November 12, 2008

Peter Meyer, Director
NCHS Research Data Center
3311 Toledo Road, Suite 4113
Hyattsville, MD 20782

To Whom It May Concern:

This letter is to request permission to use confidential-use data from the National Health Interview Survey (NHIS) through the NCHS Research Data Center. The data to which access is requested is necessary to conduct a research project that studies the effects of female labor force participation on adult and childhood obesity.

Attached you will find the research proposal including abstract, my curriculum vitae, a detailed summary of proposed research, a complete list of data requested, user supplied data, software requirements and shell tables for the report.

Please feel free to contact me at 1-333-123-4567 or irsearcher@state.edu if you have any questions.

Sincerely,

Ima Researcher, Assistant Professor
Department of Economics
State University
Street,
City, State Zip

B. PROJECT TITLE
"The Effects of Female Labor Force Participation on Adult and Childhood Obesity"

C. ABSTRACT

Obesity is a major public health problem in the U.S. Obesity rates in the last 20 years have more than doubled for adults, and more than tripled for children. Understanding the causes behind the increase in obesity rates is fundamental for devising policies aimed at stopping the increase (and eventually decreasing) those rates.

The objective of this project is to assess whether a causal relationship exists between increases in female labor force participation and increases in adult and childhood obesity. We exploit a "natural experiment", the expansion of the Earned Income Tax Credit. This change in benefits affected differentially the labor supply decisions of families with children versus childless families and families with two or more children versus one child families. This allows us to identify the effects of increased female labor supply on the obesity rates of both mothers and children.

D. FULL PERSONAL IDENTIFICATION AND INSTITUTIONAL AFFILIATION

Ima Researcher is an assistant professor of Economics at State University. Dr. Researcher is a labor economist specialized in the analysis of public policies for the low income population and on econometric methods for program evaluation. Her recent research has focused primarily on the determinants and effects of governmental manpower training policies, on the intergenerational effects of welfare dependency, and on enhancements to non-experimental methods for estimating treatment effects. She has papers recently published or accepted for publication in The Review of Economics and Statistics, and in Biometrika. Her contact information is:

Ima Researcher
Department of Economics
State University
P.O. Box 1234567
City, State 99999-1234
Phone: 1-333-123-4567
Fax: 1-333-123-5678
Email: iresearcher@state.edu

Research Team

Scientist 2, is an associate professor of Economics at State University. He is an expert in macroeconomic theory. He has studied how optimal behavior regarding consumption and saving decisions of households is affected by changes in government policies. He has published several articles in scholarly refereed journals such as Journal of Economic Theory, Economic Theory, Journal of Economic Dynamics and Control,

Macroeconomic Dynamics and Journal of Economics and Finance. His contact information is:

Scientist 2
(contact information)

Scientist 3
(contact information)

Doctoral Student
(contact information)

Programmer
(contact information)

E. DATES OF PROPOSED TENURE AT RDC

Our funding covers the period July 2008 to December 2009. The exact dates in which we will be at the RDC are difficult to pinpoint exactly, given that we will need to wait until the appropriate data is created at the RDC, plus we will work in drafts of our research project, to receive feedback on our results. We anticipate probably three one-week trips to the RDC if our proposal is approved. The first one would be in December of 2008 to get the initial results. A second trip would occur at some point during the spring of 2009, and a third trip would occur during the summer of 2009.

F. SOURCE OF FUNDING

A grant from the Institute for Research on Poverty (IRP) under the IRP-USDA RIDGE Program will provide the funding for all the costs associated with using the services at the RDC as well as the travel expenses associated with the project.

G. BACKGROUND STUDY

The objective of this project is to assess whether a causal relationship exists between increases in female labor force participation and increases in adult and childhood obesity. In particular, we will concentrate on one of the most vulnerable groups, low education mothers and their children.

