Institute for Research on Poverty
Discussion Paper no. 1293-05

# A Cautionary Tale:
# Using Propensity Scores To Estimate the Effect of Food Stamps on Food Insecurity

Christina M. Gibson-Davis
Duke University
E-mail: cgibson@duke.edu

E. Michael Foster
The Pennsylvania State University

March 2005

Funding for the first author was provided by the USDA's small grants program, administered through the Institute for Research on Poverty at the University of Wisconsin–Madison. Previous versions of this paper were presented at the Association for Public Policy Analysis and Management annual meeting, November 2004, Atlanta, GE, and the USDA's Economic Research Service Conference, December 2004, Washington, D.C.

IRP Publications (discussion papers, special reports, and the newsletter *Focus*) are available on the Internet. The IRP Web site can be accessed at the following address: http://www.ssc.wisc.edu/irp/

**Abstract**

We use propensity scores to evaluate the effect of Food Stamps on food insecurity, a measure of inadequate food supply. Data come from the Early Childhood Longitudinal Study–Kindergarten Cohort. Propensity scores offer an advantage over traditional linear regression methods because they address omitted variable bias associated with Food Stamp use by matching similarly situated treatment and control group members, and estimating mean differences within these matched groups. We find that the program does not decrease the probability of being food insecure, but it may lessen the severity of the problem. We also note that propensity scores rest on several stringent assumptions and should be employed with caution.

**A Cautionary Tale:**
**Using Propensity Scores To Estimate the Effect of Food Stamps on Food Insecurity**

## INTRODUCTION

Randomized experiments are considered the "gold standard" of evaluation methods. They ensure, assuming correct implementation, that differences in outcomes between the treatment and the control group are caused by the treatment of interest and not by some preexisting characteristic. They therefore avoid most threats to external validity and are generally preferred to other types of evaluation methods (Shadish, Cook, and Campbell, 2001).

In some instances, however, randomized experiments are infeasible, either because randomization is impossible (for example, assigning children to different types of parenting approaches) or due to ethical considerations (denying pregnant women vitamins to evaluate a prenatal nutrition program). When a randomized experiment is unlikely or improbable, researchers must rely on nonexperimental methods to evaluate the outcome of interest.

One case where researchers need to employ nonexperimental methods is in evaluations of the Food Stamp program. Food Stamps, which in 2003 provided assistance to 9.2 million households, including 5 million households with children, is the largest federal food program and the cornerstone of federal food assistance (Cunnyngham and Brown, 2004). Its goal is to ensure that low-income families have sufficient resources to purchase a nutritiously adequate diet. A measure of Food Stamp effectiveness is its impact on food insecurity, a USDA-designed module of 18 items that classifies families as being either food secure (adequate food supply) or food insecure (inadequate food supply) (Bickel et al., 2000). Reducing levels of food insecurity is an important goal, particularly for children, as those who are food insecure are more likely to suffer from a range of academic and behavioral deficits (Kleinman et al., 1998; Murphy et al., 1998; Weinreb et al., 2002; Dunifon and Kowaleski-Jones, 2003).

However, the problem in analyzing the impact of Food Stamps on food insecurity is that unmeasured or unobserved characteristics are likely to be correlated with both Food Stamp use and food

security (Jensen, 2002; Currie, 2003). This introduces bias, which might either understate or overstate the program's impact. For example, most research indicates that those who use Food Stamps are disadvantaged relative to those who do not (Daponte, Sanders, and Taylor, 1999; Cunnyngham and Brown, 2004); this disadvantage may increase the likelihood that these families are also food insecure. If this is the case, then unless analyses of the program are properly adjusted, simple comparisons are likely to understate the program's impact. The bias may also operate in the opposite direction. As discussed below, not all families who are eligible enroll in the program. It may be that those who enroll are better organized or otherwise advantaged over those who qualify but do not enroll. In that case, the program may appear more effective than it actually is.

This problem of omitted variable bias is not confined to the evaluation of Food Stamps but applies more generally in any nonexperimental evaluation of a program. One common method used to address this issue is multivariate regression. Regression, however, suffers from two problems. First, it is often heavily dependent on functional form (for example, whether the relationship between the covariates and the outcome of interest is assumed to be linear). In this example, we assume the effect of Food Stamps to be constant: the program would be just as effective at reducing mild food insecurity as it would be at reducing severe food insecurity. This may not be a plausible assumption, as Food Stamps is designed as a safety net and may be better able to help those who are in more dire need (Currie, 2003).

Functional form is particularly important in instances where the two groups of interest differ substantially in terms of their characteristics (this is known as the "common support problem," common support referring to overlapping distributions of their characteristics). If the two groups are very dissimilar, analogous to comparing apples to oranges, caution should be exercised in making comparisons between the two groups. However, regression hides the common support problem, because it does not quantify the similarities (or dissimilarities) between the two groups.

Fortunately, recent developments in nonexperimental methodology provide some alternative techniques for evaluating a nonrandomized program such as Food Stamps. One such method is propensity

scores (Rosenbaum and Rubin, 1983). Under key assumptions, propensity scores approximate a randomized experiment by creating a "matched" treatment and control group, who are, save for treatment status, comparable. When the groups are compared on an outcome, any resulting differences should reflect the treatment and not preexisting characteristics (Heckman, Ichimura, and Todd, 1997, 1998; Dehejia and Wahba, 1999, 2002). Propensity scores have been used in many contexts, including evaluations of child care, welfare-to-work programs, mental health services, and drop-out prevention programs (Hill, Brooks-Gunn, and Waldfogel, 2002; Foster, 2003; Agodini and Dynarski, 2004; Michalopoulos, Bloom, and Hill, 2004). This area is an active one in statistics, and the methods continue to be refined (Heckman and Navarro-Lozano, 2004; Imbens, 2004).

In this paper, we use propensity scores to examine the effect of the Food Stamp program on food insecurity. Data come from the first and second waves of the Early Childhood Longitudinal Sample–Kindergarten Cohort (ECLS-K), a nationally representative dataset of over 21,000 children. To the best of our knowledge, we are the first to use propensity scores to evaluate Food Stamps, and we advance the literature in two important ways. First, we use statistically rigorous methods to evaluate Food Stamp use among a sample of households with young children. Children make up the largest group of Food Stamp recipients (Cunnyngham and Brown, 2004), and also suffer when households are food insecure (Dunifon and Kowaleski-Jones, 2003; Stormer and Harrison, 2004). Thus, Food Stamps could be an important program in promoting child well-being. Second, we compare the advantages and disadvantages of propensity scores to more traditional linear regression models. We therefore provide an illustration of the use of this method, highlighting how it might be used by other researchers.

We first outline the Food Stamp program and food insecurity, and discuss the connection between the two. We then explain propensity scores, and illustrate the basic concepts behind the method. Next, we describe our data and methods, and then present our results. We conclude by discussing our findings, and describe the advantages and disadvantages to using propensity scores.

# BACKGROUND

## The Food Stamp Program

Food Stamps originated in the Great Depression as an experimental program to help families supplement their food buying power. That version of the program ended in 1943, and the program did not become federal law until nearly twenty years later with the passage of the Food Stamp Act of 1964 (U.S. Department of Agriculture, 2004b). Since then, when $75 million was appropriated to serve 500,000 individuals, the program has grown enormously in size and scope. In 2003 the federal government spent $23.9 billion on Food Stamps, serving 9.2 million households, more than half of which contained children. The number of Food Stamp recipients has increased steadily since 2000, although rates are still far lower than they were in the mid-1990s (Cunnyngham and Brown, 2004).

The Food Stamp program provides a grant to be used for the purchase of any food item, with the exceptions of vitamins, hot meals, alcohol, cigarettes, or non-food-related household items, such as soaps, paper products, or personal hygiene items. The money is provided in the form of a debit card, which can be used in grocery stores or other approved vendors.[1]

To be eligible to receive Food Stamps, households without an elderly or disabled member must have a gross monthly income that is less than 130 percent of the federal poverty line ($18,810 in 2003 for a family of four) and have assets less than $2,000.[2] They are automatically eligible if they receive General Assistance, Supplemental Security Income (SSI), or Temporary Assistance for Needy Families (TANF). If families qualify, then households receive enough funds to be able to purchase the foods that constitute

---

[1]One concern with providing families with a debit card, as opposed to actual food goods, is that they will not increase the amount of money spent on food, but will use the additional funds to purchase non-food-related items. However, research indicates that Food Stamps does increase the amount spent on food; it is estimated a household will spend an additional $0.17 to $0.47 on food for every dollar received in Food Stamps (Fraker, 1990; see also Kramer-LeBlanc, Basiotis, and Kennedy, 1997).

