BLS WORKING PAPERS

U.S. Department of Labor
U.S. Bureau of Labor Statistics
Office of Productivity and Technology

Time Spent Exercising and Obesity:
An Application of Lewbel's Instrumental Variables Method

Charles Courtemanche
University of Kentucky, NBER, & IZA

Joshua C. Pinkston
University of Louisville

Jay Stewart
U.S. Bureau of Labor Statistics & IZA

Working Paper 521
January 2020

Time Spent Exercising and Obesity: An Application of Lewbel's Instrumental Variables Method[1]

Charles Courtemanche
University of Kentucky, NBER, & IZA

Joshua C. Pinkston
University of Louisville

Jay Stewart
Bureau of Labor Statistics & IZA

January 2020

[1] Contact Josh Pinkston at josh.pinkston@louisville.edu or (502) 852-2342. The views expressed in this paper are the authors' and do not necessarily reflect those of the U.S. Bureau of Labor Statistics.

Abstract

This paper examines the role physical activity plays in determining body mass using data from the American Time Use Survey. Our work is the first to address the measurement error that arises when time use during a single day—rather than average daily time use over an extended period—is used as an explanatory variable. We show that failing to account for day-to-day variation in activities results in the effects of time use on a typical day being understated. Furthermore, we account for the possibility that physical activity and body mass are jointly determined by implementing Lewbel's instrumental variables estimator that exploits first-stage heteroskedasticity rather than traditional exclusion restrictions. Our results suggest that, on average, physical activity reduces body mass by less than would be predicted by simple calorie expenditure-to-weight formulas, implying compensatory behavior such as increased caloric intake.

JEL Codes: I10, C21

Keywords: obesity; weight; exercise; physical activity; heteroskedasticity

1 Introduction

Despite a large body of research investigating interventions that may slow or reverse the well-documented rise in obesity, researchers still debate whether physical activity is effective at producing lasting weight loss. At issue is not whether caloric expenditure lowers weight if caloric intake is held constant, but whether exogenously induced increases in exercise lead to offsetting increases in calories consumed.[2] Many studies of exercise interventions have been small and non-representative (e.g., obese men, older women, hypertensive adults), and the results tend to vary across groups. A meta-analysis by Ross and Janssen (2001) finds that exercise interventions result in less weight loss than is predicted by standard models of calories burned. Thorogood et al. (2011) present another meta-analysis of fourteen studies that suggests aerobic exercise leads to modest reductions in weight and waist circumference, but not enough for aerobic exercise alone to be considered an effective weight loss therapy.

The small, non-representative nature of these exercise interventions has motivated research using large, nationally representative, observational datasets. For instance, Dunton et al. (2009), Kolodinsky et al. (2011), and Patel et al. (2016) document a negative association between time spent in physical activities and weight using the same dataset as we do here: the American Time Use Survey. However, these studies each suffer from two important problems.

The more obvious problem, which is widely recognized, is that individuals' exercise habits could be endogenous. Exercise may make obesity less likely, but obesity can also make

[2] This issue was the subject of a *Time* cover story entitled "Why Exercise Won't Make You Thin" (Cloud, 2009), which provided multiple anecdotes of compensatory eating. It also cited a study of almost 500 overweight middle-aged women in which the treatment groups, which were randomly assigned different amounts of exercise with a personal trainer, did not lose significantly more weight than the control group after six months (Church et al., 2009).

exercise more difficult. Furthermore, both physical activity and body mass may be influenced by an unobserved variable like self-discipline.

The other critical problem is measurement error in time-use variables. Ideally, we would have accurate information about the average amount of time individuals spend on various activities over a long period. In reality, researchers have either inaccurate measures covering a long period of time or more accurate measures from a short period of time. Retrospective surveys like the Behavioral Risk Factor Surveillance System or National Health Interview Survey that ask about physical activity during the past, say, 30 days introduce recall errors and provide ample room for social desirability bias.[3] Time diaries provide more accurate information, but tend to only cover a randomly chosen day on which one's level of exercise might be far from typical. Even if the resulting measurement error is random, it would lead to attenuation bias when time use is a right-hand side variable, such as when examining the effect of exercise on weight. Therefore, previous estimates that ignore this measurement error cannot even be interpreted as non-causal associations between physical activity and body mass.

Both endogeneity and measurement error could be addressed using instrumental variables, but valid instruments that predict long-run time use are difficult to find. In the absence of traditional instruments, we address these issues using an approach developed by Lewbel (2012) that exploits heteroskedasticity in mismeasured or endogenous explanatory variables to construct instrumental variables. This estimator replaces traditional exclusion restrictions with assumptions about the covariance of certain variables with the error terms. These covariance assumptions can be tested using familiar first-stage F-statistics and tests of overidentifying

[3] For example, Courtemanche and Zapata (2013) present estimates from the 2001, 2003 and 2005 BRFSS that suggest people exercise over 90 minutes per day, which is similar to the average minutes *per week* suggested by the ATUS time-diary data.

restrictions. However, we also discuss when these assumptions are more (or less) plausible, and we present alternative tests of our identifying assumptions.

Our results suggest that the effects of physical activity on body mass are nuanced. Time spent exercising (defined as physically active leisure) reduces body mass and the probability of being obese for women; however, we do not find evidence that exercise lowers the body mass of men, possibly due to changes in muscle mass or effects of exercise on appetite. On the other hand, time spent walking or biking that is not leisure (e.g., commuting or walking a dog) reduces the body mass of both men and women. When effects do emerge, they are smaller than would be predicted by simple calorie expenditure-to-weight formulas, implying some compensatory behavior.

While these results have obvious implications for the debate on the causal effect of exercise, they also contribute to the economics literature on how time use in general influences obesity. Cutler et al. (2003) argue that increased caloric intake associated with time-saving innovations in food processing, preparation, and preservation can help explain the rise in obesity. Along similar lines, Chou et al. (2004), Courtemanche et al. (2016), Currie et al. (2010), and Dunn (2010) document positive associations between the prevalence of restaurants – which reduces the time required to consume food – and BMI; however, Anderson and Matsa (2011) argue that the effect may not be causal. Lakdawalla and Philipson (2007) estimate a link between the physical intensity of a man's occupation and his body weight. Several studies find that maternal work hours are associated with an increase in childhood obesity or related behaviors.[4]

[4] These studies include Anderson et al. (2003), Ruhm (2008), Courtemanche (2009), Fertig et al. (2009), Liu et al. (2009), Morrissey et al. (2011), Cawley and Liu (2012), Morrissey (2012), Ziol-Guest et al. (2013), Abramowitz (2016), and Courtemanche et al. (2019).