There are several potential explanations for the rapid increase in the weight of the American population observed over the past three decades. One hypothesis is that increasing female labor force participation is related to rising obesity through changes in time allocation and food consumption. Chou, Grossman and Saffer (2003) suggest that women devote more time to work and less to food preparation, increasing their reliance on convenient food and fast food (which is high in caloric content). This affects children

also, as shown by Anderson, Butcher and Levine (2003), who find that a child is more likely to be overweight the higher the hours per week worked by her mother over the child's life. Furthermore, Anderson, Butcher and Schanzenbach (2007) document that the correlation in weight outcomes between parents and children has increased since the early 1970s, so that as much as 40% of the increase in children's weight can be explained by changes in parent's weight.

An alternative hypothesis is that technological change caused food prices to decrease, and also transformed the type of work people perform, from physically demanding to sedentary jobs. Lower food prices increase the consumption of food, which translates into higher calorie consumption, while the increase in sedentary jobs implies a lowering in calories spent (Lakdawalla, Philipson and Bhattacharya, 2005; Philipson and Posner, 2003). Technological change could also operate through the decline in the time cost of prepared foods. Cutler, Glaeser, and Shapiro (2003), suggest that the lower time cost of prepared foods is behind the decline in cooking times and home meals, and behind the increase in the consumption of prepared food observed in the data.

Using U.S. aggregate data Gomis-Porqueras and Peralta-Alva (2007) study the implications of the decline in both the monetary and the time (relative) cost of prepared foods on adult calorie intake. In their framework, the time channel operates by declines in income taxes and the gender wage gap which increase female labor supply and the opportunity cost of cooking at home, thus decreasing the time spent cooking at home. Their results suggest that up to two thirds of the increase in the consumption of calories of the average adult can be explained by this channel.

In this project we will study whether the aggregate relationship between female labor force participation and obesity is confirmed by the individual-level data. The challenge is to identify the causal effect of labor force participation on obesity. Previous studies have illustrated a correlation between the two variables for adults (see for example Bleich, Cutler, Murray, and Adams, 2007), and for small children of high socioeconomic status mothers Anderson, Butcher and Levine (2003) established a causal effect of hours worked on childhood obesity. In our project we will rely on changes in the Earned Income Tax Credit (EITC) in the 1980s and 1990s as a source of exogenous variation in female labor participation of low education single (and married) mothers, to identify the effects of participation on both adult obesity and childhood obesity.

Changes in the EITC implied that the maximum benefit increased in real dollars from 1986 to 1987 by 50%, from 1990 to 1991 by 25%, from 1993 to 1994 by 63%, and from 1994 to 1995 by 20%. More importantly, the changes affected differentially the incentives of taxpayers with different number of children.

Starting in 1991 the EITC credit has been different for one-child taxpayers versus two-or-more¬children taxpayers. The difference in the credit was very small up to 1994, when the difference increased 25% in favor of taxpayers with more than one child. Starting in the same year, childless taxpayers became eligible for a small credit (in the order of $300 maximum).

To understand, and quantify, the causal effects of changes in female labor force participation on obesity we will exploit expansions in the Earned Income Tax Credit (EITC) in 1987 and 1993, which have been credited with increasing the labor force participation of low education single mothers. This provides us with a credible empirical strategy to study the effects of labor force participation on obesity, because by making comparisons between groups and across time we will be able to control for other confounding factors that might be related to changes in obesity. The empirical analyses will use more than two decades of information from the National Health Interview Survey (NHIS).

Because the EITC can be received only by working taxpayers (primarily with children), it generates changes in the incentives to work for low income individuals. In particular for single parents (mostly women), there should be unambiguous incentives to increase labor force participation. Working mothers must trade off the advantages of greater income against the disadvantages of less time for home food production and supervision of children's activities. The main hypothesis of this project is that the increased labor force participation by single mothers caused the increase in weight problems for both the working mothers and their children. These women devote more time to work and less to food preparation, increasing their reliance on convenient food and fast food (which is high in caloric content), thus consuming more calories.

In addition, we will explore the effects of the expansion of the EITC on married women, although the employment effects in that case can be of ambiguous sign.