[2]Households with an elderly or disabled member are exempt from the gross income test. There are additional restrictions on other types of individuals, including able-bodied adults without dependents and legal aliens.

the Thrifty Food Plan (TFP), a USDA-designed basket of goods that provides a nutritiously adequate diet (Currie, 2003). The amount is adjusted for household composition and geographic location; in fiscal year 2003, average monthly benefits per household in the continental United States ranged from $146 in Connecticut to $299 in Arizona, with a U.S. average of $195 (U.S. Department of Agriculture, 2004a).

Food Stamps reach some of the poorest households in the United States. Nearly nine out of ten Food Stamp households had gross monthly incomes below the federal poverty line, and almost half had incomes less than 50 percent of the poverty line. More than half of Food Stamp recipients are children, the majority of whom live in a household headed by a single adult (usually a female). As a result, more than one out three Food Stamp households consists of a single adult living with children (Cunnyngham and Brown, 2004).

Research indicates that not all who qualify use Food Stamps. Take-up rates are particularly low among the elderly, with much higher take-up rates among families with children (Currie, 2003). Past research also indicates that those likely to use Food Stamps are relatively disadvantaged, even among lower-income Americans. They are more likely to be young, minority, of lower educational status, unemployed, disabled, single, live in an urban area, have more children and fewer assets (Blank and Ruggles, 1996; Daponte et al., 1999; Cunnyngham and Brown, 2004).

## Food Insecurity

Since 1995, the USDA has collected information on food security in the United States by means of an annual supplement to the Current Population Survey. Defined as “access by all people at all times to enough food for an active, healthy life” (Nord, Andrews, and Carlson, 2003), food security is assessed by the Food Security Survey module, an 18-item questionnaire that measures the adequacy of the

household's resources to purchase the types and quantities of foods that it wants.[3] The module is ordered by degree of severity. The module measures a continuum of food-related behaviors moving from uncertainty about food availability ("*Worried food would run out before I/we got money to buy more*") to changing the types of foods purchased (*"Couldn't afford to eat balanced meals"*) to reducing adult food intake ("*Adult(s) cut size of meals or skipped meals*") to reducing child food intake ("*Children have not eaten for a whole day because their wasn't enough money for food*") (see Appendix Table A.1 for a list of the 18 items). A Rasch scale, a type of factor analysis that recognizes that the indicators are categorical, is then used to classify households into one of three categories: food secure, food insecure without hunger, or food secure with hunger. Households must respond positively to at least three items to be classified as food insecure, and eight items to be considered food insecure with hunger. Being food insecure without hunger signifies that a household has experienced some reduction in food intake due to inadequate resources, and those that are food insecure with hunger have forgone enough food to experience hunger (Bickel et al., 2000).

In 2002, 89 percent of all American households were classified as food secure. Of the 11 percent who were food insecure, 7.6 percent were food insecure without hunger and 3.5 percent were food insecure with hunger. Although not a high percentage, these rates nevertheless indicate that 12.1 million U.S. households were food insecure at some point during the previous year, including 3.5 million households that experienced hunger. Both numbers represent a slight increase from 2001, though rates remain lower than when food security was first measured in 1995. Not surprisingly, the poor are more likely to be food insecure (33.7 percent), and single-parent families and those headed by a racial or ethnic minority member are particularly at risk. Slightly less than half (43 percent) of all children in low-income families are food insecure, including 11 percent that are classified as food insecure with hunger. About

[3]Households without children are asked a 10-item subset, but the scale is constructed such that comparisons can be made between households with and without children.

one-third of households who are food insecure are also receiving Food Stamps (Nord, Andrews, and Carlson, 2003).

The literature examining the effect of food insecurity on child well-being is limited, but generally consistent. Overall, food-insecure children perform less well on measures of psychosocial functioning, and those classified as insecure with hunger are particularly at risk (Murphy et al., 1998; Weinreb et al., 2002). Children living in food-insecure households are more likely to be aggressive, anxious, hyperactive, have poorer social skills, and have more mental health problems (Kleinman et al., 1998; Murphy et al., 1998; Dunifon and Kowaleski-Jones, 2003; Stormer and Harrison, 2004). Food insecurity does not directly affect cognitive functioning (Stormer and Harrison, 2004; Slack and Yoo, 2004) but it may indirectly affect academic well-being by increasing the likelihood that a child will be placed in special education, repeat a grade, or be absent or tardy from class (Kleinman et al., 1998; Murphy et al., 1998). These findings are consistent with the theory that in more developed countries, hunger does not impair physiological development, but it can interfere with a child's socioemotional well-being (Kleinman et al., 1998)

## The Connection between Food Stamps and Food Insecurity

Theoretically, Food Stamp recipients should not be food insecure, as their monthly grant should be sufficient for them to purchase the goods that constitute the Thrifty Food Plan (TFP). However, the only assumption made by the Food Stamp program is that recipients can purchase the TFP; whether they do so is up to the choices and preferences of the individual. For example, recipients may run out of money for food if they do not smooth their food consumption throughout the month. Alternatively, participants could purchase foods not in the TFP, which may be more to their liking but also more expensive. Finally, recipients may be constrained by geography and transportation in their choice of food stores, and may not be able to obtain food at the same prices as those assumed by the TFP.

Thus far, perhaps because the module is relatively new, only a few studies have examined the effect of Food Stamps on food insecurity, and the results are inconsistent across samples and methods.

Cook et al. (2004) found that Food Stamps reduced, but did not eliminate, food insecurity among a sample of urban parents; Oberholser and Tuttle (2004) found no such relationship in their telephone survey of Food Stamp recipients in Maryland. While each study used multivariate regression, neither used additional statistical methods to address the problem of omitted variable bias. Studies that have addressed omitted variable bias, through an instrumental variable approach, have also found mixed results.[4] Borjas (2004) looked at reductions in food insecurity among households receiving either cash assistance, Food Stamps, or Medicaid; he found that a 10 percentage-point decrease in households receiving public assistance would increase the number of food-insecure households by 5 percentage points. Kabbani and Yazbeck (2004) used data from the Current Population Survey (CPS) and found no effect of Food Stamps on food insecurity among households with children ages 5 to 18.

The Use of Propensity Scores

Program evaluation often consists of estimating the counterfactual—that is, what would have occurred had the treatment or program not been in place (Shadish et al. 2001). That is, we would like to know how person C experiences food insecurity with Food Stamps and then compare that to how that same person experiences food insecurity without Food Stamps. The difference between the two states would indicate the program's impact. Of course, we cannot do this—we cannot observe C simultaneously both receiving and not receiving Food Stamps—and therefore a true counterfactual is not possible. We must use other methods to approximate it.

In randomized experiments, the way to construct the counterfactual is to randomly assign cases to either a treatment or a control group. Assuming that randomization is executed successfully and has

[4]Two other studies have examined the effect of Food Stamps on food insufficiency, a one-item measure of food adequacy used by the USDA prior to the introduction of the food insecurity module. Again, though, the results were not consistent, as Food Stamps reduced food insufficiency in one study (Rose, Gundersen, and Oliveira, 1998) but not the other (Gundersen and Oliveira, 2001).

adequate sample size, then cases (with their associated characteristics) are distributed randomly and evenly across the two groups. The randomization process ensures that even though we do not observe C both with and without Food Stamps, we can observe C with Food Stamps and someone very similar to C (say C*) without Food Stamps. This equivalence has the effect of nullifying the effect of any preexisting characteristic on C's level of food insecurity: that same characteristic also exists for C*, and its presence in both groups ensures that it cannot account for the difference between the treatment and control groups on levels of food insecurity.[5] A key feature of randomization is that it balances both observed characteristics as well as unobserved characteristics.

With propensity scores, establishing a counterfactual consists of inferring what would have happened to the treatment group had they not received the treatment. In other words, one wants to make the treatment group as similar to the control group as possible; if the two groups are similar, then there should be no intergroup variation in outcomes. The goal is therefore to form two groups that are, on average, equivalent except that one receives the treatment and other does not. Then, any mean differences between the two will be due to the receipt of the treatment and not to some preexisting characteristic (Rosenbaum and Rubin, 1983, 1985; Dehejia and Wahba, 1999, 2002).