Together, these studies and ours suggest that time use can be an important determinant of body weight.

2 Time Use as an Explanatory Variable

The main problem that must be dealt with when data from time diaries are used as explanatory variables is that the reference period in the sample is usually different from the reference period researchers are interested in.[5] In the current context, body mass is influenced by individuals' time use over previous years, but we only have data on time use during the previous day. As Frazis and Stewart (2012) point out, this source of measurement error must be dealt with even if researchers are only interested in non-causal associations between time in various activities and a dependent variable.

A second issue in our application (and others) is that the activity of interest could be endogenous. For example, exercise may have a causal effect on body mass; but unobserved factors that affect exercise, such as willpower, likely affect body mass through other avenues. Furthermore, body mass could affect the difficulty of exercise, introducing reverse causality.

A common approach to dealing with either measurement error or endogeneity is to use instrumental variables. The nature of the measurement error in our context requires instruments that predict long-run past time use, in addition to satisfying exclusion restrictions. We have not found any traditional instruments that satisfy both of these requirements.[6] Instead, we use the method developed by Lewbel (2012) that exploits heteroskedasticity in mismeasured or endogenous explanatory variables to construct instrumental variables.

[5] See Frazis and Stewart (2012) for a thorough discussion of problems in time-use studies caused by differences in reference periods.

[6] For example, more fitness centers may be located in communities with a high proportion of people who like to exercise.

For the sake of illustration, our initial assumptions about the error terms are stronger than required by Lewbel (2012) for identification.[7] First, we assume that time use is endogenous due to an unobserved common factor, μ. The equations we wish to estimate take the form:

$$BMI = X\beta_1 + M^*\gamma + \alpha_1\mu + \nu_1, \text{ and}$$

$$M^* = X\beta_2 + \alpha_2\mu + \nu_2,$$

where M^* is time spent in an activity on the average day over the period of interest, and X are exogenous explanatory variables. We also assume that μ and ν_j $(j = 1, 2)$ are conditionally uncorrelated with each other.

Observed time use on the diary day is $M = M^* + e_d$, where e_d is independent of M^*, BMI, and X. Using observed time use in place of average time use yields the following:

$$BMI = X\beta_1 + M\gamma + \varepsilon_1, \quad \varepsilon_1 = \alpha_1\mu + \nu_1 - \gamma e_d \tag{1}$$

$$M = X\beta_2 + \varepsilon_2, \quad \varepsilon_2 = \alpha_2\mu + \nu_2 + e_d \tag{2}$$

Intuitively, we can think of ν_2 as the long-run portion of the error term in equation (2), while e_d is the short-run error due to day-to-day variation in time use. As in Frazis and Stewart (2012), the above assumptions imply that e_d is independent of the long-run error term, ν_2.

Lewbel (2012) shows that heteroskedasticity in equation (2) can be used to construct instruments for endogenous or mismeasured variables. His estimator replaces traditional exclusion restrictions, which make assumptions about the coefficients in β_j, with assumptions about the covariance of certain variables with the error terms. This approach allows identification when the exclusion restrictions for available instruments are questionable, or traditional instruments are weak.

[7]This discussion roughly combines two examples discussed in Lewbel (2012).

Let Z denote a vector of exogenous variables.[8] Lewbel (2012) shows that $(Z - \bar{Z})\varepsilon_2$ are valid instruments for M under two assumptions:

$$\mathrm{Cov}(Z, \varepsilon_2^2) \neq 0 \qquad \text{(A1)}$$

$$\mathrm{Cov}(Z, \varepsilon_1\varepsilon_2) = 0. \qquad \text{(A2)}$$

In other words, Z is correlated with the heteroskedasticity in equation (2), but uncorrelated with the covariance between the error terms in equations (1) and (2). We can then obtain a consistent estimate of γ using 2SLS or GMM.

A sufficient condition for these assumptions to hold is for Z to be correlated with v_2^2, the heteroskedasticity associated with long-run time use, but conditionally independent of both μ^2 and e_d^2. Intuitively, this sufficient condition implies that Z is independent of day-to-day variation in time use, which is critical if we want to predict long-run time use instead of short-run variation.

As an example, consider rainfall as a potential Z variable. Long-run average rainfall could affect long-run time use, especially in outdoor activities, while also being conditionally independent of variation in time use yesterday from the long-run average (e_d). On the other hand, rainfall on the diary day is likely to predict time use on that day, making it correlated with the day-to-day variation that causes our measurement error. Long-run average rainfall, therefore, is more likely to satisfy (A2) than rainfall on the diary day is.

Although the assumptions made so far about ε_1 and ε_2 are sufficient for identification, they are stronger than is required by Lewbel (2012).[9] (A1) requires only that the error term in the

[8] In many applications, including the example in Lewbel (2012), Z is a subset of X; however, Lewbel points out that this is not required.

[9] For example, the variance in the day-to-day error, e_d^2, could vary with discipline or other unobserved factors without compromising identification.

time-use equation, ε_2, have heteroskedasticity that varies with some exogenous variable(s). The constructed instruments will be stronger when this covariance is higher, and weaker as it approaches zero. This assumption is easily tested using standard tests for heteroskedasticity, and is reflected in the F-statistic for $\left(Z-\overline{Z}\right)\varepsilon_2$ in first-stage regressions; however, it is important to note that those tests tell us nothing about whether a variable in Z is correlated with long-run or short-run components of the error term.

Assumption (A1) is easily satisfied in time-diary data. The structure of time-diary data, including the heteroscedasticity, is similar to that of the expenditures data Lewbel (2012) uses to demonstrate his approach. The existence of zeroes in the data due to activities (or purchases) not occurring during the reference period implies heteroskedasticity.[10]

Heteroskedasticity in time-use variables helps with identification because typical minutes spent in an activity are likely to be higher when the variance of the residual in the time-use equation is larger.[11] For example, if we consider two people who exercise every other day, the variance of the residual is larger for the person who exercises for two hours each time than the person who exercises for only 15 minutes. We illustrate this in Section 4.1 by comparing average time use and the standard deviation of residuals across groups in our sample.

Assumption (A2) ensures that the constructed instruments, $(Z-\bar{Z})\varepsilon_2$, are uncorrelated with ε_1 and are valid instruments. As Lewbel (2012) points out, any variable that is a valid instrument for M will satisfy assumption (A2), but the reverse is not true. A variable in Z can satisfy (A2) even if it is correlated with ε_1 (and thus not a valid instrument).