Policy Relevance

Given the sharp increase in adult and childhood obesity, research like the one in this project is needed to better understand why increases in obesity rates have occurred. This is an important first step for finding solutions to the nation's health problem. In particular, in order to promote specific behavioral changes, we need to understand the motivation behind these behaviors. This is key for developing policies that promote behavioral changes to counteract the sharp increases in adult and childhood obesity of recent decades. In particular, obesity rates among low income women and their children are higher than for any other group. If labor force participation can explain an important part of the changes in weight problems for those groups, then the results from this research can help tailor intervention campaigns as well as determine how to efficiently allocate funds to reduce obesity rates. For example, if women are more likely to eat foods prepared away from home in order to be able to work, then it is important to provide information and guidance on how to eat healthily.

H. DATA REQUIREMENTS

Cases to be included
We will use for our analyses data from the National Health Interview Survey (NHIS) from 1982 to 2005. We require such an extensive number of years because we think it is

important to have enough information before and after the EITC changes in 1987 and 1993. Starting in 1982 provides us with several years for the baseline before any tax changes. Conducting the analyses up to 2005 (even though in some specifications we may only run regressions up to 2000) offers us the opportunity of studying the evolution of take-up rates in the program, which previous studies have found to be increasing over time.

We have constructed already an analysis file based on the publicly available data for the NHIS. That analysis file has all the individuals in the NHIS pooled from 1982 to 2005. As the definitions of the different variables of interest change across years of the NHIS, we put our emphasis in making sure that all the variables of our interest are comparable across time. We have also merged family-level information into our analysis file, according to each individual's family identifier. This file, before dropping observations we will not use in our analyses, has 2,469,650 observations and 167 variables, with a total size of approximately 750mb (in Stata 9 format).

Our analyses will be different for adults and children (where children are defined as those individuals with age less or equal to 18 in the calendar year of the survey). For adults we will consider only individuals in the ages 20 to 50 years old. This gives us a total of 1,092,150 adults. In addition, we will impose the condition that all adults considered are not disabled and are not students. Furthermore, we will drop from our analysis file all families for which the family structure cannot be determined appropriately, in particular the number of children (and who is likely to be the parent or guarding of those children), given that the number of children is the key variable associated to the amount of EITC benefits. These constraints imply that our final analysis file will have in the order of 1,000,000 observations for adults (once we account for missing values for several of the key variables in our analyses the final number may be a lit bit less than that).

For certain regressions we will use data on children of ages 12 to 18. We have identified approximately 200,000 observations in the NHIS for those ages, with valid information on our variables of interest. Therefore, in total, our analysis file will have approximately 1.2 million observations, including adults of ages 20 to 50 and children of ages 12 to 18.

Variables to be included

We have put special attention in making as comparable as possible the variables in different years of the NHIS. Of course the biggest problem is the switch from the 1982-1996 period to the 1997-2005 period, where the definitions of many of the variables suffer many changes. In addition, many of the variables suffer minor changes even within these two periods. Still, we believe we have done a good job in keeping consistency across time. From the individual files in different years we have created a series of variables: Demographic variables (age, race, civil status and gender); education variables (we classified education in three groups as: less than a high school degree, high school degree, at least some college; also we identify current students); whether the person is foreign born (only available since 1988); dummies for region and MSA size (which we

will not use once we have access to the restricted data geographic information); and a variable indicating whether the individual is able to work or not.

The issue of comparability is even more important for the key dependent variables for our study. First, we created an employment indicator, for which we had to carefully analyze the many changes in the employment-related variables in particular in 1997 and thereafter. We believe we have achieved the right definitions, and we do not see any unusual changes in the percentage of adults employed in the years when the survey instruments changed. Second, we used height and weight of the individuals (from 1982 to 1996), and the provided BMI variable (from 1997 to 2005), to create a variable indicating Body Mass Index in the whole analyzed period. We restrict our analyses to BMI values between 15 and 50, which eliminates well below 1% of the cases and deals well with outliers. We created indicators for whether a person is overweight (BMI≥25) and whether a person is obese (BMI≥30), based on the BMI variable.