To create these comparable groups, one needs to match on those characteristics that differentiate the treatment group from the control group. This is the function of the propensity score. It represents the predicted probability of participating in the treatment, based on the observed and measured characteristics used in the prediction equation. Each member of the sample receives a propensity score, which will range between zero and one, representing either a low or high likelihood of receiving the treatment. Once the propensity score has been calculated, a treatment-group member can be paired to a control-group member who has a similar propensity score. By matching, propensity scores capture all the observed and measured

---

[5]This is an overly simplistic example, as random assignment does not rely on one treatment group member being equivalent to one control group member. Rather, it assumes that, on average, the two groups are equal.

differences between the treatment and control groups. They further avoid the computationally burdensome task of trying to match cases across multiple dimensions, because cases are matched only on a single variable (Rosenbaum and Rubin, 1983). It is important to note that matching is done only on observed characteristics; unlike randomization, propensity scores cannot match on unobserved characteristics.

Instead, propensity scores address the bias from unobserved characteristics by assuming that all factors related to selecting into a treatment (for example, all those reasons why people choose to use Food Stamps) have been observed and measured. If these factors have been taken into account, then conditional on these factors, an outcome is said to be "strictly ignorable" (Rosenbaum and Rubin, 1983) of treatment status—that is, an outcome is not influenced by those same factors that also influenced someone to take up a treatment. This assumption, which is variously known as "selection on observables" (Heckman and Robb, 1985), the "unconfoundness assumption" (Imbens, 2004), or the "conditional independence assumption" (Black and Smith, 2004), is critical to the validity of propensity scores. If this assumption is violated—if there are unobserved characteristics that influence both the use of Food Stamps and food insecurity—then the treatment and control groups may differ in unobserved ways, and between-group differences may reflect those characteristics rather than the treatment (Bryson, Dorsett, and Purdon, 2002).

Another important assumption of propensity scores is that the effect of treatment can be measured only for those individuals who have a matching case in the other group. This area of overlap in propensity scores is known as the "common support" (Imbens, 2004; Smith and Todd, in press). This assumption means, essentially, that an appropriate counterfactual can be found for each case. For cases with extreme propensity scores (very close to zero or one), the opposite response (among matched individuals) to the treatment may not be observed. For example, a family participating in Food Stamps may have such a high propensity to participate that there is no comparable family with the same propensity score who chooses not to participate. In this case, a counterfactual cannot be established (there is no C* for C), and the effect

of treatment cannot be measured. This violates the common support assumption, and when cases lie outside the region of common support, they are typically dropped from the analyses (Bryson et al., 2002).[6]

Calculating the Propensity Score

In determining what covariates are used in the propensity score model, theoretically one should include any characteristic that is related to both the treatment and the outcome. This might encourage one to err on the safe side—that is, to include anything that might potentially be confounded with the treatment effect. Indeed, Heckman and colleagues have shown that models with more covariates tend to be less biased than those that are smaller (Heckman, Ichimura, and Todd, 1997, 1998).

Yet having a propensity score model that includes a wide range of covariates has its disadvantages. First, although including extraneous variables does not influence the bias of the matching estimates, it does introduce more variance (Bryson et al., 2002). Second, and perhaps most important, including more covariates makes defining the common support region much more difficult (Smith and Todd, in press). That is, the better the model is at predicting participation, the more likely the propensity scores are "correct"—that is, for the treatment group, propensity scores will be close to one. For the control group, they will be close to zero. In that case, the area of overlap—the common support—can be quite small.

Unfortunately, there is little guidance as to how to strike a balance behind these competing tensions (Smith and Todd, in press). Construction of the propensity score thus involves a trade-off between minimizing the bias through the inclusion of many covariates, yet risking violating the common

[6]One other assumption underlying propensity scores is that the stable unit treatment value assumption (SUTVA) has not been violated. The SUTVA stipulates that the treatment effect for any given individual does not depend on the size of the treatment, or its effects on other members (Sianesi, 2001). We do not test this assumption in this paper, but there is little a priori reason it would be a significant factor in Food Stamp receipt.

support region because the two groups are so dissimilar (Black and Smith, 2004; Smith and Todd, in press).

Matching Comparable Cases

Numerous methods exist for matching on the propensity scores (see Smith and Todd, in press, for a discussion of matching estimators); most work has shown that for samples of sufficient size, the alternative methods produce similar results. However, a brief explanation of two types of methods can illustrate how they make trade-offs between bias versus efficiency. The most straightforward matching method is called "single nearest neighbor," which, just as the name implies, matches a treatment case with the nearest control case. This method uses a one-to-one matching metric, meaning that control cases not sufficiently close to treatment cases are discarded. Discarding these cases minimizes bias, because only those cases that most closely resemble each other are used in matching. At the same time, however, this practice may increase the variance of the estimates, as reliance on a small pool of matches introduces more error than if a larger pool (even if it included poorer matches) had been used (Smith and Todd, in press).

Instead of using only a single control case, other methods use multiple comparison cases. Kernel matching, for example, uses several comparison group members, pairing a treatment case with the weighted average score of all control cases within a certain distance (kernel is the name of the weighting function, and the distance is determined by the bandwidth of the kernel). A variant on kernel matching is stratification matching. Stratification matching also pairs treatment cases with a weighted average of control cases, except that the matching distance is defined by the user rather than by a kernel function.[7] One of the key differences between kernel and stratification methods is that in kernel matching, control

[7]Specifically, the matching distance is defined by dividing the common support region into N strata, with the number of strata determined such that there are no significant differences on covariates within a stratum.

cases that are closer to the treatment case are given more importance (via weights) than those that are father away. For both of these methods, using all of the control cases reduces the variance of the matching, as one has more information on which to match. However, it also increases the bias by using poorer quality matches (Smith and Todd, in press).

In sum, several different methods for matching exist, but they should asymptotically produce the same estimates. Given smaller sample sizes, the point estimates and their standard errors may differ. By looking at the assumptions behind the methods, one can hypothesize as to how the estimates may vary. Nearest neighbor should produce the least biased estimates, but also the largest standard error, whereas kernel and stratification matching may be more biased, but they will have smaller standard errors. It is unknown which would produce the largest mean square error in any given application.

## DATA

### The ECLS-K

Data come from the Early Childhood Longitudinal Sample–Kindergarten Cohort (ECLS-K), a nationally represented dataset collected by the U.S. Department of Education. This dataset is both large (over 21,000) and longitudinal (children are followed from kindergarten through third grade). The goal of the survey is to collect sufficient data so that the factors affecting a young child's health, schooling, and cognitive and behavioral outcomes can be fully understood. The ELCS-K collects data on numerous child, parent, and community domains, and surveys the mother, father, and the child's teacher. The data used in this paper come from the first wave of data, collected in the fall of the child's kindergarten year, and the second wave, collected the following spring. We restricted our sample to the 5,052 households who had incomes less than 130 percent of the federal poverty line, approximating the gross income requirements for Food Stamp receipt. Of those 5,052 households, 4,564 (90 percent) answered both rounds of the survey. After deleting cases that were missing information on the key dependent and independent variables, we have a final sample size of 4,410 households.

Variables of Interest

*Food Stamps.* We used two dichotomous variables to measure Food Stamp receipt, both taken from the first wave of the study. The first variable measures households in which at least one member had received Food Stamps during the previous twelve months. The second variable measures households in which at least one member had received Food Stamps for at least six of the past twelve months.[8] Slightly more than half (53 percent) of our sample received Food Stamps, and 43 percent of the sample had used Food Stamps for six months or more (see Table 1).[9]

*Food Insecurity.* Food Insecurity was measured at wave two, in the spring of the child's kindergarten year. As described in more detail above, it is measured by the Food Security Survey module, which classifies families as food secure, food secure without hunger, and food secure with hunger (Nord et al., 2003). Because only 4 percent of ELCS-K sample was classified as food insecure with hunger, we used a dichotomous variable representing those who are food insecure versus those who are not. Twenty-three percent of our sample was classified as food insecure (see Table 1). We also created a continuous variable that measured, among those who answered any food insecurity question positively, the amount of food insecurity present.[10] This variable was standardized to have a mean of 0 and a standard deviation of 1.

---

[8]This latter measure is taken from a question that asked how long the household had been receiving Food Stamps, using the following response scale: "1 to 2 months," "3 to 5 months," "6 to 8 months," or "8 to 12 months." Given that the scale intervals are not uniform, we created a dichotomous measure measuring Food Stamp receipt of six months or greater

[9]Our take-up rate of 53 percent is lower than what would be expected among a low-income sample (U.S. Committee on Ways and Means, 2004). We likely underestimate participation because we have no information on assets, and many households included in our sample may not qualify because their assets exceed the eligibility threshold (Daponte et al., 1999).

[10]Given the Rasch construction of the scale, the interval between answering no items affirmatively and one item affirmatively is unknown. Our level variable therefore only includes households who answered at least one item affirmatively, meaning that the range of the measure was from 1 to 17.