Continuing with our example, it is possible that long-run average rainfall is a valid

[10] See Keen (1986) for a discussion of heteroskedasticity in expenditure data. See Stewart (2013) for a discussion of similarities between time-use and expenditure data.
[11] See Rigobon (2003) and Berg et al. (2013) for related discussions of this intuition.

instrument for M. It is also possible that average rainfall affects the availability of indoor entertainment or other factors, which would make it invalid as a traditional instrument. But correlation with local indoor entertainment would not necessarily cause average rainfall to violate (A2). Lewbel's constructed instrument, therefore, can provide a second chance for a variable that may not be valid as a traditional instrument by isolating part of the variance in that variable that does not violate traditional exclusion restrictions.[12]

Fortunately, (A2) can be tested using standard tests of over-identifying assumptions. In what follows, we also use difference-in-Hansen tests to examine the exogeneity of subsets of our constructed instruments. We find some comfort in the fact that many of the variables one would expect to violate (A2), such as indicators for having young children, are rejected by these tests; however, we acknowledge that over-identification tests have shortcomings. As a result, we focus on Z variables that seem the most plausible intuitively, and we also present less-formal tests of our identifying assumptions.

3 Data

Our data come primarily from the Eating & Health (E&H) supplement to the 2006-2008 ATUS.[13] The ATUS is a time-diary survey that asks respondents to sequentially describe their activities, which are translated into over 400 detailed activity codes, during a 24-hour period that we refer to as the diary day.[14] For each episode, the ATUS collects the start and stop times, who else was present, and where the respondent was. The ATUS also contains demographic

[12] We also find that the constructed instruments are often stronger predictors of time use than the original Z is.
[13] A more complete description of the ATUS can be found in Hamermesh, Frazis, and Stewart (2005) or Frazis and Stewart (2012).
[14] If respondents report doing more than one thing at one time (e.g., cooking while talking to a child), only the primary (or "main") activity is coded. However, traveling is always considered the primary activity, even when done in conjunction with another activity. The diary day starts at 4am "yesterday" and ends at 4am "today."

information for all household members and labor force information (including labor force status and usual hours worked) for the respondent and the respondent's spouse or unmarried partner. The ATUS interviews one person per household and each respondent is interviewed only once about the day that precedes the day of the interview.

The E&H module, which was sponsored by the Department of Agriculture's Economic Research Service, collects information about eating and drinking as secondary activities, participation in SNAP and school meal programs, and whether the respondent usually does the shopping and meal preparation for the household. Respondents are also asked about their general health and to report their height and weight, which allows calculation of the body mass index (BMI).[15]

Since the work of Cawley (2002, 2004), it has been common practice in the economics literature on obesity to use validation data to correct for the tendency of survey respondents to misreport height and weight.[16] Typically, measured height and weight are regressed on polynomials of reported height and weight in the National Health and Nutrition Examination Survey (NHANES), and the resulting coefficient estimates are used to predict measured values in the primary sample.

Courtemanche, Pinkston and Stewart (2015) (CPS in what follows) demonstrate that the standard validation approach is inappropriate in most samples used to study obesity in the social sciences because the misreporting of height and weight is sensitive to survey context.[17] We apply an alternative correction developed by CPS that is robust to differences in misreporting across

[15] BMI = weight in kilograms divided by height in squared meters.
[16] As noted by Cawley (2002) and Rowland (1990), respondents tend to underreport weight and overreport height.
[17] See Courtemanche, Pinkston and Stewart (2015) for a discussion that compares data from BRFSS and the ATUS to NHANES data. The most obvious reason that survey context differs between the ATUS and NHANES is that ATUS respondents are interviewed by phone while NHANES respondents are interviewed in person prior to a physical examination in which they expect to be measured.

surveys, as long as the conditional expectations of actual measures are still increasing in their reported values in both samples. The implementation of the CPS correction is similar to the standard validation approach, but percentile ranks of reported values (instead of the reported values themselves) are used to predict measured values of height and weight.[18]

Our primary interest is in how time engaged in physical activities influences body mass and the probability that an individual is obese. Specifically, we focus on physically active leisure (exercise) and biking or walking that is not reported as leisure. Our biking or walking variable would include travel by foot or bicycle and walking a dog.

Our definition of exercise uses the mapping of ATUS activity codes to metabolic equivalents (METs) provided by Tudor-Locke, et al (2008). METs reflect the energy expended in an activity relative to the energy expended while at rest, which is assigned a MET of 1. We define exercise as any leisure activity having a MET value of 3 or higher, meaning that the activity requires at least three times the energy of being at rest.[19]

Our instrumental variables and some control variables come from supplementary sources. We use data on average surface temperatures and precipitation from NOAA for each MSA.[20] Our MSA-level measures of employment or establishment density in sports instruction, fitness centers, and restaurants (full-service or fast-food) come from the Quarterly Census of

[18] Following CPS, we append the 2006-2008 ATUS to the 2005-2006 and 2007-2008 waves of the NHANES. We then regress measured height (or weight) on a cubic basis spline of the percentile rank in reported height (weight), as well as a cubic polynomial in age using the NHANES observations of the combined data. Finally, we use the estimated coefficients to generate predicted values for both the NHANES and ATUS observations.

Because reporting patterns differ by sex and race, we run fully interacted regressions that are equivalent to separate regressions for each of 6 gender × race (white, black, and other) categories. We use sample weights so that the data from each survey are representative of the same populations. The sample restrictions mentioned elsewhere in this paper are not imposed on the ATUS data until after we correct BMI for measurement error.

[19] Third-tier ATUS activity codes could include a number of different activities, as evidenced by the examples listed in the ATUS coding lexicon. Tudor-Locke, et al (2008) assign the average MET value of the example activities to each third-tier activity code, which may place too much weight on relatively rare example activities. Fortunately, our definition of exercise appears to be robust to any distortions introduced by this averaging.

[20] Source: www.ncdc.noaa.gov/oa/climate/climatedata.html

Employment and Wages (QCEW).[21] Finally, we include data on MSA population, metro area density and median family income from the Census Bureau.

The sample is restricted to respondents between the ages of 20 and 64 who live in an identifiable MSA. The estimation sample has 11,109 women and 9,337 men with non-missing values of BMI, time use and other key variables. All estimates use ATUS sample weights.

Table 1 presents basic summary statistics for the sample. The average respondent in our sample is 41 years old with a (CPS-correction-adjusted) BMI just over 28. Nearly 62% of women and 73% of men in our sample are classified as overweight. Despite the difference in overweight status by gender, the incidence of obesity is around 33% for both women and men.