Restricted-use data requested and why it is necessary

There are three reasons why we need to access restricted-use data from the NHIS at the Research Data Center. First, for all our analyses, we think it is necessary to control as well as possible for differences in labor market conditions, local regulations, local services, local availability of prepared-food establishments, etc. Systematic geographic differences in any of these factors could be a confounder when trying to identify the effects of EITC changes on employment and obesity. One way to deal with this issue is to run regressions that include fixed effects at some geographic level. Our interest would be to run regressions with county fixed effects. This can only be done by accessing the restricted-use version of the NHIS. Note that, of course, we do not need to know exactly in which county each individual is located, we only need a variable that identifies people belonging to the same county (for example a "pseudo" FIPS variable, or similar). In this way, the county fixed effects will control for any factors that are county-specific and are related to the employment and calorie-consumption decisions of the individuals, and that do not change over time.

We believe that the county fixed effects regressions will be very helpful, but still are worried that these fixed effects may not be enough. In particular, there are many factors that could be changing over time which we would like to control for. We will include in our regressions year fixed effects, but those only deal with national-level changes over time. In addition, we would like to control for a variety of factors that are time-varying, but for which the timing of the variation can be very different across counties. For this, we would like to merge into the NHIS analysis file, publicly available data at the county (or state) level that may be very important in eliminating omitted-variable biases. In this way, we would bring to the RDC an external dataset at the county-year level composed of a variety of publicly available data. We plan to include measures of local labor market conditions (like average employment and earnings in different sectors, from the Census Bureau's Covered Employment and Wages, CEW, data), measures of the average generosity of the welfare state program and tax rates at the state level. In addition we will

try to obtain variables that measure the density of fast-food establishments at either the county or state level. We are in the process of obtaining this last variable.

In addition to being able to run county-fixed effects regressions and to add county and state level time-varying variables in the analyses, our last reason to request access to the restricted-use version of the NHIS is that we want to analyze data on obesity for children, not only for adults. However, weight and height (and thus BMI) information for children is not available in the public-use version of the NHIS. Through an email exchange with Dr. Eve Powell-Griner at the Division of Health Interview Statistics we have learned that there are great concerns about the quality of the BMI information for children, and that this is why it has not been included in the public-use versions of NHIS. The recommendation of Dr. Powell-Griner is that if we use BMI data for children at all we should only use it for teenagers of age 12 and above. Dr. Powell-Griner explicitly mentioned that she thought that even the data for children 12 and up may not be of good enough quality. We agree that we will need to thread carefully when using this data but we still believe that our methodology can deal with the potential problems with the BMI information for children. First, we will use CDC's guidelines to classify children as overweight and obese, according to CDC's percentile cutoffs for their age and gender (85th percentile for overweight, and 95th percentile for obese or "at risk"). So, this should decrease (at least a bit) concerns around BMI values being subject to measurement error. Second, because the coefficients of interest in our regressions will be those coming from comparing children in families with one child versus families with two or more children, any measurement error would cause a real problem only if the degree of measurement error differs for different family sizes. We think that it is reasonable to assume that any measurement error problems are probably not related to family size, nor are related to the timing of the changes in the EITC. Therefore, even though we will keep in mind the concerns of Dr. Powell-Griner, we believe that the BMI information for children will still be valuable for our project.

In summary, we plan to bring to the RDC our analysis file of pooled observations across years of the NHIS, including of course all identifiers, and a dataset at the county-year-level (or state-year-level when appropriate), so it can be merged by FIPS by the RDC staff with our analysis file. In addition, we would like to have a "pseudo" FIPS identifier that allows us to run county fixed effects regressions. Finally, for the children 12-18 in our analysis file, we would like to obtain their BMI information (or height and weight, if that simplifies the process).

Reasons to Keep the Data for More than a Year

In Economics the average time from submission to publication of a paper is 3 years (and it is not uncommon that it takes much longer). Given this lengthy process and that reviewers typically ask for revisions, we feel it is crucial to request that the analysis files created by the RDC staff for our project not to be erased until our research has been published. By keeping the files longer we will avoid incurring the costs of creating everything again for a subsequent revision.

I. METHODS FOR THE STUDY

Analytic strategy and statistical methods

This research project follows the strategy of Hotz, Mullin and Scholz (2005) of comparing the changes in labor force participation and obesity rates of single mothers with one child, versus single mothers with two or more children, before and after the 1994 EITC expansion for families with more than one child. In addition, we follow the strategy of Meyer and Rosenbaum (2001) of comparing single women with children versus single women without kids, before and after the 1987 EITC expansion. If the results show that the effects on labor supply are similar to the ones found in the literature, it will validate the application of this methodology to the study of changes in obesity rates of adults and children.