**Table 1**
**Descriptive Statistics**
**ECLS-K: Households with Incomes < 130% of Poverty Line**

| | Mean | SD |
|---|---|---|
| Received Food Stamps in past year | .53 | |
| Received Food Stamps 6 months in past year | .43 | |
| Classified as food insecure | .23 | |
| Amount of food insecurity[a] | 3.82 | (2.15) |
| Mother is white | .33 | |
| Mother is black | .26 | |
| Mother is Hispanic | .28 | |
| Mother is other race/ethnicity | .13 | |
| Father is different race | .08 | |
| Mother's age | 30.9 | (8.40) |
| Father's age | 34.7 | (7.81) |
| Household language is English | .75 | |
| Household size | 5.05 | (1.81) |
| Age of oldest sibling | 9.37 | (4.11) |
| Parents are married | .42 | |
| Parents are cohabiting | .12 | |
| Single parent | .46 | |
| Household education: No high school diploma | .25 | |
| Household education: high school diploma | .41 | |
| Household education: some college | .28 | |
| Household education: BA or higher | .06 | |
| Father did not work in past year | .10 | |
| Household income ($) | 14,117 | (7,499) |
| Mother received WIC during pregnancy | .73 | |
| Household received TANF in past year | .32 | |
| Household had serious financial problems | .37 | |
| No. of times child has moved | 2.40 | (1.63) |
| Household in large city | .27 | |
| Household in medium-sized city | .24 | |
| Household in large suburb | .18 | |
| Household in medium-size suburb | .07 | |
| Household in small town/rural area | .24 | |
| *observations* | *4,410* | |

[a]Among those who were classified as food insecure; n = 1,729.

Additional Covariates

The models also controlled for a number of demographic characteristics that previous research has linked to both Food Stamp receipt and food insecurity (Blank and Ruggles, 1996; Daponte et al., 1999; Nord et al., 2003; Cunnyngham and Brown, 2004). These included race/ethnicity, age, household structure, education and employment, income, financial stability, and geographic location. The demographic characteristic questions were asked of both the mother and the father; in households with only one parent, we imputed the value based on information provided by the other parent. If this information was not available, we also created a dummy variable to represent single-parent households. Descriptive statistics of all covariates are presented in Table 1.

*Race/Ethnicity and Age.* Race/ethnicity was measured through a series of dummy variables, where the omitted race category was white.[11] We also included a dummy variable for the father being of a different race than the mother. Age of mother and father is a continuous measure of their age.

*Household Structure.* Household structure includes the number of related members in the household and the age of the oldest child. Marital status was divided into three categories: married, cohabitating, or single-parent (married was omitted category).

*Educational Status and Employment.* The level of education measures the highest level of the resident parent's education, using four dummy variables: no high school diploma (omitted category), high school diploma only, some college attendance, or a bachelor's degree or higher. Employment included controls for mothers and/or fathers who were not working. In this case, we included in the not working category people who said that they were either unemployed, looking for a job, or out of the labor force, as these categories were too small to constitute their own categories.

---

[11]Mothers who were classified as being of another race includes 30 mothers who reported being non-Hispanic and of more than one race.

*Household Income and Financial Stability.* We used the log of the household's income to needs ratio, which calculates the relationship between the family's income and the official poverty line. Previous public assistance use was measured by two variables that asked if anyone in the household had received Temporary Assistance for Needy Families (TANF) in the past twelve months, and if the mother had used the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) when the child was an infant. The measure of financial problems comes from a one-item question that asked if, since the child was born, the household had any serious financial problems.

*Geographic Location.* We used a series of dummy variables to classify the household as residing in one of five types of neighborhoods, as classified by the Census Bureau. Large cities are Central Metropolitan Statistical Areas (CMSA) with populations greater than 250,000, medium-sized cities are CMSAs or MSAs with populations less than 250,000, large suburbs are towns on the outskirts of large cities, medium-sized suburbs are suburbs on the outskirts of medium-sized cities, and towns/rural areas include places that are not designated as CMSAs or MSAs (the latter serves as the omitted category).

## METHODS

Standard analyses of the relationship between Food Stamps and food insecurity involve estimating a multivariate regression equation like (1):

$$(1) \qquad \Pr(\text{Food Insecurity}_i=1 \mid X_i) = \beta_0 + \beta_1 FS_i + \Sigma\gamma_k X_{ki} + e_i$$

where the probability of the $i^{th}$ family experiencing food insecurity (that is, $\text{Food Insecurity}_i = 1$) is determined by the use of Food Stamps, or $FS_i$ (where 1 indicates use of Food Stamps and 0 indicates no use), and the *k* other factors ($X_k$s) that influence food insecurity. The coefficient of interest is $\beta_1$, which reflects the estimated change in the probability of being food insecure associated with family participation in Food Stamps. Estimates of $\beta_1$ will be unbiased if the error term is uncorrelated with $X_k$; although this could happen for several reasons, the main concern is that there are unobserved or omitted variables that might influence both Food Stamp participation and food insecurity. The extent of bias depends on the

correlation between unobserved determinants of Food Stamp usage and food insecurity and the unexplained variance in the outcome.

To address this potential endogeneity, we employ propensity scores. Propensity scores assume that, conditional on a given set of covariates X, the nonparticipation outcome ($Y_0$) is independent of the treatment (T):

(2) $$(Y_0 \perp T) \mid X$$

This equation formally states the conditional independence assumption and essentially identifies the counterfactual for the treatment group—for example, what their level of food insecurity would have been in the absence of Food Stamp receipt (Heckman et al., 1998; Smith and Todd, in press).

Instead of matching directly on X, we match on the predicted probability of using the treatment. This probability is calculated as follows:

(3) $$\Pr(\text{Food Stamps}_i=1 \mid X_i) = \beta_0 + \Sigma\gamma_k X_{ki} + e_i$$

Equation (3) is used to generate a predicted probability for each case. We then match treatment and control cases, using the three matching methods described above: nearest neighbor, kernel, and stratification matching.[12] With nearest neighbor matching, we match with replacement, which means that a control member can be matched with more than one treatment member.[13] Any cases not falling within the common support region were dropped. For kernel matching, we used a Gaussian kernel with a bandwidth of 0.6.[14] Following the recommendation of Dehejia and Wahba (1999, 2002), we ensured that our sample had been matched by dividing the common support region into strata and testing for significance difference on the covariates within each. As some covariates were not balanced, we added

---

[12]We tried alternative matching schema, including multiple nearest neighbor and local linear (a version of kernel matching). These alternative methods did not substantively change the results.

[13]Matching with replacement is preferred over the alternative, matching without replacement, in which control cases are discarded as soon as they are matched with a treatment case. In matching without replacement, as matching progresses, control units are matched with very dissimilar treatment cases, resulting in very bad matches.

[14]We tried alternative specifications for the bandwidth, but they did not affect the findings.

interaction terms until balance was achieved. The creation of these interaction terms was a theoretical and served only to balance our sample within strata. Finally, the data were randomly sorted to ensure that matching was not derived from the order of the observations.

Once the data were matched, we used two different approaches to calculate the impact of Food Stamps. First, we compared the difference in mean levels of food insecurity between the treatment and the control group. Second, we employed one of the "doubly robust" methods suggested by Imbens (2004), and used propensity scores as weights in a regression of food insecurity on Food Stamps. Using propensity scores as weights should ensure that the treatment indicator (for example, the use of Food Stamps) is not confounded with the potentially endogenous observed covariates (Hirano and Imbens, 2001; Imbens, 2004).

## RESULTS

### Calculating the Propensity Score

As noted above, the literature does not provide definitive guidance on how the propensity score should be calculated. A larger model guards against broader forms of omitted variable bias, as it is less likely to violate the conditional independence assumption, but a greater number of covariates may narrow the range of common support. In view of these competing tensions, we constructed two probability models, and the results are presented in Table 2. The first column presents the results from the smaller model (designated Model 1) and includes basic demographic information (race/ethnicity, age, household composition, education, and income). The second column includes all of these covariates but includes some additional variables potentially related to food insecurity and Food Stamp use: current public assistance receipt, residential instability, presence of financial problems, and the geographic location of the household (designated Model 2). Both models met the balancing criterion, as described above—that is, among groups stratified by the propensity score, the treatment and control groups did not differ in terms of the covariates used to calculate the propensity scores.

