3$*	Dummy, Respondent unmarried
Pregnant	Dummy, Respondent was pregnant at examination
$Race_1$	Dummy, Hispanic race
$Race_2$	Dummy, Ethnicity is non-Hispanic White Race
$Race_3$	Dummy, Ethnicity is non-Hispanic Black Race
$Race_4$*	Dummy, Other ethnicity
Smoker	Dummy, Respondent smokes
PSU_{1-2}	Dummies representing the pseudo-primary units
$Strata_{1-15}$	Dummies representing the pseudo-strata

* These variables were dropped from estimations to avoid perfect multicollinearity

Table 2. Results of the propensity score (Probit) estimation

Variables	Coefficient	% reduction of bias	Prob.
Constant	-4.383***		
Age	0.169***	93.6	0.745
Age^2	0.003***	96.7	0.836
Age^3	0.000**	98.3	0.907
Alcohol	0.869	92.7	0.945
$Alcohol^2$	-1.061	82.8	0.912
$Alcohol^3$	0.206	61.1	0.857
Age Alcohol*	-0.004	72.5	0.847
Age^2* *Alcohol*	0.000	-39.5	0.726
*Age** $Alcohol^2$	0.007	66.8	0.841
Chronic	0.182	86.2	0.839
DocDiab	0.190*	84.9	0.880
$Educ_1$*	0.376**	94.4	0.864
$Educ_2$	0.574***	87.8	0.626
$Educ_3$	0.177	56.4	0.947
WIC_hh	0.708***	91.4	0.632
$Hsize_1$	-0.259	68.5	0.502
$Hsize_2$	-0.150*	-154	0.170
$Hsize_3$	-0.012	58.7	0.400
Inc_1	1.112**	51.9	0.393
Inc_2	0.687	23.1	0.404
Male	-0.044	80.3	0.748
$MarStat_1$	-0.197	60.9	0.216
$MarStat_2$	0.337**	65.3	0.681
Pregnant	0.079	12.2	0.198
$Race_1$	0.012*	70.5	0.430
$Race_2$	0.304	87.6	0.774
$Race_3$	0.935***	75.2	0.221
Smoker	0.229**	87.4	0.745
PSU_{1-2}[1]			
$Strata_{1-15}$[1]			

*,**,*** statistically significant at the 10%, 5% and 1% level respectively

[1] Results for these variables are omitted to save space but are available upon request.

Table 3. Unconfounded Baseline Estimates (UBEs)

ATT	*SE*	*p-value*	*OFF SUPPORT*	*TREATED*	*CONTROL*	*TREATED OFF SUPPORT*	*CONTROL OFF SUPPORT*
			Number of observations				
0.105	0.040	0.01	10	335	823	10	0

Table 4. Misclassified Baseline Estimates (MBEs)

% MISCLAS	*ATT*	*SE*	**Off support**	**MIN off support**	**MAX off support**	*Treated*	*Control*	*Treated off support*	*Control off support*	**MAX treated off support**	**MAX control off support**
						Number of observations					
0	0.105*	0.040	10	-	-	335	823	10	0	-	-
5	0.084*	0.043	10	2	27	376	781	10	0	27	0
10	0.075**	0.043	10	2	29	417	740	10	0	29	0
15	0.068	0.043	11	1	40	458	700	11	0	40	0
20	0.062	0.043	11	0	42	499	659	11	0	42	0
25	0.057	0.042	12	0	42	540	618	12	0	42	0

** (*) Statistically significant at the 5% (10%) level.

Table 5. Confounded Baseline Estimates (CBEs)

CONFOUNDER	α	ε	*ATT (SE)*	pr_{11}	pr_{10}	pr_{01}	pr_{00}	**Off support**	**MIN off support**	**MAX off support**	***Treated off support***	***Control off support***	**MAX treated off support**	**MAX control off support**
								Number of observations						
Neutral	1.00	1.00	0.101** (0.042)	0.5	0.5	0.5	0.5	10	6	17	10	0	10	0
Confounder like...														
Male	0.67	0.78	0.106** (0.043)	0.28	0.45	0.37	0.47	17	8	29	17	0	17	0
Chronic	1.87	0.78	0.123** (0.041)	0.09	0.09	0.16	0.09	17	4	30	17	0	17	0
$Educ_1$	0.97	0.56	0.109** (0.042)	0.12	0.15	0.21	0.22	17	7	31	17	0	17	0
$Educ_2$	1.17	2.18	0.097** (0.044)	0.31	0.4	0.24	0.21	14	3	34	14	0	14	0
$Educ_3$	1.26	0.98	0.120** (0.042)	0.27	0.26	0.3	0.98	17	7	25	17	0	17	0
$MarStat_1$	1.06	0.59	0.106** (0.043)	0.3	0.26	0.4	0.38	17	7	30	17	0	17	0
$MarStat_2$	2.07	1.27	0.110** (0.043)	0.31	0.21	0.3	0.18	17	4	28	17	0	17	0