Table 2 presents summary statistics for the time-use variables used in the main estimation, as well as sleep and market work for comparison. In each case, averages taken with and without zeroes are included. For example, women exercise for less than 11 minutes per day on average, but those who report exercise on their diary day average nearly 70 minutes on that day. These differences reflect the fact that only 16% of women report any exercise on the diary day. In contrast, nearly all respondents report sleep, and the averages are similar regardless of the treatment of zeroes.

4 Applying Lewbel (2012) to Time-Diary Data

Any exogenous variable can be included in our vector of Z variables, as long as it satisfies assumptions (A1) and (A2). Lewbel (2012) and many applications of his method include all available exogenous variables in Z. In our application, the exogenous variables include location

[21] Source: http://www.bls.gov/cew/data.htm. Counts of employees or establishments in each industry are converted to numbers per 100 square miles to better reflect ease of access in each MSA. These variables are set to zero when missing because missing values primarily reflect BLS confidentiality rules that restrict disclosure for small cells.

characteristics, some of which might be suggested as traditional instruments; and individual characteristics like age and number of children. We view instrumental variables constructed from such personal characteristics with a great deal of skepticism. Instead, we focus on using Lewbel's method to improve identification based on MSA characteristics such as weather and prices.[22]

4.1 Heteroskedasticity and Assumption (A1)

The first requirement for the use of Lewbel's constructed IV is heteroskedasticity in the endogenous or mismeasured variables. In many contexts, the existence of heteroskedasticity is purely an empirical question. As discussed in Section 2, heteroskedasticity is expected a priori in time-diary data. This aspect of time-diary data, therefore, makes them particularly well suited to Lewbel's (2012) method.

The results in Table 2 confirm our expectations of heteroskedasticity in time-use variables. In addition to average minutes spent in each activity (with and without zeroes), the table presents χ^2 statistics from Breusch-Pagan tests for heteroskedasticity.[23] Heteroskedasticity is most pronounced for time spent walking or biking for reasons other than leisure, with $\chi^2(1)$ statistics of 6,980 for women and 2,304 for men. In contrast, the $\chi^2(1)$ statistics are below 150 for sleep, and even smaller for market work. The smallest χ^2 statistic in the table, for the test of heteroskedasticity in market work for men, has a p-value of 0.15. All of the other tests have p-values below 0.0001.

[22] See Hogan and Rigobon (2003) in addition to Lewbel (2012) for relevant discussions.
[23] These tests, which have one degree of freedom, are based on regressions of each time-use variable on the same MSA and individual characteristics used as explanatory variables in the regressions presented in Section 5. Tests based on regressions using different explanatory variables produce similar results.

Figure 1 illustrates how heteroskedasticity can help with identification. The graphs compare average minutes spent exercising and biking or walking with the standard deviation of residuals for that activity within year and state cells for women and men. As discussed by Frazis and Stewart (2012), average minutes spent on the diary days is the same as the average minutes on a typical day for any subpopulation because the day-to-day variation averages out.

Consistent with our discussion in Section 2, higher variance in the residuals of a time-use regression is associated with more minutes spent in that activity. The correlation coefficients of average minutes in an activity and the standard deviations of residuals within the relevant group are over 0.85 in each case, and all of the p-values are less than 0.0001.[24]

4.2 The Validity of Constructed Instruments

Assumption (A2) requires that the variables in Z be uncorrelated with the covariance between error terms in the time-use and BMI equations. This assumption is essential if $(Z - \bar{Z})\varepsilon_2$ are to be valid instrumental variables. Although (A2) can be tested using standard tests of overidentifying restrictions, we do not rely solely on those tests. Especially when a large number of variables are included in Z, overidentification tests may fail to reject instruments constructed using variables that do not satisfy (A2).

As described in Section 2, an implication of (A2) in our context is that the Z variables should be correlated with variation in long-run time use, without being correlated with day-to-day variation or with unobserved individual characteristics. We argue that MSA characteristics such as average weather and access to fitness centers are more likely to satisfy (A2) than individual characteristics like age or education are. Some of the local-area characteristics we

[24] Estimates of correlation coefficients and the linear fits shown in Figures 1A and 1B are weighted to account for the size of the state & year cells.

focus on may seem like potentially valid traditional instruments; however, they tend to be weak instruments in practice, or they require questionable exclusion restrictions.

Our application, therefore, is consistent with discussions in Lewbel (2012) and Hogan and Rigobon (2003) about using heteroskedasticity to improve identification based on local-area characteristics. A potential instrument included in Z satisfies (A2) under more general assumptions than are required by traditional exclusion restrictions. Furthermore, the Lewbel-style instrument is usually a stronger predictor of time use than the Z variable is by itself.

When we examine the effects of physical activities on body mass in the next section, we first present OLS estimates to provide a frame of reference. We then present estimates using Lewbel's approach with different sets of variables included in Z. We progress from specifications that include the full set of exogenous variables in Z to specifications that limit Z to variables we view as more likely to satisfy traditional exclusion restrictions. As a result, we can examine how coefficients and test statistics change as we move from identifying assumptions we are most skeptical of to the assumptions we believe are most plausible.

5 Results

All of the regressions that follow include a cubic polynomial in age, as well as dummy variables for year, race, and education level.[25] We also include the following MSA characteristics: region indicators; population and population per square mile; the unemployment rate; median income; average annual temperature, average annual rainfall, and frequency of days with more than half an inch of precipitation. Finally, we include counts per 100 square miles of fitness centers, jobs in sports instruction establishments, fast-food restaurants, and full-service restaurants.

[25] Our results are robust to the inclusion of controls for marital status, family income, number of children, and age of children; however, we exclude those variables from our preferred specifications due to possible endogeneity.

To alleviate concerns that instrumental variables based on heteroskedasticity may be weaker than suggested by first-stage *F*-statistics, we estimated all of our models using both 2SLS and Fuller modified LIML estimators. The Fuller estimates are more robust to weak instruments than 2SLS, which means that differences between 2SLS and Fuller estimates would suggest weak instruments. We saw no such differences in estimates for the activities we discuss below, so we present results only from the more robust Fuller estimators.

5.1 Main Results

Table 3 presents estimates of the effects of exercise, defined as physically active leisure, on body mass. The OLS coefficients suggest that minutes of exercise yesterday are associated with lower body mass for women, but the analogous estimates for men suggest little (if any) association. The OLS coefficient in column (1) implies that 30 minutes of exercise on the diary day is associated with BMI being a little over half a point lower for women.

