

An Assessment of the National  
Representativeness of Children in the 2007  
Panel Study of Income Dynamics

Denise Duffy  
Narayan Sastry

Panel Study of Income Dynamics  
Survey Research Center, Institute for Social Research  
November 2012  
(Revised)

This project was supported by funding from the *Eunice Kennedy Shriver* National Institute for Child Health and Human Development (Grants HD069609 and HD033474).

## **Abstract**

There are many ways that nationally-representative samples drawn for panel surveys can become non-representative over time. For example, selective attrition and non-response, immigration, and shortcomings in the initial sample may lead to a divergence of the panel sample from the corresponding population it seeks to represent. Although initial cross-sectional weights together with attrition weights in later waves may correct for some sources of non-representativeness, they do not correct for all possible sources. Furthermore, a periodic reassessment of sample representativeness is important for understanding the cumulative effects of attrition and immigration and determining whether additional measures are necessary to correct for declining representativeness and coverage.

In this technical report, we assess the national representativeness of the 2007 sample of children in the Panel Study of Income Dynamics (PSID)—a total of approximately 7,100 children aged 0–17 years of age. We evaluated the national coverage of the PSID child sample using microdata from the American Community Survey (ACS) and a new statistical technique called generalized boosted regression models. The ACS provides a gold-standard comparison based on its extremely high (98%) response rate, excellent data quality and completeness, and large sample sizes (approximately 700,000 children aged 0–17 years in 2007). We constructed a reasonably consistent set of covariates across PSID and ACS to describe children on the basis of their age, race/ethnicity, sex, poverty status, geographic region, and having a foreign-born parent. The generalized boosted models provided flexible, non-parametric estimates for assessing the relationship between our dependent variable (an indicator of whether an observation in the pooled ACS-PSID sample came from PSID) on covariates and their interactions. We also used the ACS sample to identify cases that are logically excluded from PSID (specifically, children whose parents are both post-1997 immigrants to the U.S.), and assess the size and composition of this group.

Our results reveal that PSID child sample provides good representation of the corresponding national population with coverage of approximately 97% of the U.S. population of children in 2007 and reasonable balance for most groups, although with some exceptions.

## Introduction

Sample selectivity can challenge the unbiasedness and generalizability of results and conclusions obtained from analyses of survey data. Even with high baseline and wave-to-wave response rates—and the use of well-designed weights—selectivity may emerge in panel studies through the cumulative effects of non-response and attrition. In addition, on-going studies may lose representation due to the effects of in-migration to the study area if there is not a mechanism for in-migrants to be regularly added to the sample. A recent study by Clark et al. (2009) highlights the significance of immigrants for studies of children and families. The share of children from immigrant families increased from 6% in 1960 to 24% 2007. Among certain racial and ethnic groups, the percentage of children from immigrant is much higher: in 2007, 89% of children in Asian and Pacific Islander families and 64% in Hispanic families had one or more immigrant parents (Clark et al., 2009). Panel studies that do not incorporate new in-migrants may omit a significant segment of the population. Over time, attrition and new immigration may together lead to a divergence of a study sample from the corresponding population it seeks to represent. An assessment of a survey's representativeness is hence likely to have considerable value for data users and for consumers of study results, and may also help those designing and running the survey to understand and address its limitations.

In this technical report, we assess the national representativeness of the 2007 sample of children included in the Panel Study of Income Dynamics (PSID) using data from an independent gold-standard cross-sectional survey in the U.S.—the American Community Survey (ACS)—together with a new statistical approach known as generalized boosted models. PSID is a nationally-representative panel of U.S. families that began in 1968 and had, by 2007, collected data on the same families and their descendents for 35 waves over 39 years. In 2007, the PSID comprised of approximately 8,500 family units with a total of 24,000 individual family members—including 7,100 children aged 0–17 years of age. The survey is the world's longest-running household panel survey and is widely used, with more than 3,200 peer-reviewed

publications based on the data (McGonagle et al., 2012). The ACS provides a gold-standard comparison survey based on its extremely high (98%) response rate, excellent data quality and completeness, and large sample sizes (approximately 700,000 children aged 0–17 years in 2007). The goal of this paper is to develop a better understanding of the PSID's child sample properties. We expect that, despite the various challenges to representativeness, the coverage of the PSID child sample is likely to be quite good based on its high response rates, a 1997 immigrant refresher sample, and a mechanism to include new immigrants who marry into or join existing study households.

The PSID has achieved high response rates for most of its history (Schoeni et al., forthcoming), although lower-income families have had higher cumulative attrition (Fitzgerald et al., 1998a and 1998b; Fitzgerald, 2011). Assessments of the representativeness of PSID has focused on analyzing attrition, examining the bias in estimated parameters due to attrition, and comparing weighted estimates from the PSID with those from benchmark national surveys. In a recent study focusing on intergenerational analysis using the PSID, Fitzgerald (2011) found little evidence that cumulative attrition led to biased estimates. Comparisons with national surveys include assessments of weighted PSID estimates for income, earnings, and transfers against the March Current Population Survey (Gouskova et al., 2010; Beckett et al., 1988; Duncan and Hill, 1989), wealth against the Survey of Consumer Finances (Bosworth & Anders, 2008), expenditures against the Consumer Expenditures Survey (Li et al. 2010), health status and behaviors against the National Health Interview Survey (Andreski et al. 2009; Fitzgerald, 2011), and time use against the American Time Use Survey (Cornman et al., 2011). In all of these analyses, the findings indicate that the weighted PSID estimates are closely comparable to the parallel set of estimates and support the conclusion that the weighted PSID data achieves a convincing level of national representativeness. Taken together, these findings provide evidence that the estimates based on the PSID are largely unbiased by the effects of non-response, attrition, and non-coverage.

Essentially all of the previous assessments of the PSID sample representativeness have focused on outcomes for adults. The adult sample comprises of original sample members and, increasingly over time, adult children of the original sample members who enter the sample as primary respondents when they split-off to form their own economically-independent households. In the case of the PSID child sample, however, none of the children are selected directly for the study—in most cases, they are two or even three generations removed from the household that was original sampled for the study. As a consequence, an assessment of the representativeness of the child sample is a useful test of the long-term validity of the PSID's study design.

There have been no direct assessments for the PSID of its coverage error—that is, the extent to which the PSID sample captures all segments of the corresponding population (Groves, 1989). Non-coverage emerges if there are individuals in the target population for whom there is no mechanism for selection into the sample. Surveys typically place some bounds on the sampling frame and define the study's target population accordingly. For instance, individuals who are institutionalized or reside in group quarters are usually excluded from surveys of the general population. For PSID, children born to parents who both immigrated to the U.S. after 1997 are not covered by the sampling frame. However, children for whom one parent was in the country before 1997 but the other parent was a post-1997 immigrant do enter the sample through marriage. A separate source of error is due to non-response at baseline or to attrition. The cumulative effects of these two types of non-response might affect the representation of respondents in particular population segments.