Similar analyses can be performed for the evolution of Body Mass Index (BMI), rates of overweight and obesity, for adults and children. So, differences-in-differences regressions of the following form will be examined:

$$Yi = \beta 0 + \beta 1\ Xi + \beta 2\ Dkids + \sum \gamma t\ Dyear\_t + \sum \delta t\ (Dyear\_t * Dkids) + \eta c + ui, tt$$

where i refers to an individual; the variable Yi will represent alternatively an indicator variable for whether the person is employed, the BMI of the person, an indicator for whether the person is overweight, or an indicator for whether the person is obese. The vector Xi will include a variety of individual covariates like age, race and ethnicity, number of children of different ages, education levels, plus a series of county- and state-level variables. The variable Dkids will be an indicator for whether the family of person i is a two-or-more children family (for the specifications where the comparison group will be one-child families), or it will be an indicator for whether the family of person i has any children (for the cases where the comparison group will be women with no children). The t variables Dyear_t represent indicator variables for the year in which observation i is observed (of course there will one year omitted from the regressions), and ηc represent the county fixed effects. The coefficients of interest here are the δt coefficients associated to the interaction terms between the kids dummy and the year dummies. These coefficients identify the differences in the average dependent variable for the two groups identified by the variable Dkids. If the effects of the EITC expansion are as expected, those coefficients should be positive and statistically significant for all the dependent variables, only in the years after the corresponding EITC expansion. That is, after 1987 for the expansion that affected all families with children (when Dkids represent whether the woman has children or not), and after 1993 for the expansion that affected differentially families with one child versus families with more than one child. Of course, for studying childhood obesity we will be able to rely only on the second source of variation, so for children outcomes, only the 1994 expansion will be relevant. The logic for the expected positive signs is that EITC differentially increases the labor force participation of the women in the "treatment" group, which also increases their BMI, and the prevalence of overweight and obesity. This affects both adults and children.

With our proposed methodology we can to assess whether a causal relationship exists between increases in female labor force participation and increases in adult and childhood obesity. In particular, we exploit a "natural experiment", the expansion of the Earned Income Tax Credit. An important advantage of our approach is that by comparing types of families at the same point in time, we are able to hold constant confounding factors (like price changes and technological changes) that have been proposed as explanatory for the increase in obesity. In that sense, unless these factors differentially affect families with different number of children, our strategy presents a clean way of understanding the effects of female labor force participation on obesity.1 Finally, the key assumption for the differences-in-differences method to be valid is that the "treatment" and "comparison" groups have not changed in composition over time. Comparisons of one-child versus two-or-more-children families should be less subject to this concern.

Software requirements

The software used for all the analyses will be Stata. All the datasets provided by the researchers (the constructed analysis file based on the publicly available NHIS data, and the additional data) will be in Stata format.

Sampling weights and standard errors

All summary tables and regressions will use sampling weights (the type called "pweights" or probability weights in Stata), that take into account the probability of being selected into the sample for each observation. The natural weight variable to use is "wtfa", the final weight for each unit in the individual-level files of the NHIS. However, there is a complication introduced by the fact that starting in 1997 the BMI variable is only available for individuals in the "sample adults" file. Therefore, when working with BMI and the overweight and obesity indicators, the appropriate weight is "wtfa_sa", the final weight for individuals in the "sample adults" file of the NHIS. Because "wtfa_sa" is available only since year 1997 on, a solution to weight appropriately the observations, is to use "wtfa" for observations prior to 1997 and "wtfa_sa" for observations on or after 1997. We constructed such variable, calling it "weight_mixed".

Footnote:
1 Technological changes in home food production have the potential of affecting differentially families of different size. However, we believe that at least in the case of one versus two or more children, these effects should be a minor concern.

When running the employment regressions, this is not an issue, because we use the employment information in the "persons" files that contains all the individuals, and "wtfa" is available for all the observations. However, to make the employment and obesity regressions completely comparable, we want to select the same sample, which implies using starting in 1997 only the individuals in the sample adults file and weighting by "weight_mixed", even when using the employment variable as a dependent variable.