**Table 2**
**Predicted Likelihood of Ever Using Food Stamps**
**(ECLS-K: Households with Incomes <130% of Poverty Line)**

| | Model 1 | Model 2 |
|---|---|---|
| Mother is black | 0.380*** | 0.222*** |
| | [0.139] | [0.072] |
| Mother is Hispanic | 0.443 | -0.214** |
| | [0.572] | [0.090] |
| Mother is other race/ethnicity | 0.044 | -0.023 |
| | [0.087] | [0.114] |
| Father is different race than mother | 0.056 | 0.325 |
| | [0.103] | [0.301] |
| Age of mother | 0.013*** | 0.036*** |
| | [0.003] | [0.008] |
| Age of father | 0.020*** | 0.015*** |
| | [0.003] | [0.004] |
| Household language is English | 0.342*** | 0.180** |
| | [0.077] | [0.082] |
| No. of related members in household | 0.037** | 0.080*** |
| | [0.017] | [0.019] |
| Age of oldest sibling | 0.012 | 0.069* |
| | [0.008] | [0.037] |
| Parents are cohabiting | 0.316*** | 0.017 |
| | [0.076] | [0.081] |
| Single-parent household | 0.228* | 0.114 |
| | [0.121] | [0.132] |
| Household education: High school diploma | -0.158** | -0.089 |
| | [0.066] | [0.076] |
| Household education: Some college | -0.257*** | -0.153* |
| | [0.078] | [0.089] |
| Household education: BA or higher | -0.719*** | -0.498*** |
| | [0.120] | [0.137] |
| Mother not working | 0.519*** | 0.402*** |
| | [0.052] | [0.059] |

(table continues)

**Table 2, continued**

| | Model 1 | Model 2 |
|---|---|---|
| Father not working | 0.895*** | 0.641*** |
| | [0.090] | [0.094] |
| Household income (log) | -0.119*** | 1.456*** |
| | [0.031] | [0.213] |
| Mother received WIC | | 0.460*** |
| | | [0.060] |
| Household received TANF in past year | | 1.441*** |
| | | [0.075] |
| Household had serious financial problems | | 0.217*** |
| | | [0.057] |
| No. of times child has moved in lifetime | | 0.029* |
| | | [0.018] |
| Household in large city | | 0.086 |
| | | [0.084] |
| Household in medium sized city | | 0.154* |
| | | [0.079] |
| Household in large suburb | | -0.224*** |
| | | [0.086] |
| Household in medium sized suburb | | -0.124 |
| | | [0.099] |
| *Observations* | *4410* | *4410* |

**Notes**: Standard errors in brackets.
*significant at 10%; **significant at 5%; ***significant at 1%.

Results are consistent with previous research; those who are at higher levels of disadvantage are more likely to use Food Stamps (Blank and Ruggles, 1996; Currie, 2003). Model 1, which has fewer covariates, indicates that parents who are younger, not working, and have lower levels of education are more likely to use Food Stamps, as are parents who are cohabiting or single. Once the additional covariates are added (Model 2), age and marital status are no longer significant, but past and current use of public assistance and the presence of financial problems predict Food Stamp use.

The Common Support Problem

The different model specifications have implications for the common support region, as illustrated in Figure 1. The graphs plot the matched sample by their propensity scores, indicating the frequency of control group cases (below the line) and treatment cases (above the line) at .05 increments of their score (matching method in this instance was kernel). Cases that violated the common support region are also indicated. Figure 1 corresponds to Model 1 (the smaller model) and Figure 2 to Model 2 (the larger model). The common support condition mandates that the frequency of treatment and control groups should be reasonably symmetric about the line, indicating that the size of the treatment and control groups are roughly comparable.

In the model with fewer covariates (Figure 1), the distribution is somewhat symmetric, the majority of cases for both the treatment and control group having a propensity score between .04 and .78. However, introducing additional covariates in the larger model (Figure 2) moves the treatment group sharply to the right and the control group to the left. The gap between the average treatment and control score nearly doubles, from .22 in Model 1 (.63 versus .41, respectively) to .41 in Model 2 (.72 versus .31). The number of treatment cases with very high propensity scores (greater than .875) rises from 9 percent in Model 1 to 43 percent in Model 2. The number of control cases with very low propensity scores (less than .125) also increases, from 8 percent to 21 percent. In sum, because the larger model does a better job of predicting Food Stamp receipt, it makes the control and treatment groups less comparable and the imposition of a common support region less feasible. In terms of our results, Figure 2 indicates that we

**Figure 1**
**Graph of Propensity Score of Treatment and Control Cases:**
**Model 1**

Treated: Mean propensity score: .629
Control: Mean propensity score: .405

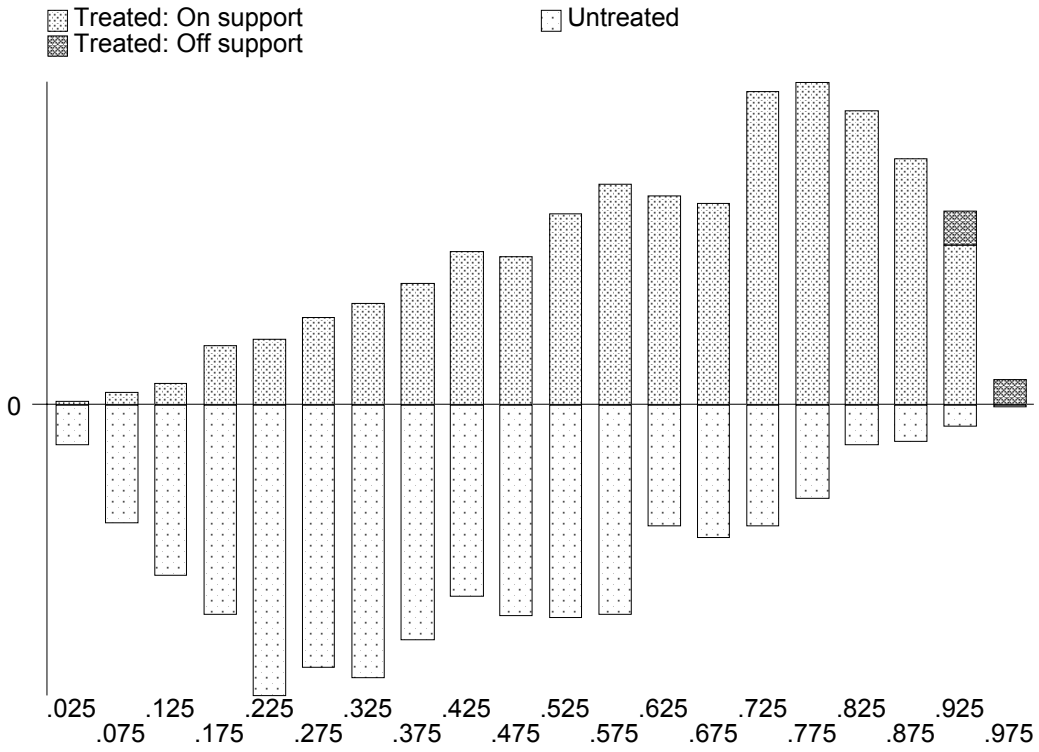
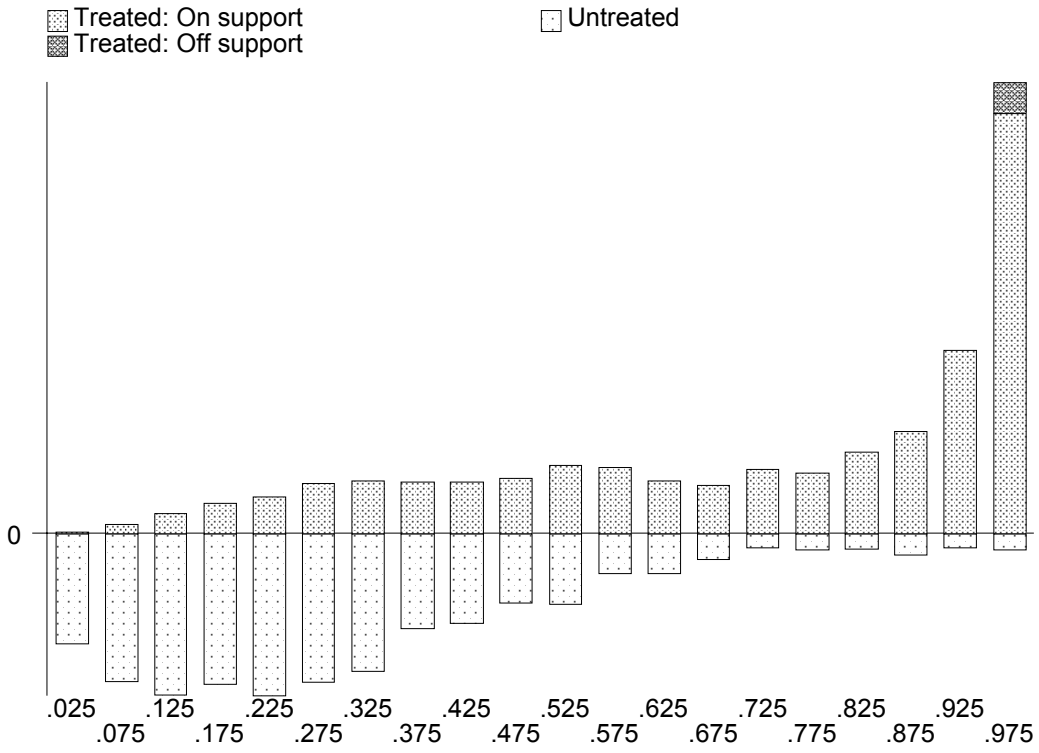


**Figure 2**
**Graph of Propensity Score of Treatment and Control Cases:**
**Model 2**

Treated: Mean propensity score: .719
Control: Mean propensity score: .307



will have a smaller sample of families that can be matched to their counterparts in the other group, but we will have more confidence that this subset is comparable.