The approach we use to investigate the coverage of children by PSID is similar to the techniques used to assess sampling frames through non-response analyses (see Groves, 2006) and to methods used to analyze and adjust for coverage bias in Internet surveys (Schonlau et al., 2009; Lee and Valiant, 2009), in landline telephone surveys (Blumberg and Luke, 2007; Peytchev, Carley-Baxter, and Black, 2011), and due to study design features such as not

following migrants who leave the study area (Basic and Rendtel, 2007). In each case, there is another survey or sampling frame with better coverage against which the more select study sample is compared. The goal in general is to develop a set of weights that allow the more select survey to produce estimates for the entire target population. Because of this aim, the literature in this area tends to focus on the effects of non-coverage on the bias and variance associated with estimated parameters—typically, means, proportions, and other descriptive statistics—using methods reviewed by Groves (2006). Our focus in this paper is somewhat different; we are interested in understanding the relative and absolute coverage by the PSID of the population of children in the U.S. Our goal is to describe the relationship of the survey to the overall child population and to characterize the size and attributes of the non-target population, which may be the focus of a future refresher sample. This difference in motivation reflects the fact that analyzing non-coverage for the PSID child sample allows us to identify an important segment of the population (children of recent immigrants) that is not captured by the study design, rather than investigating more mundane choices about the type of telephone service or whether to obtain Internet access at home.

The remainder of the paper is organized as follows. In the next section, we describe the data and statistical methods. We then present our results, which include descriptive findings, the regression model results, and an assessment of size and characteristics of children not captured by the study. The final section discusses the results and presents some conclusions and implications of our findings.

## **Data and Methods**

In this section, we provide a brief overview of PSID and describe the American Community Survey. We then identify the specific child and family characteristics that we use in our analyses. Finally, we outline the statistical methods that we use to assess the representativeness of children in PSID using data from ACS.

## Panel Study of Income Dynamics

PSID is a nationally-representative, longitudinal study of families in the U.S. that began in 1968. The original 1968 PSID sample was drawn from two independent samples: a nationally representative sample of 2,930 families designed by the Survey Research Center at the University of Michigan (the “SRC sample”) and an over-sample of 1,872 low-income families from the Survey of Economic Opportunity (the “SEO sample”). The oversampling of families who were poor in the late 1960s resulted in a sizable subsample of African Americans. These two samples combined constitute a national probability sample of U.S. families as of 1968.

PSID has followed all persons living in PSID families in 1968, as well as their birth or adoptive descendants. Adult children become primary respondents when they form their own economically-independent family units. People who married into PSID families after 1968 are treated as non-sample persons: information about them is collected while they are living in PSID families, but they are not tracked and interviewed if they leave a sampled household. However, all children who are born to or adopted by an eligible PSID respondent (regardless of their spouse/partner’s PSID sample status) have information collected about them, and they are tracked even if they leave a sample household.

Based on the original rules for selection and for following respondents and their offspring, PSID would not include, as sample persons, individuals who arrived in the U.S. after 1968. To address this limitation, a representative sample of approximately 500 immigrant families was added to PSID in 1997/1999. This new sample is known as the 1997 PSID immigrant refresher sample and included families who immigrated to the U.S. after 1968 and families of individuals born in 1969 or later to people who were not living in the U.S. in 1968. These recent immigrants were not represented in the PSID sample before 1997—except as non-sample individuals if they married into or joined a PSID sample household. There has not been a PSID immigrant refresher sample since 1997 and, as a consequence, the study’s target population does not include post-1997 immigrants or their children. However, PSID does track

(as non-sample persons) post-1997 immigrants who married into or joined PSID families.

Immigrant families headed by two post-1997 immigrants have no mechanism to be selected into the PSID sample, although the study is considering adding a new immigrant refresher sample to capture households comprised entirely of individuals who have arrived in the U.S. after 1997.

### American Community Survey

ACS is the U.S. Census Bureau's replacement for the long-form in the decennial census (U.S. Census Bureau, 2009). ACS uses a series of monthly national samples that are fielded on a continuous basis. Full implementation of ACS for housing units began in 2005 with group quarters added in 2006. The ACS Public Use Microdata Sample (PUMS) data that we use represents a 1% national sample that comprises of about 1.3 million housing units and about 3 million person records. The ACS PUMS files include a variety of modifications (such as record swapping, top- and bottom-coding, and subsampling of ACS households) to protect respondent confidentiality.

ACS is primarily a mail survey, although there is a telephone follow-up for non-respondents to the mailed questionnaire and an in-person follow-up for non-respondents to the telephone interview. ACS includes 25 housing and 42 population questions, covering topics such as basic demographic characteristics, schooling, employment, disability, commuting, and dwelling characteristics. The ACS questionnaire is generally completed by one household respondent, who is a member of the household at least 18 years of age.

ACS achieves very high response rates because of its status as a government survey in which households are required to participate, the use of multiple modes (mail, telephone, and in-person), and the Census bureau's extensive experience and expertise in data collection. In 2007, interviews were completed in 98% of cases (U.S. Census Bureau, 2009). Furthermore, data quality and completeness is very high in ACS—equaling or outperforming the Census long form sample and the Current Population Survey on a variety of relevant measures (National Research Council, 2007). Population coverage by ACS is also very high.



## Covariates

To assess the representativeness of sample of children in the PSID compared with the ACS, we chose a set of geographic, demographic, and economic variables that are relatively consistently measured across the two surveys.

A key question for the analysis is whether the PSID adequately reflects national race and ethnic diversity, as well as immigrant composition. We obtained measures of race and ethnicity from both surveys and created the following categories: non-Hispanic whites (henceforth, “whites”), non-Hispanic blacks (“blacks”), Hispanics of any race, non-Hispanic American Indians (“American Indians”), non-Hispanic Asians (“Asians”), and non-Hispanic other races including multiple races (“others”). In the ACS, race and ethnicity is asked of every individual in the household. In the PSID, however, race and ethnicity is not asked of all children. The study began collecting race and ethnicity for children in 1997 and we have child-specific reports of race and ethnicity for approximately 77% of PSID children. For the remaining 23% of PSID children for whom we do not have a direct report of their race and ethnicity, we impute the category based on race and ethnicity of the household head or spouse.

Parent immigration status is more difficult to capture consistently between the two surveys. The ACS collects information on place of birth for all individuals. To identify each child’s parents, we applied the family interrelationship mapping algorithm developed by the Integrated Public Use Microdata Series (IPUMS) project (Ruggles et al., 2010). IPUMS provides the person numbers of the individuals present in the household who are most likely to be each child’s mother and father (including adoptive parents). In most (97%) cases, the mapping is straightforward and based on Census family relationship information. In some cases, however, IPUMS used additional data on household composition, age, and other demographic variables to identify a child’s most likely mother and father (IPUMS, 2010). We created a variable that identifies a child in the ACS as being born to an immigrant if one or both of the parents was not born in the U.S. In PSID, immigration status is not asked directly. However, we developed a

proxy indicator for a child having one or more foreign-born parents. A child is presumed to have a foreign born parent if either: (1) the child is from the 1997 immigrant refresher sample or (2) the Head grew up in a foreign country.