When we use Lewbel's approach with the largest set of Z variables, the estimated effect of exercise on BMI falls slightly relative to the OLS coefficient; however, it rises as we restrict the set of Z variables to those that we believe produce more plausible instruments. The smaller coefficients in specifications that use all available Z variables could be explained by some of the larger set of instruments being invalid, or due to the well-known downward bias that can result from using a large number of instruments (especially if some of those instruments are weak). We address both potential problems by using fewer and more plausible instruments.

The bottom set of estimates in Table 3 use instruments that are constructed using the density of fitness centers in the MSA, jobs in sports instruction and weather variables. The estimated effect of exercise on BMI for women increases in magnitude to −0.032 (0.016), which suggests 30 minutes of exercise on the typical day lower BMI by nearly 1. The estimated effects

of exercise on the probabilities of being overweight or obese also increase in magnitude, but are no longer statistically significant.

In contrast to the results for women, we find no evidence that exercise affects the BMI of men in Table 3. However, this does not imply that exercise does not have other health benefits. It is possible that exercise simply increases muscle mass for men as much as it reduces body fat. It's also possible that men are more likely to increase caloric intake in response to exercise than women are. Without data on body composition or calories consumed, we cannot rule out either possibility.

On the other hand, biking or walking for reasons other than exercise is associated with lower body mass for both men and women. The OLS coefficients in Table 4 from the BMI regressions are −0.025 (0.005) for women and −0.028 (0.005) for men. The coefficients in the linear probability models for obesity are also statistically significant above any conventional level for both men and women, as is the coefficient in the model for overweight status among men. This suggests that men and women may not view these activities as exercise *per se* and therefore may not completely offset these calories burned by eating more.

The Lewbel IV estimates in Table 4 again suggest larger effects as we use fewer and more plausible instruments. In the final set of estimates, where we only use long-run weather variables to construct our instruments, the coefficients in the BMI equations are −0.035 (0.015) for women and −0.050 (0.020) for men.[26] These coefficients suggest that averaging 30 minutes of biking or walking per day lowers the BMI of women by more than 1, and lowers the BMI of men by more than 1.5. Furthermore, 30 minutes of biking or walking per day lowers the probability of a man being overweight by roughly 16 percentage points.

[26] Specifically, we use average annual temperature, average annual rainfall, and the frequency of days with more than half an inch of rain.

The results in both Tables 3 and 4 suggest that the bias in OLS regressions caused by measurement error in the time diary data is more severe for these activities than the bias from endogeneity. We would expect the measurement error introduced by using time yesterday in place of time on the typical day to bias coefficients toward zero. On the other hand, bias from either reverse causality or unobserved factors like discipline would likely make OLS coefficients more negative than the true causal effects.[27] The fact that our preferred IV estimates in Tables 3 and 4 are more negative than the corresponding OLS coefficients is consistent with the bias due to measurement error being larger than the bias from endogeneity.

5.2 Testing Assumptions

The Hansen *J*-tests in Table 4 reject the validity of using the full set of exogenous variables to construct instruments in the equations for BMI and overweight status for men. Furthermore, difference-in-Hansen tests (not shown) often reject the validity of instruments constructed using personal characteristics, even in cases where the Hansen test does not reject overidentification.[28] Despite the fact that tests of overidentification have more power when fewer instruments are used, we never reject the validity of our preferred instruments. This supports our view that some variables result in more plausible constructed instruments than others, and suggests that researchers should apply Lewbel (2012) with care.

Baum and Lewbel (2019) point out that a violation of the assumption (A2) that $\mathrm{Cov}(Z, \varepsilon_1\varepsilon_2) = 0$ would imply heteroskedasticity with respect to Z in equation (1), the BMI regression. They suggest using the test for heteroskedasticity developed by Pagan and Hall

[27] If being heavier makes physical activity more difficult, we would expect negative OLS coefficients in Tables 3 and 4 even if physical activity had no effect on body mass. Unobserved discipline would likely be correlated with increased physical activity, as well as other behaviors that would affect BMI.
[28] The difference-in-Hansen results also suggest that the rejections we see in Hansen tests is not due to random chance.

(1982) for regressions with endogenous regressors; however, they also note that there could be heteroskedasticity in the BMI regression for reasons that are unrelated to (A2). Therefore, testing for heteroskedasticity in the BMI regressions cannot *reject* (A2), but it may provide reassurance that (A2) is plausible.

We view the results of these heteroskedasticity tests as consistent with (A2) overall. When we test for heteroskedasticity that is correlated with our preferred Z variables in BMI regressions, we fail to reject homoscedasticity in most cases.[29] In contrast, we strongly reject homoscedasticity every time we test for heteroskedasticity associated with variables outside of our preferred Z variables, or when we expand Z to include more variables. The one case in which we find evidence of heteroskedasticity associated with our preferred Z is the BMI regression for women with exercise as the endogenous variable; however, evidence of heteroskedasticity associated with regressors that aren't in Z is also stronger in this regression than in any other, which increases the likelihood that the heteroskedasticity we find is benign.[30]

6 Concluding Remarks

The impact of time use on the likelihood of becoming obese is an important, but under-researched area. One of the reasons is that the ideal data do not exist. Ideally, we would have reliable data on long-run time use, such as average time spent exercising. Retrospective survey questions may include reported long-run time use, but such reports are subject to recall and

[29] Results (not shown) are available on request. The Pagan/Hall test can be performed in Stata using ivhettest.ado, which was written by Mark Schaffer; however, we modified the ado file to work with sample weights, and are responsible for any mistakes. We only considered these tests for the BMI regressions to avoid the heteroskedasticity that is inherent in linear probability models.

[30] The differences in heteroskedasticity tests is especially large across gender, with test statistics being up to four times larger for women than for men.

social desirability biases. Time-diaries, while more accurate, cover only one day and may be a poor representation of individuals' long-run time use.

In addition to measurement issues, time use is likely endogenous. We expect physical activity to reduce BMI, but being overweight or obese may also make exercise more difficult. Or unobserved factors, such as discipline, may affect both BMI (perhaps through eating habits) and inclination to exercise.

A common solution to both of these issues is to use instrumental variables. But it is often difficult to find instruments that are both strong and truly exogenous. We address these problems by using the heteroskedasticy-based IV procedure proposed by Lewbel (2012), which replaces traditional exclusion restrictions (assumptions about coefficients) with assumptions about the covariance of error terms. Time-diary data are well-suited to Lewbel's method because, with the large number of zero-value observations, errors are naturally heteroskedastic. As a result, they are similar to expenditure data, which Lewbel uses to illustrate his method.