Using both models, we next calculated the differences in food insecurity between the treatment and control groups. The difference was calculated in five ways: nearest neighbor matching, kernel matching, strata matching, regression with weighted propensity scores, and a conventional regression to serve as a comparison. In the regression models, we used a logistic model for the dichotomous outcome of food insecurity, and Ordinary Least Squares for the continuous measure. The standard errors for the matching methods are based on 500 bootstrapped replications. Note that the number of control observations will not be the same across estimations. Nearest neighbor matching will discard any control observations that cannot be matched to a treatment group, and all three matching methods discard any cases that violate the common support assumption.

Table 3 presents the results on the dichotomous outcome of being food insecure (top panel) and the amount of food insecurity (bottom panel) for the smaller and larger models:

The table indicates that the relationship between Food Stamps and the food security outcomes depends on the model specifications. For the dichotomous outcome of food insecurity, Model 1 yields statistically significant and negative estimates, indicating that receiving Food Stamps increases the likelihood of being food insecure. This pattern is consistent across different estimations, including the logistic regression model. The matching estimates range from .054 in kernel matching to .076 in nearest neighbor matching; both are smaller than the robust weighted regression estimate at .298 (p-value < .01 for all three). However, once additional covariates are added (Model 2), the estimates become negative, and none are statistically significant at conventional levels.

For the amount of food insecurity, Model 1 estimates are negative, but not significant. However, we find negative and significant effects of Food Stamp receipt for Model 2. The effect sizes range from a 14 percent decrease in the robust regression (p-value < .01) to a 26 percent decrease in nearest neighbor

**Table 3**
**Difference in Food Security Outcomes, by Matching Estimators and Linear Regression:**
**Treatment = Ever Received Food Stamps**
**(ECLS-K: Households with Incomes < 130% of Poverty Line)**

| | Model 1 | | | | | Model 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Matching Method | Difference | SE | T-score | Treatment | Control | Difference | SE | T-score | Treatment | Control |
| **Food Insecurity** | | | | | | | | | | |
| Nearest neighbor | .076*** | (.018) | 4.25 | 2,321 | 935 | -.040 | (.043) | -.931 | 2,321 | 692 |
| Kernel | .054*** | (.016) | 3.37 | 2,321 | 2,072 | -.010 | (.029) | -.340 | 2,321 | 2,084 |
| Stratification | .056*** | (.018) | 3.14 | 2,321 | 2,072 | -.015 | (.032) | -.459 | 2,321 | 2,084 |
| Weighted regression | .298*** | (.092) | 3.23[a] | 2,321 | 2,089 | -.102 | (.115) | -.890[a] | 2,321 | 2,089 |
| Logistic regression | .309*** | (.087) | 3.56[a] | 2,321 | 2,089 | -.013 | (.101) | -.130[a] | 2,321 | 2,089 |
| **Amount of Food Insecurity** | | | | | | | | | | |
| Nearest neighbor | .035 | (.074) | .47 | 1,051 | 386 | -.259*** | (.117) | -2.62 | 1,051 | 290 |
| Kernel | -.065 | (.064) | -1.02 | 1,051 | 677 | -.228** | (.106) | -2.16 | 1,051 | 646 |
| Stratification | -.052 | (.069) | -.75 | 1,051 | 677 | -.251** | (.116) | -2.16 | 1,051 | 646 |
| Weighted regression | -.086 | (.065) | -1.33 | 1,051 | 678 | -.144** | (.073) | -2.67 | 1,051 | 678 |
| Ordinary least squares | -.069 | (.060) | -1.15 | 1,051 | 678 | -.151** | (.064) | -2.36 | 1,051 | 678 |

**Notes**: All models estimated using the same set of covariates. The standard errors for the matching results were bootstrapped 500 times. Sample sizes will differ across models because of differences in the matching and regression methods.
*significant at 10%; **significant at 5%; ***significant at 1%
[a]Estimated with a z-score.

(p-value < .05). In general, the differences across the estimates are small, especially relative to their standard errors.

We also examined whether receiving Food Stamps for more than six months was significantly related to changes in the food security outcomes; these results are presented in Appendix Table A.2. We found the same overall pattern of effects as we found in Table 3, except that the estimates were generally not significant. For being food insecure, the smaller model (Model 1) estimates were positive, and the larger model (Model 2) estimates were negative. As for amount of food insecurity, both models indicated a negative relationship, but generally not a statistically significant one.

As one final estimation strategy, we examined the effect of Food Stamps for those whose propensity score is in the "thick" part of the distribution (for example, between .33 to .67). Propensity scores in this region are less likely to be biased even if the conditional independence assumption does not hold (Black and Smith, 2004); intuitively, this simply means that unobservable factors are more important for cases where the likelihood of receiving treatment is very high (that is, > 67 percent) or very low (< 33 percent). Thus, if we have identified a true effect of Food Stamps, we would expect this effect to be strong in this restricted sample.

The results for the larger model specification are presented in Table 4 (smaller model estimates not shown). Because of the additional restriction, the sample size has dropped from 4,410 to 1,212 cases (although the number of control cases will again vary). Despite this reduction in sample size, however, these results confirm our earlier findings, as they indicate the exact same pattern: no significant effect of receiving Food Stamps on being food insecure, but a significant reduction for the amount of food insecurity. For the amount of food insecurity, the matching estimates, save for nearest neighbor, have all increased in size and significance level from those that used the full range of the propensity score.

## DISCUSSION

We use propensity scores to investigate the effect of Food Stamps on food insecurity. We use this nonexperimental method because a randomized experiment in the Food Stamp program seems highly

**Table 4**
**Difference in Food Security Outcomes by Matching Estimators and Linear Regression:**
**Treatment = Ever Received Food Stamps**
**"Thick" Propensity Score Region**
**(ECLS-K: Households with Incomes < 130% of Poverty Line)**

| | Model 2 | | | | |
|---|---|---|---|---|---|
| Matching Method | Difference | SE | T-score | Treatment | Control |
| **Food Insecurity** | | | | | |
| Nearest neighbor | .005 | (.035) | .15 | 584 | 322 |
| Kernel | -.012 | (.028) | -.42 | 584 | 628 |
| Stratification | .007 | (.025) | .29 | 584 | 628 |
| Weighted regression | -.062 | (.155) | -.40[a] | 584 | 628 |
| Logistic regression | -.063 | (.154) | -.41[a] | 584 | 628 |
| **Amount of Food Insecurity** | | | | | |
| Nearest neighbor | -.240* | (.124) | -1.94 | 263 | 142 |
| Kernel | -.272*** | (.096) | -2.83 | 263 | 229 |
| Stratification | -.257*** | (.099) | -2.60 | 263 | 229 |
| Weighted regression | -.236*** | (.091) | -2.59 | 263 | 229 |
| Ordinary Least Squares | -.242*** | (.090) | -2.67 | 263 | 229 |

**Notes**: All models estimated using the same set of covariates. The standard errors for the matching results were bootstrapped 500 times. Sample sizes will differ across models because of differences in the matching and regression methods.
*significant at 10%; **significant at 5%; ***significant at 1%.
[a]Estimated with a z-score.

unlikely; moreover, conventional estimates of Food Stamp effects may be biased, given that the use of Food Stamps and food insecurity likely share common, unmeasured determinants. We find no effect of Food Stamps on the likelihood that a household will be food insecure, as the estimates were small and not consistent across the model specifications. This result is in keeping with that of Kabbani and Yazbeck (2004) and Oberholser and Tuttle (2004), but is contrary to that of Borjas (2004). We note, however, that Borjas examined the combined effect of receiving three types of public assistance (TANF, Food Stamps, or Medicaid), and perhaps his expansive definition accounts for the discrepancy. Previous research has not examined the effect of Food Stamps on levels of food insecurity; when we did so, we found that Food Stamps lessoned the amount of food insecurity with an effect size between 14 and 26 percent [a small effect size, according to Cohen's (1988) classification]. This result is consistent with the intent of the Food Stamp program as a safety net, and so it may be more effective in aiding those in more desperate need (Currie, 2003).