To measure household economic status, we used the ratio of family income in the previous year to the corresponding poverty threshold based on each family's size and age composition from the 2006 Census Needs Standard (U.S. Census Bureau, 2006).

The final set of covariates from both the ACS and PSID includes the following: child age, sex, and race/ethnicity; whether the child had a foreign-born parent; the ratio of household income to the poverty threshold; and Census Division, which distinguishes nine broad regions of the country (New England, Mid-Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific).

#### Methods for Comparing Sample Means

For comparisons of means between ACS and PSID, we examined effect sizes and the statistical significance of tests of equal means for the two samples. The advantage of examining effect sizes is that they provide insights into the substantive significance of the differences in means between the ACS and PSID. Comparisons based on statistical significance are influenced by the extremely large sample sizes in the ACS which can reject the equivalence of even substantively minor differences in means of a couple of percentage points.

We calculated the effect size as the absolute difference between the PSID and ACS means divided by the standard deviation, a measure known as Cohen's  $d$  statistic:

$$d = \frac{|\bar{X}_{ACS} - \bar{X}_{PSID}|}{s_{pooled}}$$

where  $\bar{x}$  is the mean and  $s_{pooled}$  is the standard deviation of the pooled sample. The effect size is a standardized measure of the difference between means reported in standard deviation

units. Widely used interpretation guidelines (Cohen, 1988) indicate that a small effect size is about 0.2, medium is about 0.5, and large is about 0.8.

We calculated and compared both unweighted and weighted means. Comparisons based on unweighted means are strongly influenced by each study's sample design—such as oversampling of particular population subgroups—which may limit the value of assessing coverage using this approach. On the other hand, it is straightforward to construct weights (e.g., using post-stratification) in a way that brings two populations into close alignment and hence makes comparisons of weighted means uninformative. In other words, there are limitations associated with comparisons of both weighted and unweighted means. Nevertheless, the comparisons of the two sets of means provide some basic information for assessing how well the PSID child sample achieves national representativeness. In particular, the results should help to identify segments of the PSID child sample which are under- or over-represented.

#### Generalized Boosted Regression Models

We employed a statistical technique called generalized boosted modeling (GBM) to assess the national representativeness of the 2007 PSID against the 2007 ACS. Originally developed for propensity models, boosted regression is a non-parametric, adaptive modeling algorithm that can estimate the nonlinear relationship between a dependent variable (in our case, whether or not an observation came from the PSID), and a large number of covariates (Friedman, 2001). Boosted regression is appealing for our analysis because of its flexible non-parametric approach, its prediction accuracy, and its ability to achieve a high degree of balance between the attributes of the treatment group (PSID children) and the control group (ACS children).

Boosted propensity models seek to estimate the log-odds of treatment assignment (i.e., the probability of being in the PSID),

$$g(x) = \log \left[ \frac{p(x)}{1 - p(x)} \right]$$

using an aggregation of regression trees. Initially, the algorithm sets  $g(x)$  equal to the average log-odds of treatment assignment. The algorithm then searches for a small adjustment,  $h(x)$ , that improves the fit of the model. The function  $h(x)$  is a regression tree of the residuals from the prior model as a function of the covariates. In addition, the algorithm selects a random subsample of the data on which to fit the model at each iteration, a feature that has been shown to reduce bias (McCaffrey, Ridgeway, and Morral, 2004).

McCaffrey et al. (2004) found that, compared to propensity models based on logistic regression, generalized boosted models exhibited greater prediction accuracy and better balance among covariates. Depending on the variable selection criteria used, the prediction error for the logistic regression model was 10%–16% higher than that for the generalized boosted model. In terms of balance, the GBM was also better able to remove differences in covariates between the two groups. Both the overall average effect size and the maximum effect size for any one covariate were larger for the logistic regression model.

For our analysis, we also estimated a parallel set of models using logistic regression. We found that the results for both models were similar, although the GBM models provided a noticeably better approach for including higher-level interaction effects that resulted in unstable parameter estimates when estimated using logistic regression.

We estimated the GBM models and conducted balance assessments using the TWANG (Toolkit for Weighting and Analysis of Nonequivalent Groups) package in the R statistical environment (Ridgeway, McCaffrey, and Morral, 2010).

## **Results**

We begin this section with a descriptive comparison of the sample of children in PSID and ACS. Next, we present results from the propensity analysis based on the GBM model. We then use this model to identify and describe segments of the population least likely to be

included in PSID. Because of potentially important differences in relationships by race and ethnicity, we next present results from separate models for whites, blacks, and Hispanics. Finally, using the ACS data alone, we identify and describe the subgroup of children who are excluded by design from the PSID sample.

### Descriptive Comparison of Children in PSID and ACS

In Table 1 we present and compare descriptive summary statistics for all children in the 2007 ACS and the 2007 PSID, based on both unweighted and weighted means. Effect sizes are shown for the comparisons of unweighted and weighted means but the statistical significance of observed differences is only presented for the weighted means.

The largest difference between PSID and ACS based on the unweighted mean effect size is for black children (0.69) who are substantially overrepresented in PSID (40% of sample) compared to the ACS (11%). This result is due to the PSID's oversample of poor and predominately black households as part of its original sample design (Brown, Duncan, and Stafford, 1996). Corresponding to the oversample of blacks is the underrepresentation of other race and ethnic groups, although for the unweighted means the effect size is moderate (0.36) for whites and small for Hispanics (0.22) and Asians (0.17).

The PSID weights clearly—and by construction—provide a major adjustment for the oversample of blacks and the relative undersampling of the other race/ethnic groups. In the weighted data, all of the effect sizes are classified as small although there are statistically significant differences between the ACS and PSID in the percentage of children who are white, Hispanic, and Asian. Notably, there is not a statistically significant difference between the ACS and PSID in the weighted representation of blacks.

There are differences between PSID and ACS in the representation of children of foreign-born parents. The PSID weights reduce the gap in representation of children of foreign-born parents, but they do not eliminate it. The effect size of these differences is small to moderate, both in the unweighted and weighted comparisons; however, this difference has the

largest effect size for all of the weighted contrasts and is statistically significant. Additionally, children in the lowest income quartile are overrepresented in the unweighted data (reflecting the PSID sample design), but are underrepresented in the weighted data which perhaps reflects higher attrition among this group. The opposite pattern is observed for children in the highest income quartile. There are minor differences in regional representation between the PSID and the ACS, with small and, especially in the weighted data, very small effects sizes.

In summary, the largest differences between the PSID and the ACS in representation of children in the unweighted comparison reflect the oversampling of blacks and the underrepresentation of children of foreign-born parents. These differences are reduced greatly in the weighted comparisons. In principle, the weighted differences could be eliminated almost entirely by constructing post-stratified weights for PSID to match the characteristics of the ACS. However, there remains a segment of the child population in ACS that is not included in PSID (children of post-1997 immigrants) which ought to be excluded before constructing such weights.

Because of the substantial variation by race/ethnicity in the percentage of children with foreign-born parents, the one-way comparisons of means in Table 1 could be obscuring important differences between the PSID and ACS within race/ethnic subgroups. In Table 2, we present differences in means by race and ethnicity for the three largest race/ethnic groups: whites, blacks, and Hispanics.