Essentially, Lewbel's procedure requires a variable that is correlated with heteroskedasticity in the first-stage regression but independent of the covariance between error terms of the first- and second-stage regressions. Variables that satisfy traditional exclusion restrictions also satisfy this covariance assumption; however, variables that do not satisfy the exclusion restriction can still satisfy this covariance assumption. Therefore, variables that may not be valid as traditional instruments get a "second chance" via Lewbel's constructed IV approach.

Our results differ somewhat for men and women. We find that time spent exercising reduces BMI for women, but has no statistically significant effect for men. It is not clear whether this is due to men gaining muscle mass or increasing caloric consumption in response to

excercise. In contrast, time spent biking or walking for reasons other than exercise reduces BMI for both men and women, with the effects for men being larger.

Coefficients from our preferred models are consistently larger than OLS coefficients, which suggests measurement error in our time-use variables introduces more bias than reverse causality or other sources of endogeneity. The results from our preferred models are also stronger than those from models that use larger, less intuitively appealing, sets of instruments. More importantly, our overidentification tests never reject the validity of our preferred instruments, but often reject instruments constructed using those variables (e.g., individual age) which no one would suggest as a traditional instrumental variable.

While our preferred IV estimates suggest larger effects of physical activity on BMI than OLS estimates do, they still suggest "real world" effects that are more modest than might be expected based purely on calories burned. For example, an additional 30 minutes per day of either type of physical activity we consider would lower the BMI of women in our sample by 1 (or 3.5%) on average. Biking or walking for 30 minutes more per day would lower the BMI of the average man by 1.5 (over 5%). At average heights (5'4" for women and 5'9" for men), these reductions in BMI would be equivalent to 6 pounds of weight loss for women and 10 pounds for men. In contrast, the average man who started walking briskly for 30 minutes per day might expect to lose twice as much weight based on online calorie calculators, and the average woman might expect to lose 2.5 times more.[31] These results provide support for the hypothesis of compensatory calorie intake in response to an exogenously induced change in physical activity.

[31] For example, Harvard Medical School presents tables of estimated calories burned by people of three different weights during 30 minutes of various activities at this link: https://www.health.harvard.edu/diet-and-weight-loss/calories-burned-in-30-minutes-of-leisure-and-routine-activities
Our back-of-the-envelope calculations for the average man and woman are based on the Harvard estimates for a person weighing 185 and 155 pounds, respectively. Since the average weights in our sample are 195 and 161 pounds, a naïve person may actually view our calculations as conservative.

References

Abramowitz, Joelle (2016): "The Connection between Working Hours and Body Mass Index in the U.S.: A Time Use Analysis." *Review of Economics of the Household* 14(1): 131-154.

Anderson, M.L. and Matsa, D.A. (2011): "Are Restaurants Really Supersizing America?" *American Economic Journal: Applied Economics* 3: 152-188.

Anderson, Patricia M., Kristin F. Butcher, and Phillip B. Levine (2003): "Maternal employment and overweight children." *Journal of Health Economics* 22(3): 477-504.

Burkhauser, Richard and John Cawley (2008): "Beyond BMI: The value of more accurate measures of fatness and obesity in social science research." *Journal of Health Economics* 27(2): 519-529.

Cawley, John (2002): "Addiction and the consumption of calories: Implications for obesity." Unpublished manuscript, Cornell University.

Cawley, John (2004): "The impact of obesity on wages." *Journal of Human Resources* 39(2): 451-474.

Cawley, John and F. Liu. (2012): "Maternal employment and childhood obesity: A search for mechanisms in time use data." *Economics and Human Biology* 10(4): 352-364.

Chou, Shin-Yi, Michael Grossman, Henry Saffer (2004): "An Economic Analysis of Adult Obesity: Results from the Behavioral Risk Factor Surveillance System", *Journal of Health Economics* 23: 565-587.

Church, Timothy S., Corby K. Martin, Angela M. Thompson, Conrad P. Earnest, Catherine R. Mikus, and Steven N. Blair (2009): "Changes in Weight, Waist Circumference, and Compensatory Responses with Different Doses of Exercise among Sedentary, Overweight Postmenopausal Women." *PLoS One* 42(2): e4515.

Cloud, John (2009): "The Myth About Exercise." *Time*, August 17, 2009, 174(6): 42-47.

Courtemanche, Charles (2009): "Longer Hours and Larger Waistlines? The Relationship Between Work Hours and Obesity," *Forum for Health Economics and Policy* 12: Article 5.

Courtemanche, Charles, Joshua C. Pinkston, and Jay Stewart (2015): "Adjusting Body Mass for Measurement Error with Invalid Validation Data." *Economics and Human Biology* 19: 275-293.

Courtemanche, Charles, Joshua C. Pinkston, Christopher Ruhm, and George Wehby (2016): "Can Changing Economic Factors Explain the Rise in Obesity?" *Southern Economic Journal* 82(4): 1266-1310.

Courtemanche, Charles, Rusty Tchernis, and Xilin Zhou (2019): "Maternal Work Hours and Childhood Obesity: Evidence using Instrumental Variables Related to Sibling School Eligibility." *Journal of Human Capital* 13(4): 553-584.

Currie, J., DellaVigna, S., Moretti, E. and Pathania, V. (2010): "The Effect of Fast Food Restaurants on Obesity and Weight Gain." *American Economic Journal: Economic Policy* 2: 32-63.

Cutler, David M., Edward L. Glaeser, and Jesse M. Shapiro (2003): "Why Have Americans Become More Obese?" *Journal of Economic Perspectives* 17(3): 93-118.

Dunn, R. (2010): "Obesity and the Availability of Fast-Food: An Analysis by Gender, Race/Ethnicity and Residential Location." *American Journal of Agricultural Economics* 92: 1149-1164.

Dunton, G.F., D. Berrigan, R. Ballard-Barbash, B. Graubard, and A.A. Atienza (2009): "Joint associations of physical activity and sedentary behaviors with body mass index: results from a time use survey of US adults." *International Journal of Obesity* 33: 1427-1436.

Faberman, R. Jason (2010): "Revisiting the role of home production in life-cycle labor supply." No. 10-3. Federal Reserve Bank of Philadelphia.

Fertig, A., G. Glomm, and R. Tchernis (2009): "The connection between maternal employment and childhood obesity: inspecting the mechanisms." *Review of Economics of the Household*,7(3): 227-255.

Frazis, Harley, and Jay Stewart (2007): "Where does the time go? Concepts and measurement in the American Time Use Survey." *Hard-to-measure goods and services: Essays in honor of Zvi Griliches*. University of Chicago Press. 73-97.