Indeed, we have run regressions already where we use the employment variable as dependent variable and we alternatively use "wtfa" as weight variable, or "weight_mixed" as weight variable (and on the smaller sample) and the results of the regressions are very similar. This implies that not much of a problem is caused by going from using all (valid) individuals in the persons file, versus using (valid) individuals in the sample adults file starting in 1997 (which was expected, given that the adults in the sample-adults file where selected at random).

In addition in all the regressions we will calculate heteroskedastic-robust standard errors (using what is known also as the "White-Huber" or "Sandwich" estimator). We are aware that the NHIS follows a complex sampling design, which would be ideal to take into account in the estimation of standard errors (using the "svy" or the "cluster" commands in Stata, for example). However, given that we are analyzing NHIS surveys over such a long period of time it becomes impossible to do so, taking into account the changes in the definitions of PSUs over the different sampling designs of the survey. We believe our approach of sampling-probability weighting and calculating heteroskedastic-robust standard errors is a good solution to take into account the sampling design.

J. TABLE SHELLS, EQUATIONS, AND TEST STATISTICS

We anticipate that disclosure risk analysis will be easy under this project. We only plan to produce some summary statistics tables (where each cell will be big enough), and regression results, where we believe there will not be any disclosure risks. We have done some preliminary analyses, and even when looking at the smallest groups in which we are interested in, by year, the smallest cell size is close to 100 observations.

The desired output will consist of the following tables:

Table 1: Summary statistics providing the mean and standard deviations of all the variables used in the regressions.

Table 2: Summary statistics describing the employment rates for women ages 19-50 by education, number of children and marital status, by year.

Table 3: Summary statistics describing the average BMI, overweight and obesity rates for women ages 19-50 by education, number of children and marital status, by year.

Table 4: The same as Table 3 but for the children ages 12-18 of the women included in Table 3.

Table 5 (and above): Several tables with the results from running the regressions. The focus will be on the coefficients associated to the year dummies*family size interactions.

References
(Use any standard bibliographic or reference format)

K. APPENDICES

1. Curriculum Vitae for each person who will participate in the research team

Sample CV (Use any standard biographical format).

Ima Scientist
Curriculum Vitae

November 12, 2008

CONTACT INFORMATION
Address, Phone number, Fax number, email address, website.

PERSONAL INFORMATION

EDUCATION

AREAS OF SPECIALIZATION

PROFESSIONAL EXPERIENCE

AWARDS

GRANTS

CONFERENCE PRESENTATIONS

SEMINAR PRESENTATIONS

JOURNAL REFEREE

MEMBERSHIP IN PROFESSIONAL ORGANIZATIONS

PUBLICATIONS

Curriculum Vitae for Second researcher.

Curriculum Vitae for Third researcher.

Curriculum Vitae for programmer.

Etc.

2. Data dictionary for analysis file constructed using publicly available NHIS data from 1982 to 2005.

[RDC Note to Applicants:

The combined codebook shown below, is what the researcher submitted. While it is a good start, this kind of summary list is not the format we wish you to use. Because the variable names may change over the years, you should send us a codebook of the public use variables you will supply for each year, or survey cycle, for each data file you will use. You will need this type of codebook to create the data files anyway, so it is helpful to you and to us for you to do this at the research proposal stage.]

As explained above, an analysis file was created pooling the publicly available NHIS data from 1982 to 2005. Without dropping any observations due to sample selection the analysis file contains 2,469,650 observations and 167 variables. The size of the file, in Stata 9 format, is almost 750mb. However, we anticipate that the analysis file that we will take to the RDC will be substantially smaller (in the order of 300mb), because due to sample selection we will have only around 1.2 million observations (1 million adults 20-50 years old, and approximately 200,000 children 12-18 years old). Below is the list of variables in our analysis file.