Based on the unweighted means, the correspondence between the ACS and PSID samples for whites, blacks, and Hispanics is quite close for all of the characteristics that we examined. The largest difference in means between ACS and PSID (based on effect sizes) are for the percentage of children with a foreign-born parent—although this is only the case for whites ( $ES=0.16$ ) and blacks ( $ES=0.38$ ). For both of these groups, the effect sizes are small-to-moderate. The PSID actually has a higher fraction of Hispanic children with a foreign-born

parent (65%) than the ACS (62%), although this result could reflect shortcomings in the procedure we used to identify children of foreign-born parents.

The geographic dispersion of the PSID sample is interesting, because it reflects not just the original sample design (including the oversampling of poorer households), but also the prevailing migration patterns over the past four decades. Among Hispanic children in the PSID, 43% are located in the Pacific region, versus 34% of Hispanic children in the ACS. Hispanic children in the PSID are less represented in the West South Central region (which includes Texas). There are some minor differences (in terms of effect size) in the correspondence between PSID and ACS by child age and income quartile. In all cases, however, the largest of the effect sizes for these differences are classified as “small” in both the weighted and unweighted comparisons.

#### Comparison of Children in PSID and ACS: Generalized Boosted Model Results

We used the generalized boosted regression framework to estimate the probability that an observation came from the PSID sample (the treatment sample,  $y=1$ ) rather than from the ACS sample (the control sample,  $y=0$ ), as a function of demographic, economic, and geographic covariates. We then applied the resulting probability weights to the PSID sample and assessed the “balance,” or extent to which the adjusted PSID sample resembles the nationally-representative ACS sample across all covariates. We let the algorithm iterate until the average effect size difference across the covariates was minimized, and allowed up to three-way covariate interactions. For the full sample, convergence was achieved after just over 7,900 iterations. GBM does not provide a standard set of regression coefficients and standard errors. Rather, the model output comprises of a table showing the relative influence of each covariate and a view of the covariate balance between the optimally-weighted treatment group and the control group, expressed in terms of the effect size of the comparison.

Figure 1 illustrates the relative influence of each covariate on the resulting propensity score. The estimate of relative influence is based on decomposing improvements in the

generalized boosted model's likelihood function as covariates are added to the model. The measure is normalized so that the total influence across all model covariates is 100%.

The most influential covariate is black race. This result is not surprising because our analysis is based on the unweighted data, and the PSID, by design, oversampled low income, predominantly black households. Census division and age are the next most influential covariates. For both of these covariates, the relative influence is aggregated over each of the multiple outcome categories. The relatively large influence of the foreign born parent covariate is also not surprising given that PSID does not capture post-1997 immigrants. Together these top four covariates (black race, Census Divisions, age, and foreign-born parent) account for over 87% of the influence in the model.

The GBM model produces a propensity weight by minimizing the effect size of differences in the weighted means between the treatment group (PSID children) and the control group (ACS children). In Table 3, we present these weighted means, the effect sizes, and the statistical significance of the differences in the weighted means. The GBM model substantially improves the balance between these two samples, relative to the balance achieved by the sampling weights alone (see Table 1). In particular, the largest effect size is now just 0.03, for the foreign-born parent covariate. This is well below the unweighted effect size of 0.32 and the effect size of 0.21 with the sampling weights. The largest unweighted effect size of 0.69 for the black race covariate has a similar weighted effect size based on the propensity weights and the sampling weights—likely reflecting the explicit adjustment in the PSID sampling weights for the oversample of poor, predominately black families in the SEO sample. However, the effect sizes based on the propensity weights are substantially lower compared to the effect sizes based on the sampling weights for Asians, Hispanics, and whites, for poverty quartile categories, and for the Census Division categories. These results suggest that, compared to a nationally-representative sample of U.S. children, the coverage of PSID is quite good. Nevertheless, there are signs from this analysis—despite the very small effect sizes on the differences—that the



PSID sample of children does not provide complete coverage of all race/ethnic groups and underrepresents children with foreign-born parents. In addition, limitations in the number of covariates and the construction of covariate categories could obscure other significant shortcomings in national representativeness of the PSID child sample.

#### Analysis of ACS Children Least Likely to Appear in PSID

We can use the propensity scores from the GBM regression to identify observations from the ACS with the lowest probability of being included in the PSID, and compare the characteristics of this group with the nationally-representative ACS sample in order to gain some additional insights into the characteristics of children least likely to appear in PSID. To do this, we sort all observations by their probability of being included in the PSID (i.e., the propensity score), and identify observations with a low, very low, and extremely low probabilities that we define as comprising of the lowest 5%, 3.5%, and 1% of cases. Although these percentages are arbitrary, they were chosen to provide an ample number of observations in each category and multiple categories were selected to assess the sensitivity of our results to choosing a specific cut-off.

The results, presented in Table 4, show a striking—but unsurprising—result that children with a foreign-born parent were least likely to appear in the PSID. In the full ACS sample (Column 1 in Table 4), the probability of having a foreign-born parent is 23%. Among the children with a low likelihood of being in the PSID, 91% have a foreign-born parent; among those with a very low likelihood of being in the PSID, 93% have a foreign-born parent and essentially all children who have an extremely low likelihood of being in the PSID have a foreign-born parent.

For the race/ethnic groups, some results are sensitive to the specific cut-off chosen to identify children with a low likelihood of appearing in the PSID. In particular, Hispanics comprise a large fraction (36%) of children with a low likelihood of appearing in the PSID but a very small fraction (2%) of children with an extremely low likelihood. In contrast, there is a consistent

finding across the different cut-off values that that Asians comprise a large percentage of children with the lowest likelihood of being included in the PSID. Asian children comprise 4% of children in the U.S., according to the full ACS, but represent 60% of children with an extremely low likelihood of appearing in the PSID. There is also evidence from these results that children in the lowest and highest income quartiles and residing in certain regions (New England, Mid-Atlantic, and West South Central) are underrepresented in PSID. However, these results for race, income, and region are likely reflecting the characteristics of children with foreign-born parents who have a low likelihood of being in the PSID, rather than representing independent effects of these variables.

#### Comparison of Children in PSID and ACS by Race/Ethnicity

Our stratified analysis of the representativeness of PSID children by race and ethnicity was restricted to black, Hispanic, and white children in order to ensure adequate sample sizes. The results, presented in Table 5, show that there are differences by race/ethnic group in the specific covariates that have lower levels of balance between the ACS and PSID. However, as with the full sample, the generalized boosted regression weights create good balance between the two samples in terms of effect size of differences in means.