Frazis, Harley, and Jay Stewart (2012): "THE QUALITY OF DIARIES: How to Think about Time-Use Data: What Inferences Can We Make about Long-and Short-Run Time Use from Time Diaries?" *Annales d'Economie et de Statistique* 105: 231-246.

Hamermesh, Daniel S., Harley Frazis, and Jay Stewart (2005): "Data Watch: The American Time Use Survey." *Journal of Economic Perspectives* 19: 221-232.

Hamermesh, Daniel S., Caitlin Knowles Myers, and Mark L. Pocock (2008): "Cues for Timing and Coordination: Latitude, Letterman, and Longitude." *Journal of Labor Economics* 26(2): 223-246.

Jiminez-Pavon, David, Joanna Kelly, and John J. Reilly (2010): "Associations between objectively measured habitual physical activity and adiposity in children and adolescents: Systematic review." *International Journal of Pediatric Obesity* 5(1): 3-18.

Keen, Michael (1986): "Zero expenditures and the estimation of Engel curves." *Journal of Applied Econometrics* 1(3): 277-286.

Kolodinsky, Jane M. and Amanda B. Goldstein (2011): "Time Use and Food Pattern Influences on Obesity". *Obesity* 19(12): 2327-2335.

Lakdawalla, Darius, and Tomas Philipson (2007): "Labor Supply and Weight." *The Journal of Human Resources* 42(1): 85-116.

Lewbel, Arthur (2012): "Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models." *Journal of Business & Economic Statistics* 30(1): 67-80.

Liu, E., Hsiao, C., Matsumoto, T., & Chou, S. (2009): "Maternal Full-Time Employment and Overweight Children: Parametric, Semi-Parametric, and Non-Parametric Assessment." *Journal of Econometrics* 152: 61-69.

Morrissey, T. W. (2012): "Trajectories of Growth in Body Mass Index across Childhood: Associations with Maternal and Paternal Employment." *Social Science and Medicine* 95: 60-68.

Morrissey, T. W., Dunifon, R. E., & Kalil, A. (2011): "Maternal Employment, Work Schedules, and Children's Body Mass Index." *Child Development* 82: 66-81.

Mullahy, John and Stephanie A. Robert (2008): "No Time to Lose? Time Constraints and Physical Activity." NBER Working Paper 14513.

Patel, Viral C., Andrea M. Spaeth, & Mathias Basner (2016): "Relationships between Time Use and Obesity in a Representative Sample of Americans." *Obesity* 24(10): 2164-2175.

Rigobon, Roberto (2003): "Identification Through Heteroskedasticity." *The Review of Economics and Statistics* 85(4): 777-792.

Ross, Robert and Ian Janssen (2001): "Physical activity, total and regional obesity: Dose-response considerations." *Medicine & Science in Sports & Exercise* 33(6): S521-S527.

Rowland, M.L. (1990): "Self-reported weight and height." *American Journal of Clinical Nutrition* 52(6): 1125-1133.

Ruhm, Christopher (2008): "Maternal employment and adolescent development." *Labour Economics* 15(5): 958-983.

Stewart, Jay (2013): "Tobit or not Tobit?" *Journal of Economic and Social Measurement* 38(3): 263-290.

Tudor-Locke, Catrine, Tracy L. Washington, Barbara E. Ainsworth, and Richard Troiano (2008): "Linking the American Time Use Survey (ATUS) and the Compendium of Physical Activities: Methods and Rationale" *Journal of Physical Activity and Health* 6(3): 347-353.

Ziol –Guest, K., R. Dunifon, & A. Kalil (2013): "Parental Employment and Children's Body Weight: Mothers, Others, and Mechanisms." *Social Science and Medicine* 95: 52-59.