As an additional robustness check, we also examined the effect of Food Stamps for those who had a propensity score between 33 percent and 67 percent (the so-called "thick" region of the distribution). These estimates may be more reliable, because they are less sensitive to bias from unobserved or unmeasured characteristics (Black and Smith, 2004). These results confirm our earlier findings: we found no impact on the likelihood of being food insecure, but do find a reduction in the level of food insecurity.

There are two other potential explanations for the findings in the "thick" region: Food Stamps may be more effective at reducing food insecurity in this range, or food insecurity may be less well measured at the far ends of the distribution (Black and Smith, 2004). We do not believe either to be the case. There is no a priori reason to believe either that Food Stamps have a curvilinear relationship or that the sensitivity of the food security module varies with the score.

Some limitations to our paper should be discussed. First, we could only measure Food Stamp receipt by two dichotomous indicators. Previous research has indicated a dosage effect in Food Stamp

receipt, and we would have liked to investigate the effect of the amount of money received in Food Stamps (Rose et al., 1998). Unfortunately, the ELCS-K did not collect this information. As a crude proxy for dosage, we modeled the effect of receiving Food Stamps for six months or more; however, these models did not yield any significant impacts. Second, some research indicates that the food insecurity module may understate food inadequacy for households with younger children, as those of elementary age and younger are more likely to be spared food insecurity than are those who are older (Wilde, 2004). However, we controlled for age of oldest child to address this bias.

One advantage to using the ECLS-K is that it is an unusually rich dataset and includes information that may not routinely be collected. In fact, when we constructed a model with a smaller set of covariates (race, age, income, etc.) typically available in large-scale surveys, we found that Food Stamps increased the likelihood that a household would be food insecure. However, once we took advantage of measures found in the ELCS-K that may not be present in other surveys, we found a negative effect. This discrepancy highlights the need to collect thorough, comprehensive measures in large-scale surveys. We also note that propensity scores could not help us with model construction; both our smaller and larger models were found to balance the observed characteristics.

Propensity scores rely on two assumptions, conditional independence and common support. These two assumptions make competing demands on the methodology. The conditional independence assumption indicates that in order to ensure unbiased estimates, propensity scores should be based on a rich array of covariates. However, including more covariates in a model reduces the viability of the common support region, as there are fewer treatment (or control) cases for which the counterfactual can be constructed.

This tension was borne out in our results. Constructing a model with fewer covariates provided a thicker region of common support, but the model excluded certain covariates that were predictive of receiving the treatment. Adding additional covariates meant a richer model, but it also meant that the average treatment and control cases were quite dissimilar. As a result, the estimates from the model with

fewer covariates had smaller standard errors, but it is also likely that they were biased because they excluded certain key covariates.

However, this discrepancy in common support regions illustrates an advantage of propensity scores over other nonexperimental methods (Bryson et al., 2002; Black and Smith, 2004; Smith and Todd, in press). There may be some instances in which establishing a counterfactual for a certain case might not be possible, given the magnitude of the selection effects. However, the size of these selection effects is evident only because we used a nonparametric method like propensity scores; a parametric method, such as linear regression, would hide problems with the support region (Bryson et al., 2002; Black and Smith, 2004). The lack of common support does not invalidate the use of linear regression, as it can extrapolate to out-of-bounds cases. Yet if a large percentage of cases are off-support, and those cases differ systematically from on-support cases, then it is critical to identify the functional form of the model correctly.

It is also worth noting that the failure of the common support region depends to a large extent on the sample size. With a sufficiently large sample, it may be possible to find matches for any given propensity score, and the common support problem may be minimized. Even in this case, though, one must be aware that using fewer matches will reduce the efficiency of the estimates (Smith and Todd, in press).

Another advantage in using propensity scores is that, unlike linear regression, this method does not rely on a particular functional form. This is particularly salient in the case of Food Stamps and food insecurity, where there is no theoretical or empirical reason to believe that the effect of Food Stamps is constant. As mentioned above, there is good reason to believe the opposite is true, and that it will be more effective for those who are experiencing greater food insecurity.

A disadvantage to using propensity scores is that specification of the underlying probability model is crucial, and yet it is difficult to do this correctly. When we used a smaller set of covariates, we found a model that passed the matching balancing tests and had an acceptable common support region.

This model, however, was most likely incorrect, as it did not include all of the factors related to Food Stamp receipt and food insecurity. Furthermore, there is no guarantee that the larger model is correctly specified; there may still be significant omitted variable bias.

Furthermore, we also found that our estimates using regular linear regression methods were in line with the results from our propensity score models. This could mean that having a rich data set such as the ELCS-K, where many potentially confounding factors can be controlled for, could be sufficient for estimating the program's effects. We are wary of drawing this conclusion, as consistency between the linear regression and propensity score results does not guarantee that omitted variable bias is not a problem.

In sum, we conclude that propensity scores are not a magic bullet to solve the problem of omitted variable bias. The great discrepancy between treatment and control cases indicates that it may not be possible to measure and observe all the ways in which these two groups differ. It thus seems that a better use of propensity scores is to use them in conjunction with other nonexperimental methods and hope that they indicate consistent effects. This would provide some evidence that the researcher has identified a true treatment effect.

**Appendix Table A.1**
**List of Household Food Security Questions**
Answer categories in parentheses

1. "We worried whether our food would run out before we got money to buy more." (Sometimes, often, never)

2. "The food that we bought just didn't last and we didn't have money to get more." (Sometimes, often, never)

3. "We couldn't afford to eat balanced meals." (Sometimes, often, never)

4. In the last 12 months, did you or other adults in the household ever cut the size of your meals or skip meals because there wasn't enough money for food? (Yes/No)

5. (If yes to Question 4) How often did this happen—almost every month, some months but not every month, or in only 1 or 2 months?

6. In the last 12 months, did you ever eat less than you felt you should because there wasn't enough money for food? (Yes/No)

7. In the last 12 months, were you ever hungry, but didn't eat, because you couldn't afford enough food? (Yes/No)

8. In the last 12 months, did you lose weight because you didn't have enough money for food? (Yes/No)

9. In the last 12 months, did you or other adults in your household ever not eat for a whole day because there wasn't enough money for food? (Yes/No)

10. (If yes to Question 9) How often did this happen—almost every month, some months but not every month, or in only 1 or 2 months?

11. "We relied on only a few kinds of low-cost food to feed our children because we were running out of money to buy food." (Sometimes, often, never)

12. "We couldn't feed our children a balanced meal, because we couldn't afford that." (Sometimes, often, never)

13. "The children were not eating enough because we just couldn't afford enough food." (Sometimes, often, never)

14. In the last 12 months, did you ever cut the size of any of the children's meals because there wasn't enough money for food? (Yes/No)

15. In the last 12 months, were the children ever hungry but you just couldn't afford more food? (Yes/No)

16. In the last 12 months, did any of the children ever skip a meal because there wasn't enough money for food? (Yes/No)

17. (If yes to Question 16) How often did this happen—almost every month, some months but not every month, or in only 1 or 2 months?

18. In the last 12 months did any of the children ever not eat for a whole day because there wasn't enough money for food? (Yes/No)

**Appendix Table 2.A**
**Difference in Food Security Outcomes by Matching Estimators and Linear Regression:**
**Treatment = Received Food Stamps for More Than 6 Months**
**(ECLS-K: Households with Incomes < 130% of Poverty Line)**

| | Model 1 | | | | | Model 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Matching Method | Difference | SE | T-score | Treatment | Control | Difference | SE | T-score | Treatment | Control |
| **Food Insecurity** | | | | | | | | | | |
| Nearest neighbor | .043 | (.023) | 1.89 | 1,898 | 903 | -.042 | (.030) | -1.38 | 1,898 | 760 |
| Kernel | .023 | (.016) | 1.46 | 1,898 | 2,492 | -.026 | (.021) | -1.21 | 1,898 | 2,500 |
| Stratification | .026 | (.015) | 1.70 | 1,898 | 2,500 | -.022 | (.022) | -1.03 | 1,898 | 2,500 |
| Weighted regression | .051 | (.089) | .57[a] | 1,898 | 2,512 | -.164 | (.105) | -1.56[a] | 1,898 | 2,512 |
| Logistic regression | .110 | (.087) | 1.28[a] | 1,898 | 2,512 | -.062 | (.097) | -0.64[a] | 1,898 | 2,512 |
| **Level of Food Insecurity** | | | | | | | | | | |
| Nearest neighbor | -.013 | (.076) | -.17 | 845 | 376 | -.030 | (.087) | -.34 | 845 | 328 |
| Kernel | -.002 | (.057) | -.04 | 845 | 878 | -.040 | (.061) | -.66 | 845 | 870 |
| Stratification | -.006 | (.055) | -.12 | 845 | 878 | -.075 | (.058) | -1.30 | 845 | 870 |
| Weighted regression | -.130** | (.063) | -2.14 | 845 | 891 | -.138** | (.067) | -1.13 | 845 | 881 |
| Ordinary Least Squares | -.065 | (.059) | -1.11 | 845 | 891 | -.082 | (.060) | -1.36 | 845 | 881 |

**Notes**: All models estimated using the same set of covariates. The standard errors for the matching results were bootstrapped 500 times. Sample sizes will differ across models because of differences in the matching and regression methods.
*significant at 10%; **significant at 5%; ***significant at 1%.
[a]Estimated with a z-score.