The most notable finding in Table 5 is the excellent balance between the ACS and PSID for each of the three race/ethnic groups. The result is particularly striking for whites, for whom the largest effect size is only 0.04 (for foreign-born parent) and for whom there are no statistically significant differences in weighted means between the two samples. There is only one statistically significant difference in weighted means for Hispanics (for residence in the New England Census Division) and again the effect sizes for the differences in weighted means are small, with the largest values for residence in New England (0.07) and the lowest quartile of the income distribution (0.06). The balance for Hispanic children with a foreign-born parent is very good, although this may reflect a different mix between children within this category across the two samples (e.g., because we do not distinguish between children with one vs. two foreign-

born parents). Finally, for black children, there is a statistically significant difference in the weighted means between the ACS and PSID for children with a foreign-born parent. There are also a number of other statistically significant differences by region and age but, notably, not by income quartile. However, in all cases the effect sizes are small, with the largest value of 0.07 for the percentage with a foreign-born parent.

For each race/ethnic group, we also analyzed the characteristics of children in the ACS with the lowest probability of appearing in the PSID sample. We present results in Table 6 only for the “very low” likelihood category that is defined as the ACS children in the bottom 3.5% of cases ordered by their likelihood of appearing in the PSID. The results show that black and white children with the lowest probability of selection into the PSID are comprised exclusively of children with foreign-born parents. Interestingly, Hispanic children with foreign-born parents are no more likely to be in the very low probability group than in the full ACS.

There are some notable differences by region and income quartile in the distribution of children with a very low likelihood of appearing in the PSID compared to the full ACS sample. Blacks with a very low probability of appearing in the PSID are substantially more likely to be in the East North Central region (comprising of Illinois, Wisconsin, Michigan, Ohio, and Indiana); the vast majority of Hispanics with a very low probability of appearing in the PSID are also located in this region. Whites with a very low likelihood of appearing in the PSID are more likely to be residing in the West South Central region (Texas, Arkansas, Louisiana, and Oklahoma). Finally, Hispanics with a very low likelihood of being in the PSID are almost exclusively in the lowest income quartile while similar whites are more likely than not to be in the highest income quartile.

#### Descriptive Analysis of Children in the ACS Not in the PSID Sampling Frame

The final part of our analysis is to apply the PSID sampling rules to the ACS sample in order to identify the relative size and demographic and socioeconomic characteristics of children

in the ACS who—by design—would not appear in the PSID. This group comprises of children born to parents who were both born abroad and migrated to the U.S. after 1997.

In order to identify children of post-1997 immigrant parents in the ACS, we must know the immigration status of both parents. This is not possible if both parents are not present in the same household as the child. One option is to conduct this analysis only for children for co-reside with both parents. However, this restriction reduces our sample by one-third and may also lead to biased results. The approach we chose was to instead impute the missing immigration status and year of arrival for a parent when we can identify the other parent. Specifically, we calculated the probability of a second parent being foreign born, given that the first parent is also foreign born, and we calculated these probabilities separately by ancestry, education, and age. See the Appendix for details about the imputation procedure.

We present the results of this analysis in Table 7, which show that only 3.0% of the children in the ACS meet the criterion of having both parents born outside the U.S. and arriving in the country after 1997. The 2007 PSID thus, in principle, covers approximately 97% of children in the U.S.—a very high level of coverage indeed. Based on the design of the PSID, coverage was 100% in 1997 and the decline in coverage was only 3% percent per decade between 1997 and 2007.

A comparison of characteristics for children in and out of the PSID sampling frame shows that children not covered by PSID were more likely to be Asian and Hispanic. This finding is consistent with the fact that these two race/ethnic groups comprise the majority of children of recent immigrant to the U.S. (Clark et al, 2009). Children not covered by the PSID were somewhat younger than those covered by the PSID, and also lived in households with lower income. These results likely reflect the younger average age among post-1997 immigrants to the U.S.

PSID's high coverage rate among children in the U.S. reflects in part a mechanism to include children for whom only one parent was a post-1997 immigrant. In other words, all

children for whom at least one parent was living in the U.S. in 1997 are covered by the PSID sample rules; some of these children had a parent who was a post-1997 immigrant to the U.S. However, counts from the PSID sample suggest that this segment of the population is relatively small—of the 668 new spouses added to the PSID sample since 1997, only 9 were likely post-1997 immigrants.

## **Summary and Conclusions**

As panel studies age, the opportunity to study a rich set of intergenerational dynamics increases. However, the risk of non-generalizability due to attrition, non-response, and migration also increases. PSID is the longest running panel study of its kind in the world, and as such, it is critical that its national representativeness is assessed periodically. Unlike other studies which have focused on household and adult representativeness, in this technical report we assessed the national representativeness of children in the PSID, by comparing it to a nationally representative, gold-standard survey, the ACS.

We found that the PSID coverage is quite good, with some key exceptions. The PSID has the potential to cover 97% of children in the U.S. in 2007. However, among the 3% of U.S. children who are not covered by the PSID sampling design, Asian and Hispanic children of immigrant parents are over-represented. This finding is not surprising given recent trends in immigration, along with the PSID's inability to cover children born to two post-1997 immigrant parents.

While unweighted comparisons between the PSID and ACS reveal some large differences—particularly the percentages of children who are black and with foreign-born respondents—many of these differences disappear when using sample weights. We can further enhance PSID coverage by applying the GBM propensity score-derived weights. After applying the GBM weights, we found that compositional differences between the PSID and ACS based on age, race/ethnicity, sex, having a foreign-born parent, geography, and income were

considered to be small according to their effect sizes. We found that the group the PSID is least likely to cover includes Asian and Hispanic children and those with foreign-born parents. Results within race and ethnic group are not entirely consistent, however. We found that among white and black children, the PSID is least likely to cover children with a foreign-born parent. Among Hispanic children, however, the PSID is no less likely to cover children with foreign-born parents, but is less likely to cover poor children and those from the East North Central Census division.

There are several limitations to our study. First, our analysis included only a limited set of covariates that were of primary methodological interest and could be reasonably consistently defined across the ACS and PSID. Second, our PSID-based measures of children's race/ethnicity and of immigration status for parents and children relied on imputed values that may have led to classification errors as well as understated standard errors. Third, it is not possible to identify both parents in the ACS for all children; in order to use the ACS to comprehensively assess the coverage of the PSID we imputed parents' foreign-born status for children who had only one parent identified in the ACS.

As the population of children born to immigrant parents continues to grow and the duration since the last PSID immigrant refresher increases, it will be necessary to periodically re-assess the national representativeness of the various PSID samples and to consider addressing sample non-coverage through new immigrant refresher samples. These comparisons will be enhanced by the collection of more accurate data on children's race and ethnicity and parents' immigration status and year of immigration in future waves of the PSID.