Figure 1. Heteroskedasticity & Typical Time Use

Average Minutes in Activities Within State & Year

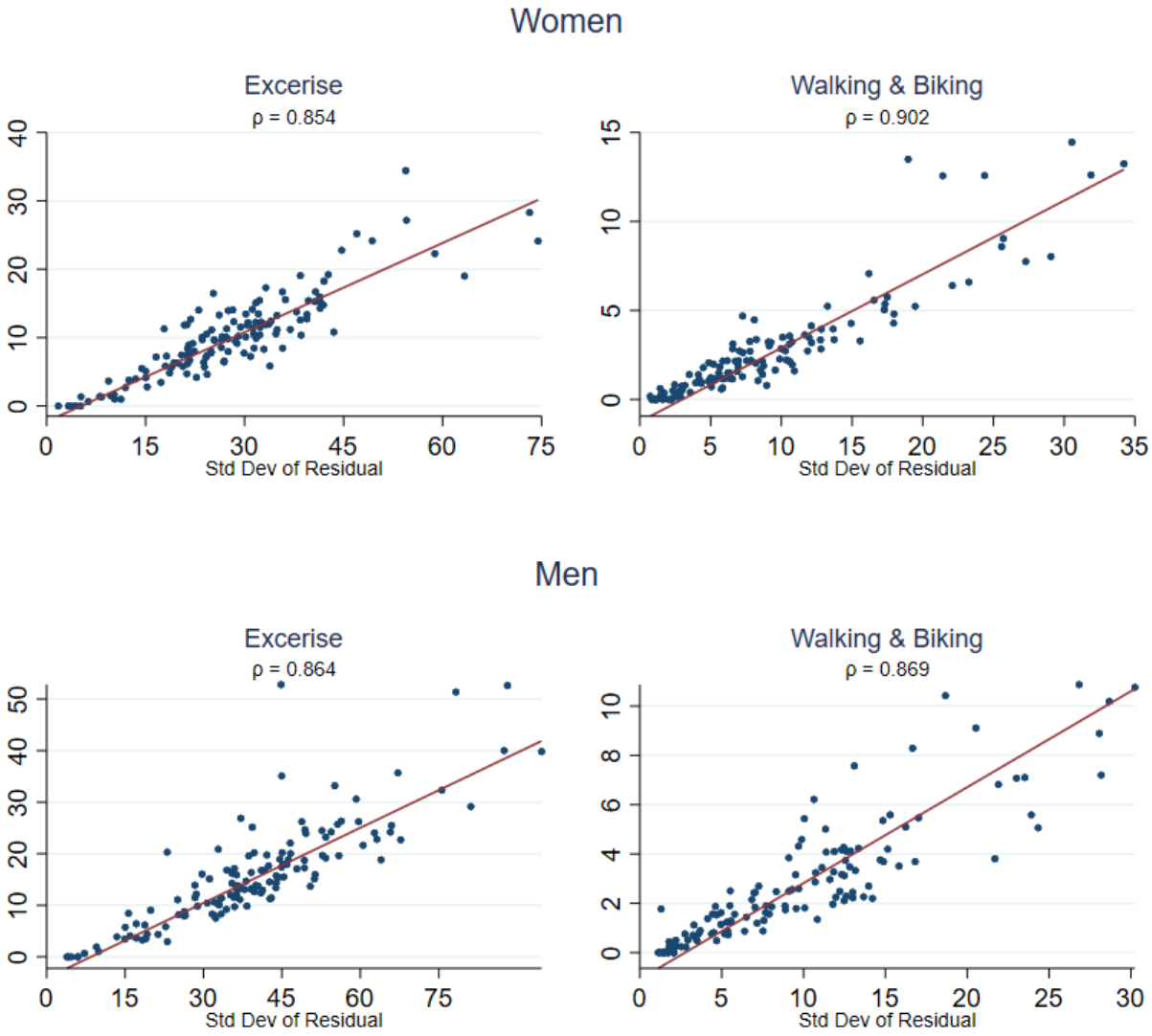

Table 1. Summary Statistics

	Women		Men	
	Mean	Std. Dev.	Mean	Std. Dev.
BMI	28.355	7.236	28.541	5.890
Overweight	0.619	0.486	0.726	0.446
Obese	0.331	0.470	0.328	0.470
Age	41.197	12.456	40.790	12.283
White	0.787	0.409	0.818	0.386
Black	0.140	0.347	0.117	0.322
Other Race/ethnicity	0.073	0.260	0.065	0.246
Observations	11,109		9,337	

Notes: All estimates use ATUS sample weights. BMI, Overweight and Obese are calculated using the CPS percentile-rank measurement error correction described in the text.

Table 2. Summary Statistics for Time Spent in Various Activities.

	Women				Men			
	Mean & Std. Dev.		Percent	Breusch-Pagan	Mean & Std. Dev.		Percent	Breusch-Pagan
	W/ Zeroes	No Zeroes	Non-zero	Het. Test $\chi2(1)$	W/ Zeroes	No Zeroes	Non-zero	Het. Test $\chi2(1)$
Exercise	10.82 (31.91)	69.7 (49.55)	15.5%	832.38	16.17 (44.04)	88.42 (64.97)	18.3%	651.44
Walking & Biking, Not as Exercise	3.564 (14.28)	24.72 (29.86)	14.4%	6,979.81	3.515 (13.53)	25.24 (27.68)	13.9%	2,304.09
Sleep	502.3 (131.84)	502.6 (131.31)	99.9%	120.18	497.3 (133.64)	497.9 (132.70)	99.9%	147.29
Market Work	230.7 (266.23)	458.8 (190.51)	50.3%	64.67	333.9 (299.10)	522 (204.14)	64.0%	2.06

Note: All times are minutes per day.

Table 3. The Effects of Exercise on Body Mass

	Women			Men		
	BMI	Overweight	Obese	BMI	Overweight	Obese
OLS	-0.0189***	-0.0010***	-0.0012***	-0.0026	< 0.0001	-0.0003*
	(0.0022)	(0.0002)	(0.0001)	(0.0018)	(0.0001)	(0.0002)
All Exogenous Variables Included in "Z"						
Lewbel IV	-0.0165***	-0.0017***	-0.0009**	0.0050	0.0002	0.0005
	(0.0048)	(0.0005)	(0.0003)	(0.0053)	(0.0005)	(0.0005)
First-Stage *F* -Stat.	92.34	92.34	92.34	58.51	58.51	58.51
Hansen *p* -value	0.313	0.678	0.513	0.697	0.827	0.723
All MSA Characteristics Included in "Z"						
Lewbel IV	-0.0277**	-0.0018*	-0.0012*	0.0001	-0.0003	0.0002
	(0.0108)	(0.0011)	(0.0007)	(0.0111)	(0.0008)	(0.0009)
First-Stage *F* -Stat.	42.62	42.62	42.62	35.68	35.68	35.68
Hansen *p* -value	0.801	0.679	0.750	0.212	0.763	0.210
Most Plausible Potential Instruments Included in "Z"						
Lewbel IV	-0.0321**	-0.0019	-0.0017	0.0025	-0.0007	0.0008
	(0.0160)	(0.0013)	(0.0011)	(0.0144)	(0.0010)	(0.0011)
First-Stage *F* -Stat.	59.49	59.49	59.49	51.41	51.41	51.41
Hansen *p* -value	0.458	0.672	0.339	0.572	0.723	0.208

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. *F*- statistics are Cragg-Donald. Exercise is in minutes per day. "Most Plausible" *Z* variables are the concentration of fitness centers, jobs in sports instruction establishments, and average weather variables. Lewbel IV models are estimated using Fuller modified

Table 4. The Effects of Biking and Walking on Body Mass

	Women			Men		
	BMI	Overweight	Obese	BMI	Overweight	Obese
OLS	-0.0246***	-0.0006	-0.0011***	-0.0282***	-0.0021***	-0.0019***
	(0.0053)	(0.0005)	(0.0004)	(0.0049)	(0.0005)	(0.0004)
All Exogenous Variables Included in "Z"						
Lewbel IV	-0.0226***	-0.0006	-0.0009**	-0.0182**	-0.0013	-0.0012**
	(0.0073)	(0.0007)	(0.0005)	(0.0088)	(0.0008)	(0.0006)
First-Stage *F* -Stat.	599.7	599.7	599.7	216.2	216.2	216.2
Hansen *p* -value	0.192	0.377	0.835	0.076	0.008	0.236
All MSA Characteristics Included in "Z"						
Lewbel IV	-0.0271***	-0.0009	-0.0014***	-0.0333***	-0.0035***	-0.0022***
	(0.0086)	(0.0008)	(0.0005)	(0.0109)	(0.0010)	(0.0007)
First-Stage *F* -Stat.	687.2	687.2	687.2	245.2	245.2	245.2
Hansen *p* -value	0.618	0.825	0.603	0.283	0.108	0.394
Most Plausible Potential Instruments Included in "Z"						
Lewbel IV	-0.0342**	-0.0012	-0.0016	-0.0500**	-0.0054***	-0.0019
	(0.0157)	(0.0015)	(0.0010)	(0.0204)	(0.0021)	(0.0015)
First-Stage *F* -Stat.	412.8	412.8	412.8	226.2	226.2	226.2
Hansen *p* -value	0.333	0.614	0.334	0.470	0.580	0.611

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. *F*- statistics are Cragg-Donald. Biking and walking are in minutes per day. "Most Plausible" *Z* variables are average temperature and rainfall variables. Lewbel IV models are estimated using Fuller modified LIML.