[1] These values are rounded averages over the 1,000 estimations

** (*) Statistically significant at the 5% (10%) level.

Table 6. Misclassified Confounded Estimates (MCEs) for 5% misclassification errors

CONFOUNDER	***α***	***ε***	***ATT (SE)***	pr_{11}[1]	pr_{10}[1]	pr_{01}[1]	pr_{00}[1]	**Off support**[2]	**MIN off support**	**MAX off support**	***Treated off support***[2]	***Control off support***[2]	**MAX treated off support**	**MAX control off support**
								Number of observations						
Neutral	1.00	1.00	0.084** (0.042)	0.5	0.5	0.5	0.5	10	1	38	10	0	38	0
Confounder like...														
Male	0.66	0.80	0.079* (0.043)	0.28	0.46	0.38	0.48	10	1	40	10	0	40	0
Chronic	1.95	0.80	0.087** (0.043)	0.1	0.09	0.16	0.09	10	1	38	10	0	38	0
$Educ_1$	0.98	0.60	0.083** (0.044)	0.12	0.16	0.21	0.22	10	0	40	10	0	40	0
$Educ_2$	1.19	2.04	0.078* (0.044)	0.3	0.38	0.24	0.21	11	0	49	11	0	49	0
$Educ_3$	1.27	0.99	0.084** (0.043)	0.28	0.26	0.3	0.26	10	1	34	10	0	34	0
$MarStat_1$	1.07	0.63	0.084* (0.044)	0.31	0.27	0.40	0.39	10	1	43	10	0	43	0
$MarStat_2$	2.12	1.24	0.079* (0.043)	0.31	0.2	0.3	0.18	10	1	41	10	0	41	0

[1] These are average percentages over all simulations since the value of the distribution parameters of the demographic variables on the treatment/outcome condition were different in each of the 1,000 simulated databases.

[2] These values are rounded averages over the 1,000,000 estimations.

** (*) Statistically significant at the 5% (10%) level.

Table 7. Misclassified Confounded Estimates (MCEs) for 10% misclassification errors

CONFOUNDER	***α***	***ε***	***ATT (SE)***	pr_{11}[1]	pr_{10}[1]	pr_{01}[1]	pr_{00}[1]	**Off support**[2]	**MIN off support**	**MAX off support**	***Treated off support***[2]	***Control off support***[2]	**MAX treated off support**	**MAX control off support**
								Number of observations						
Neutral	1.00	1.00	0.074* (0.043)	0.50	0.50	0.50	0.50	10	1	36	10	0	36	0
Confounder like...														
Male	0.66	0.82	0.070 (0.043)	0.30	0.46	0.37	0.48	10	0	43	10	0	43	0
Chronic	0.09	1.94	0.076* (0.043)	0.10	0.09	0.16	0.09	10	0	40	10	0	40	0
$Educ_1$	0.98	0.64	0.074* (0.043)	0.13	0.16	0.21	0.22	10	0	46	10	0	46	0
$Educ_2$	1.19	1.91	0.069 (0.044)	0.30	0.36	0.24	0.21	10	0	51	10	0	51	0
$Educ_3$	1.27	0.83	0.072* (0.044)	0.29	0.23	0.30	0.28	10	1	35	10	0	35	0
$MarStat_1$	1.07	0.66	0.073* (0.043)	0.31	0.29	0.40	0.39	10	0	56	10	0	56	0
$MarStat_2$	2.13	1.22	0.070 (0.043)	0.31	0.2	0.3	0.18	10	0	43	10	0	43	0

[1] These are average percentages over all simulations since the value of the distribution parameters of the demographic variables on the treatment/outcome condition were different in each of the 1,000 simulated databases.