## References

- Andreski, P., K.A. McGonagle, and R.F. Schoeni. 2009. "Analysis of the Quality of the Health Data in the Panel Study of Income Dynamics." Panel Study of Income Dynamics, Technical Paper Series, #09-02.
- Basic, Edin, and Ulrich Rendtel. 2007. "Assessing the bias due to non-coverage of residential movers in the German Microcensus Panel: An evaluation using data from the Socio-Economic Panel," *ASTA Advances in Statistical Analysis* 91:311–334.
- Beckett, Sean, William Gould, Lee Lillard, and Finis Welch. 1988. "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation," *Journal of Labor Economics* 6:472–492.
- Blumberg, Stephen J., and Julian V. Luke. 2007. "Coverage Bias in Traditional Telephone Surveys of Low-Income and Young Adults," *Public Opinion Quarterly* 71:734–749.
- Bosworth, B.B., and S. Anders. 2008. Savings and Wealth Accumulation in the PSID 1984-2005. Center for Retirement Research at Boston College, Working Paper 2008-2.
- Brown, Charles, Greg J. Duncan, and Frank P. Stafford. 1996. "Data Watch: The Panel Study of Income Dynamics," *Journal of Economic Perspectives* 10:155–168.
- Clark, Rebecca L., Jennifer E. Glick, and Regina M. Bures. 2009. "Immigrant Families Over the Life Course: Research Directions and Needs," *Journal of Family Issues* 30:852–872.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences, Second Edition*. Hillsdale, NJ: Lawrence Erlbaum.
- Cornman, J.C., V.A. Freedman, and F.P. Stafford. 2011. "Comparing Estimates of Time Use in the PSID's Supplement on Disability and Use of Time and the American Time Use Survey." Panel Study of Income Dynamics, Technical Paper Series, #11-01.
- Duncan, G.J., and D.H. Hill. 1989. "Assessing the Quality of Household Panel Data: The Case of the Panel Study of Income Dynamics," *Journal of Business & Economic Statistics* 7:441–452.

- Fitzgerald, J. 2011. "Attrition in models of intergenerational links in health and economic status in the PSID," *Berkeley Electronic Journal of Economic Analysis & Policy* 11(3): Article 2.
- Fitzgerald, J., P. Gottschalk, and R. Moffitt. 1998a. "An analysis of sample attrition in panel data," *Journal of Human Resources* 33:251–99.
- Fitzgerald, J., P. Gottschalk, and R. Moffitt. 1998b. "An analysis of the impact of sample attrition on the second generation of respondents in the Michigan panel study of income dynamics," *Journal of Human Resources* 33:300–344.
- Friedman, J.H. (2001). "Greedy function approximation: A gradient boosting machine," *Annals of Statistics* 29:1189–1232.
- Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley.
- Groves, Robert M. 2006. "Nonresponse rates and nonresponse bias in household surveys," *Public Opinion Quarterly* 70:646–675.
- Gouskova, E.P., P. Andreski, and R.F. Schoeni. 2010. "Comparing estimates of family income in the Panel Study of Income Dynamics and the March Current Population Survey, 1968-2007." Panel Study of Income Dynamics, Technical Paper Series, #10-01.
- IPUMS. 2010. "IPUMS User's Guide, Volume 1: Overview." Minneapolis: University of Minnesota.
- Lee, Sunghee, and Richard Valliant. 2009. "Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment," *Sociological Methods & Research* 37:319–343.
- Lee, S. 2006. "Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys," *Journal of Official Statistics* 22:329–349.
- Li, G., R.F. Schoeni, S. Danziger, and K.K. Charles. 2010. "New Expenditures in the PSID: Comparison with the CE," *Monthly Labor Review* 133:29–39.



- McCaffrey, Daniel F., Greg Ridgeway, and Andrew F. Morral. 2004. "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods* 9:403–425.
- McGonagle, Katherine A., Robert F. Schoeni, Narayan Sastry, and Vicki A. Freedman. 2012. "The Panel Study of Income Dynamics: Overview, Recent Innovations, and Potential for Life Course Research," *Longitudinal and Life Course Studies* 3(2): 268–284.
- National Research Council. 2007. *Using the American Community Survey: Benefits and Challenges*. Washington, DC: National Academy Press.
- Peytchev, Andy, Lisa R. Carley-Baxter, and Michele C. Black. 2011. "Multiple Sources of Nonobservation Error in Telephone Surveys: Coverage and Nonresponse," *Sociological Methods & Research* 40:138–168.
- Ridgeway, Greg, Dan McCaffrey, and Andrew Morral. 2010. "Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial for the twang package." Manuscript. RAND Corporation.
- Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. 2010. *Integrated Public Use Microdata Series: Version 5.0*. Minneapolis: University of Minnesota.
- Schoeni, R.F., F.P. Stafford, K.A. McGonagle and P. Andreski. Forthcoming. "Response rates in national panel surveys." In D. Massey and R. Tourangeau (Eds.), *The Future of Social Science Surveys: Opportunities and Challenge*. New York: Russell Sage.
- Schonlau, Matthias, 2005. "Boosted regression (boosting): An introductory tutorial and a Stata plugin," *The Stata Journal* 5(3): 330-354
- Schonlau, Matthias, Arthur van Soest, Arie Kapteyn, and Mick Couper. 2009. "Selection Bias in Web Surveys and the Use of Propensity Scores," *Sociological Methods & Research* 37:291–318.
- U.S. Census Bureau. 2006. "Poverty Thresholds 2006." Accessed 4/15/2012.  
<http://www.census.gov/hhes/www/poverty/data/threshld/thresh06.html>.

U.S. Census Bureau. 2009. *Design and Methodology: American Community Survey*.

Washington, D.C.: U.S. Government Printing Office.

**Table 1.** Comparison of Descriptive Statistics for Children in the 2007 ACS and 2007 PSID

	Unweighted Mean			Weighted Mean			Sig
	ACS	PSID	Effect	ACS	PSID	Effect	
<b>Age (years)</b>	8.9	8.4	0.09	8.6	8.5	0.01	
<b>Race/ethnicity</b>							
American Indian	0.9%	0.5%	0.06	0.8%	0.5%	0.03	***
Asian	4.2%	1.4%	0.17	4.0%	2.3%	0.10	***
Black	11.1%	39.5%	0.69	14.3%	13.9%	0.01	
Hispanic	18.2%	10.5%	0.22	20.7%	15.8%	0.13	***
Other	3.3%	3.6%	0.01	3.4%	3.6%	0.01	
White	62.2%	44.5%	0.36	56.8%	63.8%	0.14	***
<b>Male</b>	51.3%	50.5%	0.02	51.2%	51.4%	0.00	
<b>Foreign-born parents</b>	22.7%	10.8%	0.32	24.8%	16.5%	0.21	***
<b>Income:poverty</b>							
Quartile 1	25.0%	30.0%	0.11	28.5%	22.7%	0.13	***
Quartile 2	25.0%	27.4%	0.05	25.2%	23.9%	0.03	***
Quartile 3	24.9%	22.2%	0.06	23.5%	23.8%	0.01	
Quartile 4	25.1%	20.5%	0.11	22.8%	29.6%	0.15	***
<b>Census Division</b>							
New England	4.5%	2.5%	0.11	4.3%	3.8%	0.02	**
Mid-Atlantic	12.7%	9.2%	0.11	12.6%	13.5%	0.03	**
East North Central	15.8%	17.4%	0.04	15.3%	14.4%	0.03	**
West North Central	6.9%	7.6%	0.03	6.6%	7.4%	0.03	***
South Atlantic	18.2%	24.1%	0.14	18.6%	17.4%	0.03	**
East South Central	5.9%	9.0%	0.12	5.9%	6.6%	0.03	**
West South Central	12.2%	11.4%	0.02	12.6%	12.0%	0.02	
Mountain	7.6%	5.4%	0.09	7.6%	7.6%	0.00	
Pacific	16.3%	13.0%	0.09	16.6%	16.8%	0.00	