Variables used to identify individuals and families:

Yr:	year survey was taken
Psunumr:	available from 1982-1994. Pseudo primary sampling unit. Used to identify households. String.
Segnum:	available from 1982-1994. Segment number. Used to identify households. String.
Hhnum:	available from 1982-1994. Household number. Used to identify households. String.
Pnum:	available from 1982-1996. Person number. Used to identify people. String.
Famnum:	available from 1982-1994. Family number. Used to identify families. String.
Quarter1:	available from 1982-1994. Processing quarter. Used to identify households. String.
Weekcen1:	available from 1982-1994. Processing week code. Used to identify households. String.
Householdid:	available for 1982-1994. Identifies households. Generated by us. String.

Personalid: available from 1982-1994. Identifies people. Generated by us. String.

Familyid: available from 1982-2005. Identifies families. generated by us. String.

Familyid1: same as familyid, but with the year attached to the end. String.
Hhid: available in years 1995-1996. Identifies households. String.
Famtype: available for years 1995-1996. Family type. Used to identify families.
Hhx: available for years 1997-2005. Household number. Used to identify individual households. String.
Fmx: available for years 1997-2005. Family number. Used with fmx to identify individual families. String.
Px: available for years 1997-2003. Person number. Used with fmx and hhx to identify people. String.
Fpx: available for years 2004-2005. Person number. String.

Response/sample variables:
Respond: Indicates whether respondent answered for himself or by proxy. This variable is available for years 1982-1996.
Respond1: Constructed by us. It's a dummy, 1 if answered for him/herself. otherwise (=1 for individuals in sample adult/child files).

Sampadult: Dummy, 1 if person is in the sample adult file (1997-2005), 0 otherwise
Sampchild: Dummy, 1 if person is in the sample child file (1997-2005), 0 otherwise

Sampling weight variables:
Wtfa: Weight for people in Person's file.
Wtfa_sa: Weight for people in sample adult file (1997-2005)
Wtfa_sc: Weight for people in sample child file (1997-2005)
Weight_mixed: This variable is = wtfa for 1982-1996 and = wtfa_sa for 1997-2005.

Demographic variables:
Age: In years
Agesq: Age squared
Maledum: Dummy if person is male
Married: 1=married. 2=single. 3=divorced
Dummarried: Dummy for whether person is married.
Race1: 1 if white, 2 if Hispanic, 3 if black, 4 if other.
White: Dummy for race=White
Hispanic Dummy for race=Hispanic
Black: Dummy for race=Black
Otherrace: Dummy for race=other
Foreignborn: Dummy – 1 if born outside of the US (only available for years 1989-2005; for years 1982-1988 foreignborn=0 for everyone).
Foreignmissing: Dummy, 1 of foreignborn is missing (1982-1988). Note that foreignborn=0 for years 1982-1988.
Yrsinus: Number of years lived in the US. Available for years 1989-1996.

Geographic variables:

Region: Northeast, Midwest, South, West
Msasize1: Metropolitan Size. 1=more than one million habitants. 2=250000¬999,999 habitants. 3=under 250,000 habitants.
Smallcity: Dummy for city size (person lives in city under 250,000 habitants)
Mediumcity: Dummy for city size (person lives in city 250,000 -999,999 habitants)
Largecity: Dummy for city size (person lives in city with more than 1 million habitants) Previous three dummy variables valid only for years 1995-2001.

Employment variables:

Employ1: Dummy, 1 if employed.
Noablework: Dummy, 1 if unable to work.

Education variables:

Student: Dummy – 1 if a student.
Educ1: Education. 1=less than high school. 2= high school degree. 3=some college and more.
Nohsdegree: Dummy, 1 if person is no high school degree
Hsdegree: Dummy, 1 if person has high school degree (and no more)
Somecollege: Dummy, 1 if person has attended college (or more)

Height/weight/BMI variables:

Obese: 1=person is obese, 2=person is overweight, 3=person's weight is normal, 4=person is underweight. Obesedummy: Dummy, 1 if person is obese, 0 otherwise. Overweight: Dummy, 1 if person is obese or overweight.
Height: in inches.
Weight: in pounds
BMI: Body Mass Index

Family structure variables:

Nkids_0: Number of infants in the family
Nkids_15: Number of kids between ages 1 and 5 in the family
Nkids_612: Number of kids between ages 6 and 12 in the family
Nkids_1318: Number of kids between ages 13 and 18 in the family
Nkids_tot: Total number of kids in the family
Dum_nkid0: Dummy, 1 if there are kids in the person's family, 0 otherwise.
Dum_nkid1: Dummy, 1 if there is one kid in the person's family, 0 otherwise.
Dum_nkid2plus: Dummy, 1 if the person's family has 2 or more kids, 0 otherwise.
Dum_nkid3plus: Dummy, 1 if the person's family has 3 kids or more.
Dum _nkid2: Dummy, 1 if person's family has exactly 2 kids.
Adultchild: Number of adult children in the family. Cutoff: greater or equal to 19 years old.
Kidchild: Number of kid children in the family. Cutoff: less than 19 years old.
Grandchild: Number of grandchildren in the family.

Other_adult: Number of "other" adults in the family.
Other_child: Number of "other" children in the family.
Ref_standardized: 1 =person is the reference person, 2=person is the spouse of reference person, 3=person is an adult child (>=19) of reference person, 4= person is a kid child (<19) of reference person, 5=person is grandchild of reference person, 6=person is an "other adult", 7=person is "other child"
Ref: Dummy, 1 if there is a reference person in the household. 0 if no one in the household is listed as a reference person.
Spouse: Dummy, 1 if there is a spouse of the reference person in the family. 0 if no one is listed as a spouse of the reference person. Dk: Number of people for whom their relationship to the reference person is unknown.
Other: Number of "other" people in the family.
Other_adult_dum: Dummy, 1 if there is an "other adult" in the family, 0 otherwise.
Other_child_dum: Dummy, 1 if there is an "other_child" in the family, 0 otherwise.
Age_difference_issue: Dummy, 1 if age difference between the oldest child and youngest grandchild is less than 14 years, 0 otherwise.
Fam_structure: 1=Family head is single, 2=Family head is living with spouse
Valid_fam_str: 1=family head is single, there is no age difference issue, and no "other child" in the family. 2= family head is with spouse, there is no age difference issue, and no "other child in the family. 3=everybody else.

Other:
Dum1982-Dum2005: Year dummy variables.
Int1982-Int2005: Interaction dummies; year dummies*dum_nkid2plus

Household head variables:
Information for the head of household. The definitions are the same as above, the variables just have the prefix "head": head_employ1, head_bmi, head_obesedummy, head_overweight, head_white, head_hispanic, head_black, head_otherrace, head_dummarried, head_nohsdegree, head_hsdegree, head_somecollege, head_age, head_maledum, head_foreignborn.

Spouse variables:
Information for the spouse in the household. The definitions are the same as above, the variables just have the prefix "spouse": spouse_employ1, Spouse_bmi, Spouse_obesedummy, Spouse_overweight, Spouse_white, Spouse_hispanic, Spouse_black, Spouse_otherrace, Spouse_dummarried, Spouse_nohsdegree, Spouse_hsdegree, Spouse_somecollege, Spouse_age, Spouse_maledum, Spouse_foreignborn.

3. Data dictionary for researcher-supplied data

We have not constructed this data yet, we are in the process of gathering the data, so there could be some small changes. The data will be organized with one observation per county, per year, from 1982 to 2005. A FIPS identifier will be included, as well as a year identifier. The variables that we will include are all publicly available data. We will include the following variables.

Local labor market conditions (Census Bureau's Covered Employment and Wages data):
County Level:
-Employment/Population ratio
-Average real earnings
-Employment/Population ratio by sector (manufacturing, wholesale trade, retail trade, construction, services)
-Average real earnings by sector (manufacturing, wholesale trade, retail trade, construction, services)
State Level:
-Unemployment rate (BLS does not have a series on unemployment rates by county that are consistent for the whole period, only since 1990).

Welfare policy variables (DCF at DHHS data):
This data is available at the State level:
-Maximum expected welfare (AFDC/TANF) real benefits
-Maximum benefit if a person works

Income tax data (IRS):
This data is available at the State level:
-Average income tax rate
-Highest marginal income tax rate

Density of full-service and fast-food establishments (Census Bureau):
This data is easily available at the State level, but we are trying to get this information at the county level.