[2] These values are rounded averages over the 1,000,000 estimations.

* Statistically significant at the 10% level.

Table 8. Confounded Baseline Estimates (CBEs)[1]

	s=0.1 ***ε= [1.58,1.63]***	***s=0.2*** ***ε= [2.55,2.62]***	***s=0.3*** ***ε= [4.12,4.22]***	***s=0.4*** ***ε= [6.81,6.98]***
d=0.1 ***α =[1.73,1.99]***	0.087* (0.045)	0.070 (0 .047)	0.057 (0 .051)	0.040 (0.059)
d=0.2 ***α =[2.86,3.8]***	0.079* (0 .045)	0.053 (0 .048)	0.026 (0.052)	-0.006 (0.059)
d=0.3 ***α= [4.76,7.33]***	0.072* (0 .045)	0.036 (0.049)	-0.003 (0.053)	-0.052 (0.060)
d=0.4 ***α= [8.20,15.58]***	0.065 (0.047)	0.018 (0.05)	-0.033 (0.054)	-0.10 (0.060)

[1] For each of these 16 models, similar results such as those in Tables 5-7 are available upon request

* Statistically significant at the 10% level

Table 9. Misclassified Confounded Estimates (MCEs) for 5% misclassification errors[1]

	s=0.1 ***ε= [1.57,1.63]***	***s=0.2*** ***ε= [2.53,2.63]***	***s=0.3*** ***ε= [4.14,4.29]***	***s=0.4*** ***ε= [6.96,7.21]***
d=0.1 ***α =[1.73,2.10]***	0.074* (0.044)	0.060 (0.046)	0.046 (0.051)	0.026 (0.058)
d=0.2 ***α =[2.88,4.12]***	0.066 (0.045)	0.042 (0.047)	0.013 (0.051)	-0.025 (0.059)
d=0.3 ***α= [4.82,8.35]***	0.058 (0.045)	0.023 (0.048)	-0.019 (0.052)	-0.076 (0.059)
d=0.4 ***α= [8.37,19.13]***	0.051 (0.046)	0.006 (0.049)	-0.050 (0.053)	-0.12 (0.060)

[1] For each of these 16 models, similar results such as those in Tables 5-7 are available upon request.

* Statistically significant at the 10% level

Table10. Misclassified Confounded Estimates (MCEs) for 10% misclassification errors[1]

	s=0.1 ***ε= [1.61,1.64]***	***s=0.2*** ***ε= [2.49,2.63]***	***s=0.3*** ***ε= [4.12,4.35]***	***s=0.4*** ***ε= [7.12,7.56]***
d=0.1 ***α =[1.74,2.20]***	0.064 (0.043)	0.051 (0 .046)	0.035 (0 .051)	0.013 (0.058)
d=0.2 ***α =[2.93,4.56]***	0.055 (0 .044)	0.031 (0 .047)	0.021 (0.048)	-0.043 (0.058)
d=0.3 ***α= [4.95,10.02]***	0.046 (0 .044)	0.01 (0.048)	-0.036 (0.053)	-0.099 (0.059)
d=0.4 ***α= [10.66,20.63]***	0.038 (0.045)	-0.009 (0.048)	-0.070 (0.052)	-0.158 (0.059)

[1] For each of these 16 models, similar results such as those in Tables 5-7 are available upon request.

References

Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *Stata Journal*, *4*, 290–311.

Abadie, A., & Imbens, G. W. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica*, *74*(1), 235-267.

Angrist, J. & Pischke, J., 2010. The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *National Bureau of Economic Research Working Paper Series*, No. 15794.

Baum, C. (2007). *The Effects of Food Stamps on Obesity* (): The United States Department of Agriculture (USDA), Economic Research Service (ERS).

Bitler, M. P., Currie, J., & Scholz, J. K. (2003). WIC eligibility and participation. . *Journal of Human Resources*, *38*, 1139–1179.\line

Black, D. A., & Smith, J. A. (2004). How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics*, *121*(1-2), 99–124.

Bollinger, C. R., & David, M. H. (1997). Modeling Discrete Choice With Response Error: Food Stamp Participation. *Journal of the American Statistical Association*, *92*(439), 827-835.

Bray, G. A. (2004). Medical Consequences of Obesity. *J Clin Endocrinol Metab*, *89*(6), 2583-2589.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, *163*(12), 1149-1156.