Note: Tabulation of 702,044 observations of children 0–17 years of age from the 2007 American Community Survey (ACS) and 7,077 children 0–17 years of age from the 2007 Panel Study of Income Dynamics (PSID). Column labeled “Effect” shows the standardized effect size for the difference between the ACS and PSID means. \* $p < .05$ ; \*\* $p < .01$ ; and \*\*\* $p < .001$

**Table 2.** Comparison of Descriptive Statistics for Children in the 2007 ACS and 2007 PSID: Results by Race/Ethnicity

	Black							Hispanic							White						
	Unweighted mean			Weighted Mean				Unweighted Mean			Weighted Mean				Unweighted Mean			Weighted Mean			
	ACS	PSID	Effect	ACS	PSID	Effect	Sig	ACS	PSID	Effect	ACS	PSID	Effect	Sig	ACS	PSID	Effect	ACS	PSID	Effect	Sig
<b>Age (years)</b>	9.2	8.8	0.10	8.8	9.2	0.08	***	8.4	7.7	0.04	8.0	8.2	0.03		9.0	8.2	0.15	8.8	8.5	0.05	***
<b>Male</b>	50.9%	49.9%	0.02	50.9%	50.9%	0.00		51.0%	50.5%	0.02	51.2%	50.7%	0.01		51.5%	50.9%	0.01	51.4%	51.5%	0.00	
<b>Foreign-born parent</b>	14.7%	1.4%	0.38	14.1%	3.3%	0.31	***	61.7%	65.3%	0.12	63.2%	67.0%	0.08	**	8.4%	4.0%	0.16	9.1%	4.7%	0.15	***
<b>Income:poverty</b>																					
Quartile 1	45.7%	45.8%	0.00	48.2%	43.8%	0.09	***	41.3%	36.4%	0.13	44.3%	38.5%	0.12	***	16.6%	14.3%	0.06	18.1%	13.9%	0.11	***
Quartile 2	25.5%	30.3%	0.11	25.6%	28.1%	0.06	***	30.0%	33.6%	0.08	29.8%	30.7%	0.02		23.7%	23.6%	0.00	23.7%	21.5%	0.05	***
Quartile 3	17.5%	16.1%	0.04	16.3%	19.2%	0.08	***	17.6%	19.2%	0.04	16.4%	18.0%	0.04		28.6%	28.5%	0.00	28.0%	26.5%	0.03	*
Quartile 4	11.2%	7.8%	0.11	9.9%	8.8%	0.04	**	11.1%	10.7%	0.01	9.5%	12.7%	0.11	***	31.0%	33.6%	0.05	30.2%	38.1%	0.17	***
<b>Census Division</b>																					
New England	2.1%	0.2%	0.13	2.1%	0.4%	0.12	***	2.4%	0.9%	0.10	2.4%	0.4%	0.13	***	5.6%	4.8%	0.03	5.7%	5.5%	0.01	
Mid-Atlantic	14.2%	17.1%	0.08	13.5%	14.7%	0.04	*	8.7%	3.3%	0.19	9.9%	2.5%	0.25	***	14.0%	13.0%	0.03	13.2%	15.3%	0.06	***
East North Central	3.1%	2.8%	0.02	15.5%	17.0%	0.04	**	6.8%	10.5%	0.15	7.3%	7.4%	0.01		19.4%	19.6%	0.00	18.8%	15.4%	0.09	***
West North Central	12.4%	5.5%	0.21	3.5%	1.5%	0.11	***	2.1%	5.5%	0.23	2.3%	4.3%	0.05		9.2%	12.7%	0.12	9.1%	9.9%	0.03	
South Atlantic	35.4%	39.8%	0.13	33.3%	28.7%	0.10	***	12.7%	12.3%	0.01	12.1%	12.2%	0.00		17.3%	14.4%	0.08	17.9%	16.9%	0.02	
East South Central	10.6%	14.8%	0.00	10.2%	12.3%	0.07	***	1.1%	0.7%	0.04	1.1%	0.3%	0.08	***	6.9%	6.4%	0.02	7.0%	7.2%	0.01	
West South Central	13.8%	13.9%	0.09	13.5%	16.7%	0.09	***	21.0%	10.9%	0.25	20.8%	15.3%	0.37	***	9.6%	9.5%	0.00	9.7%	10.3%	0.02	
Mountain	2.0%	0.5%	0.11	1.9%	1.3%	0.04	***	11.4%	12.7%	0.04	11.4%	13.1%	0.05		7.3%	8.0%	0.03	7.4%	7.8%	0.01	
Pacific	6.3%	5.5%	0.04	6.5%	7.3%	0.03		33.7%	42.7%	0.19	32.7%	43.8%	0.24	***	10.7%	11.1%	0.01	11.1%	11.1%	0.00	

Note: Tabulation of 702,044 observations of children 0–17 years of age from the 2007 American Community Survey (ACS) and 7,077 children 0–17 years of age from the 2007 Panel Study of Income Dynamics (PSID). Column labeled “Effect” shows the standardized effect size for the difference between the ACS and PSID means. \* $p < .05$ ; \*\* $p < .01$ ; and \*\*\* $p < .001$ .

**Table 3.** Comparison of Weighted Means for Children in the 2007 ACS and the 2007 PSID Using Weights Based on the Generalized Boosted Model

	<u>ACS</u>	<u>PSID</u>	<u>Effect</u>	<u>Sig</u>
<b>Age (years)</b>				
<1	5.0%	4.8%	0.01	
1	5.0%	5.2%	0.01	
2	5.1%	5.1%	0.00	
3	5.2%	5.4%	0.01	
4	5.2%	5.0%	0.01	
5	5.2%	5.3%	0.00	
6	5.4%	5.3%	0.00	
7	5.4%	5.4%	0.00	
8	5.5%	5.7%	0.01	
9	5.5%	5.2%	0.01	
10	5.6%	5.9%	0.01	
11	5.6%	5.6%	0.00	
12	5.8%	5.9%	0.00	
13	5.8%	5.8%	0.00	
14	5.9%	6.0%	0.00	
15	6.1%	5.9%	0.01	
16	6.3%	6.1%	0.01	
17	6.3%	6.3%	0.00	
<b>Race/ethnicity</b>				
American Indian	0.9%	0.7%	0.02	**
Asian	4.2%	3.8%	0.02	*
Black	11.1%	11.3%	0.01	
Hispanic	18.2%	17.6%	0.02	
Other	3.3%	3.0%	0.02	
White	62.2%	63.6%	0.03	**
<b>Male</b>	51.3%	51.2%	0.00	
<b>Foreign-born parent</b>	22.7%	21.6%	0.03	**
<b>Income:poverty</b>				
Quartile 1	25.0%	24.5%	0.01	
Quartile 2	25.0%	24.7%	0.01	
Quartile 3	24.9%	25.5%	0.01	
Quartile 4	25.1%	25.3%	0.01	
<b>Census Division</b>				
New England	4.5%	4.2%	0.02	
Mid-Atlantic	12.7%	12.4%	0.01	
East North Central	15.8%	16.0%	0.01	
West North Central	6.9%	6.8%	0.01	
South Atlantic	18.2%	18.3%	0.00	
East South Central	5.9%	5.7%	0.01	
West South Central	12.2%	12.1%	0.00	
Mountain	7.6%	7.7%	0.01	
Pacific	16.3%	16.3%	0.00	