**References**

Agodini, Roberto, and Mark Dynarski. 2004. "Are Experiments the Only Option? A Look at Dropout Prevention Programs." *Review of Economics and Statistics* 86(1): 180–194.

Bickel, Gary, Mark Nord, Cristofer Price, William L. Hamilton, and John T. Cook. 2000. *Guide to Measuring Household Food Security*. Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Black, Dan A., and Jeffrey Smith. 2004. "How Robust Is the Evidence on the Effects of College Quality? Evidence from Matching." *Journal of Econometrics* 121(1): 99–124.

Blank, Rebecca M., and Patricia Ruggles. 1996. "When Do Women Use Aid to Families with Dependent Children and Food Stamps?" *Journal of Human Resources* 31(1): 57–89.

Borjas, George J. 2004. "Food Insecurity and Public Assistance." *Journal of Public Economics* 88: 1421–1443.

Bryson, Alex, Richard Dorsett, and Susan Purdon. 2002. *The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies*. London: Department of Work and Pensions.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Cook, John T., Deborah A. Frank, Carol Berkowitz, Maureen M. Black, Patrick H. Casey, Diana B. Cutts, Alan F. Meyers, Nieves Zaldivar, Anne Skalicky, Suzette Levenson, Tim Heeren, and Mark Nord. 2004. "Food Insecurity Is Associated with Adverse Health Outcomes among Human Infants and Toddlers." *Journal of Nutrition* 134(6): 1432–1438.

Cunnyngham, Karen, and Beth Brown. 2004. *Characteristics of Food Stamp Households: Fiscal Year 2003*. Alexandria, VA: U.S. Department of Agriculture, Food and Nutrition Service.

Currie, Janet. 2003. "U.S. Food and Nutrition Programs." *Means-Tested Transfer Programs in the United States*, edited by Robert Moffitt. Chicago: University of Chicago Press.

Daponte, Beth Osborne, Seth Sanders, and Lowell Taylor. 1999. "Why Do Low-Income Households Not Use Food Stamps? Evidence from an Experiment." *Journal of Human Resources* 34(3): 612–628.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448): 1053–1062.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score Matching Methods for Non-Experimental Causal Studies." *Review of Economics and Statistics* 84: 151–161.

Dunifon, Rachel, and Lori Kowaleski-Jones. 2003. "The Influences of Participation in the National School Lunch Program and Food Insecurity on Child Well-being." *Social Service Review* 77: 72–92.

Foster, E. Michael. 2003. "Is More Treatment Better than Less? An Application of Propensity Score Analysis." *Medical Care* 41(10): 1183–1192.

Fraker, Thomas M. 1990. *The Effects of Food Stamps on Food Consumption: A Review of the Literature*. Washington, DC: Mathematica Policy Research, Inc.

Gundersen, Craig, and Victor Oliveira. 2001. "The Food Stamp Program and Food Insufficiency." *American Journal of Agricultural Economics* 83(4): 875–887.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64(4): 605–654.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(2): 261–294.

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

Heckman, James J., and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." *Longitudinal Analysis of Labor Market Data*, edited by James Heckman and Burton Singer. Cambridge, MA: Harvard University Press.

Hill, Jennifer L., Jeanne Brooks-Gunn, and Jane I. Waldfogel. 2002. "Differential Effects of High-Quality Child Care." *Journal of Policy Analysis and Management* 21(4): 601–627.

Hirano, Keisuke, and Guido W. Imbens. 2001. "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services and Outcomes Research Methodology* 2: 259–278.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86(1): 4–29.

Jensen, Helen L. 2002. "Food Insecurity and the Food Stamp Program." *American Journal of Agricultural Economics* 84(5): 1215–1228.

Kabbani, Nader S., and Myra Yazbeck. 2004. "The Role of Food Assistance Programs and Employment Circumstances in Helping Households with Children Avoid Hunger." Institute for Research on Poverty, Discussion Paper 1280-04. University of Wisconsin–Madison.

Kleinman, Ronald E., J. Michael Murphy, Michelle Little, Maria Pagano, Cheryl A. Wehler, Kenneth Regal, and Michael S. Jellinek. 1998. "Hunger in Children in the United States: Potential Behavioral and Emotional Correlates." *Pediatrics* 101(1): E3. Available at: http://www.pediatrics.org/cgi/content/full/101/1/e3.

Kramer-LeBlanc, Carol S., Peter Basiotis, and Eileen T. Kennedy. 1997. "Maintaining Food and Nutrition Security in the United States with Welfare Reform." *American Journal of Agricultural Economics* 79(5): 1600–1608.

Michalopoulos, Charles, Howard Bloom, and Carolyn J. Hill. 2004. "Can Propensity Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" *Review of Economics and Statistics* 86(1): 156–197.

Murphy, J. Michael, Cheryl A. Wehler, Maria E. Pagano, Michelle Little, Ronald E. Kleinman, and Michael S. Jellinek. 1998. "Relationship between Hunger and Psychosocial Functioning in Low-Income American Children." *Journal of the American Academy of Child and Adolescent Psychiatry* 37(2): 162–170.

Nord, Mark, Margaret Andrews, and Steven Carlson. 2003. *Household Food Security in the United States, 2002*. Washington, DC: U.S. Department of Agriculture, Food and Rural Economics Division, Economic Research Service.

Oberholser, Cheryl A., and Cynthia Reeves Tuttle. 2004. "Assessment of Household Food Security among Food Stamp Recipients in Maryland." *American Journal of Public Health* 94(5): 790–795.

Rose, Donald, Craig Gundersen, and Victor Oliveira. 1998. *Socio-Economic Determinants of Food Insecurity in the United States: Evidence from the SIPP and CSFII Datasets*. Alexandria, VA: U.S. Department of Agriculture.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Rosenbaum, Paul R., and Donald B. Rubin. 1985. "Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *Journal of the American Statistician* 39: 33–38.

Shadish, William, Thomas Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton-Mifflin.

Sianesi, Barbara. 2001. *An Evaluation of the Active Labour Market Programmes in Sweden*. London: Office of Labour Market Policy Evaluation (IFAU).

Slack, Kristen Shook, and Joan Yoo. 2004. "Food Hardships and Child Behavior Problems among Low-Income Children." Institute for Research on Poverty, Discussion Paper 1290-04, University of Wisconsin–Madison.

Smith, Jeffrey, and Petra Todd. In press. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*.

Stormer, Ame, and Gail G. Harrison. 2003. "Does Household Food Security Affect Cognitive and Social Development of Kindergartners?" Institute for Research on Poverty, Discussion Paper 1276-03, University of Wisconsin–Madison.

U.S. Committee on Ways and Means. 2004. 2004 *Green Book*. Washington, DC: U.S. Government Printing Office.

U.S. Department of Agriculture, Food and Nutrition Service. 2004a. "Food Stamp Program Participation and Costs." Available at http://www.fns.usda.gov/pd/fssummar.htm.

U.S. Department of Agriculture, Food and Nutrition Service. 2004b. "A Short History of the Food Stamp Program." Available at http://www.fns.usda.gov/fsp/rules/Legislation/history.htm.

Weinreb, Linda, Cheryl A. Wehler, Jennifer Perloff, Richard Scott, David Hosmer, Linda Sagor, and Craig Gundersen. 2002. "Hunger: Its Impact on Children's Health and Mental Health." *Pediatrics* 110(E41). Available at http://www.pediatrics.org/cgi/content/full/110/4/e41.

Wilde, Parke E. 2004. "Differential Response Patterns Affect Food-Security Prevalence Estimates for Households with and without Children." *Journal of Nutrition* 134(8): 1910–1916.