Chen, Z., Yen, S. T., & Eastwood, D. B. (2005). Effects of Food Stamp Participation on Body Weight and Obesity. *American Journal of Agricultural Economics*, *87*(5), 1167-1173.

Cutler, D. M., Glaeser, E. L., & Shapiro, J. M. (2003). Why Have Americans Become More Obese?. *Journal of Economic Perspectives*, *17*(3), 93-118.

Ejerblad, E., Fored, C. M., Lindblad, P., Fryzek, J., McLaughlin, J. K., & Nyren, O. (2006). Obesity and Risk for Chronic Renal Failure. *J Am Soc*

Nephrol, *17*(6), 1695-1702.

Esposito, K., Giugliano, F., Di Palo, C., Giugliano, G., Marfella, R., D'Andrea, F., et al. (2004). Effect of Lifestyle Changes on Erectile Dysfunction in Obese Men: A Randomized Controlled Trial. *JAMA*, *291*(24), 2978-2984.

Gibson, D. (2003). Food Stamp Program Participation is Positively Related to Obesity in Low Income Women. *J. Nutr.*, *133*(7), 2225-2231.

Grundy, S. M. (2004). Obesity, Metabolic Syndrome, and Cardiovascular Disease. *J Clin Endocrinol Metab*, *89*(6), 2595-2600.

Harel, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*, *4*(1), 75 - 89.

Heckman, J., & Navarro-Lozano, S. (2004). Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models. *Review of Economics and Statistics*, *86*(1), 30-57.

Hill, A., & Roberts, J. (1998). Body mass index: a comparison between self-reported and measured height and weight. *J Public Health*, *20*(2), 206-210.

Ichino, A., Mealli, F., & Nannicini, T. (2006). From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? . *CEPR Discussion Paper 5736.*

Ichino, A., Mealli, F., & Nannicini, T. (2008). From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? . *Journal of Applied Econometrics*, *23*(3), 305-327.

Kaushal, N. (2007). Do food stamps cause obesity? Evidence from immigrant experience. *Journal of Health Economics*, *26*(5), 968–991.

Lakdawalla, D., Philipson, T., & Bhattacharya, J. (2005). Welfare-enhancing technological change and the growth of obesity. *American Economic Review*, *95*(2), 253–257.

Meyer, B. D., Goerge, R. M., & Hall, C. (2010). Errors in Survey Reporting and Imputation and their Effects on Estimates of Food Stamp Program Participation.

Millimet, D. L., & Tchernis, R. (2009). On the Specification of Propensity Scores: with Applications to the Analysis of Trade Policies. *Journal of Business & Economic Statistics*, *27*, 397-415.

Philipson, T. J., & Posner, R. A. (2003). The Long-Run Growth in Obesity as a Function of Technological Change. *Perspectives in Biology and Medicine*, *46*(3x), S87-S107.

Roberts, R. (1995). Can self-reported data accurately describe the prevalence of overweight? . *Public Health*, *109*(4), 275-284.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*(1), 33–38.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rosin, O. (2008). The economic causes of obesity: A survey. *Journal of Economic Surveys*, *22*(4), 617-647.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (). New York, NY: John Wiley and Sons.

Rubin, D. B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. . *Statistica Neerlandica*, *57*(1), 3-18.

Rubin, D. B., & Thomas, N. (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics*, *52*(1), 249-264.

Shen, Z. (2000). *Nested multiple imputation*. Unpublished PhD Thesis, Department of Statistics, Harvard University, Cambridge, MA.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. . *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Van der Steeg, J. W., Steures, P., Eijkemans, M. J., Habbema, J. D. F., Hompes, P. G., Burggraaff, J. M., et al. (2007). Obesity affects spontaneous pregnancy chances in subfertile, ovulatory women. . *Hum. Reprod.*, dem371.

Ver Ploeg, M., Mancino, L., Lin, B., & Michele Ver Ploeg, L. M. A. L. (2007). *Food and Nutrition Assistance Programs and Obesity: 1976-2002* (economic research report): USDA.\line

Whitmer, R.A., Gunderson, E., Barrett-Connor, E., Quesenberry C., Yaffe, K. 2005. Obesity in middle age and future risk of dementia: a 27 year longitudinal population based study. BMJ