**Table 4.** Descriptive Statistics for ACS Observations with a Low Probability of Appearing in the PSID

	<u>ACS (%)</u>	<u>Probability of appearing in PSID</u>		
		<u>Low</u>	<u>Very low</u>	<u>Extremely low</u>
<b>Age (years)</b>	8.9	9.7	9.7	7.5
<b>Race/ethnicity</b>				
American Indian	0.9%	2.4%	2.1%	1.6%
Asian	4.2%	31.5%	37.3%	59.8%
Black	11.1%	10.4%	12.6%	27.9%
Hispanic	18.2%	36.4%	32.5%	1.7%
Other	3.3%	4.6%	5.1%	6.7%
White	62.2%	14.6%	10.5%	2.4%
<b>Male</b>	51.3%	51.1%	52.0%	53.2%
<b>Foreign-born parent</b>	22.7%	90.6%	92.9%	99.9%
<b>Income:poverty</b>				
Quartile 1	25.0%	45.2%	42.4%	20.3%
Quartile 2	25.0%	7.8%	7.0%	11.8%
Quartile 3	25.0%	8.5%	7.1%	11.5%
Quartile 4	25.0%	38.5%	43.5%	56.4%
<b>Census Division</b>				
New England	4.5%	11.7%	12.2%	20.6%
Mid-Atlantic	12.7%	19.7%	20.6%	18.1%
East North Central	15.8%	1.5%	1.7%	2.7%
West North Central	6.9%	1.9%	1.8%	2.9%
South Atlantic	18.2%	17.9%	18.4%	22.3%
East South Central	5.9%	6.3%	7.1%	10.2%
West South Central	12.2%	29.4%	26.3%	9.1%
Mountain	7.6%	2.6%	2.2%	3.0%
Pacific	16.3%	9.0%	9.8%	11.1%
<b>Observations</b>	702,044	35,456	24,819	7,091

**Table 5.** Weighted Means with Propensity Weights from Generalized Boosted Regression Models: Results by Race/Ethnicity

	<b>Black</b>				<b>Hispanic</b>				<b>White</b>			
	<b>ACS</b>	<b>PSID</b>	<b>Effect</b>	<b>Sig</b>	<b>ACS</b>	<b>PSID</b>	<b>Effect</b>	<b>Sig</b>	<b>ACS</b>	<b>PSID</b>	<b>Effect</b>	<b>Sig</b>
<b>Age (years)</b>												
<1	4.6%	3.9%	0.03	**	5.8%	5.2%	0.02		4.7%	4.7%	0.00	
1	4.6%	5.0%	0.02		5.7%	6.1%	0.02		4.8%	4.7%	0.00	
2	4.6%	4.9%	0.01		5.7%	5.6%	0.01		4.9%	4.9%	0.00	
3	4.8%	5.7%	0.04	**	5.6%	5.5%	0.00		5.0%	5.0%	0.00	
4	4.9%	4.2%	0.03	*	5.6%	5.8%	0.01		5.1%	5.2%	0.00	
5	5.0%	4.6%	0.02		5.5%	5.7%	0.01		5.2%	5.4%	0.01	
6	5.3%	5.4%	0.01		5.5%	6.1%	0.03		5.4%	5.2%	0.01	
7	5.3%	5.5%	0.01		5.6%	5.9%	0.02		5.4%	5.4%	0.00	
8	5.4%	5.2%	0.01		5.5%	5.1%	0.02		5.5%	5.6%	0.00	
9	5.3%	5.7%	0.02		5.5%	5.3%	0.01		5.5%	5.4%	0.01	
10	5.6%	6.2%	0.03		5.5%	4.6%	0.04		5.6%	5.7%	0.00	
11	5.4%	5.5%	0.00		5.4%	6.5%	0.05		5.7%	5.8%	0.00	
12	5.9%	5.8%	0.01		5.7%	5.5%	0.01		5.9%	5.8%	0.00	
13	6.1%	5.8%	0.01		5.5%	5.3%	0.01		6.0%	6.0%	0.00	
14	6.3%	6.1%	0.01		5.5%	5.9%	0.02		6.1%	6.1%	0.00	
15	7.0%	6.2%	0.03	*	5.5%	5.5%	0.00		6.2%	6.2%	0.00	
16	6.9%	7.5%	0.03		5.6%	5.2%	0.02		6.5%	6.4%	0.00	
17	6.9%	6.6%	0.01		5.3%	5.0%	0.01		6.5%	6.5%	0.00	
<b>Male</b>	50.9%	49.8%	0.02		51.0%	51.3%	0.01		51.5%	51.6%	0.00	
<b>Foreign-born parent</b>	14.7%	12.1%	0.07	***	61.8%	63.6%	0.04		8.4%	7.3%	0.04	
<b>Income:poverty</b>												
Quartile 1	45.8%	45.7%	0.00		41.3%	38.5%	0.06		16.6%	16.6%	0.00	
Quartile 2	25.5%	25.8%	0.01		30.0%	31.2%	0.03		23.7%	23.5%	0.01	
Quartile 3	17.5%	17.4%	0.00		17.6%	18.8%	0.03		28.6%	29.1%	0.01	
Quartile 4	11.2%	11.1%	0.00		11.1%	11.4%	0.01		31.0%	30.8%	0.00	
<b>Census Division</b>												
New England	2.1%	1.3%	0.05	***	2.4%	1.4%	0.07	**	5.6%	5.2%	0.02	
Mid-Atlantic	12.4%	12.6%	0.01		8.7%	8.1%	0.02		14.0%	13.8%	0.01	
East North Central	14.2%	14.2%	0.00		6.8%	6.7%	0.01		19.4%	19.4%	0.00	
West North Central	3.1%	2.7%	0.02		2.1%	1.8%	0.03		9.2%	9.2%	0.00	
South Atlantic	35.4%	37.3%	0.04	**	12.7%	12.3%	0.01		17.3%	17.4%	0.00	
East South Central	10.6%	10.6%	0.00		1.1%	0.8%	0.03		6.9%	6.7%	0.01	
West South	13.8%	14.2%	0.01		21.0%	21.3%	0.01		9.6%	9.6%	0.00	
Mountain	2.0%	1.5%	0.03	**	11.4%	11.9%	0.01		7.3%	7.2%	0.01	
Pacific	6.3%	5.4%	0.04	**	33.7%	34.8%	0.02		10.7%	10.9%	0.01	

**Table 6.** Descriptive Statistics for ACS Observations with a Low Probability of Appearing in the PSID: Results by Race/Ethnicity

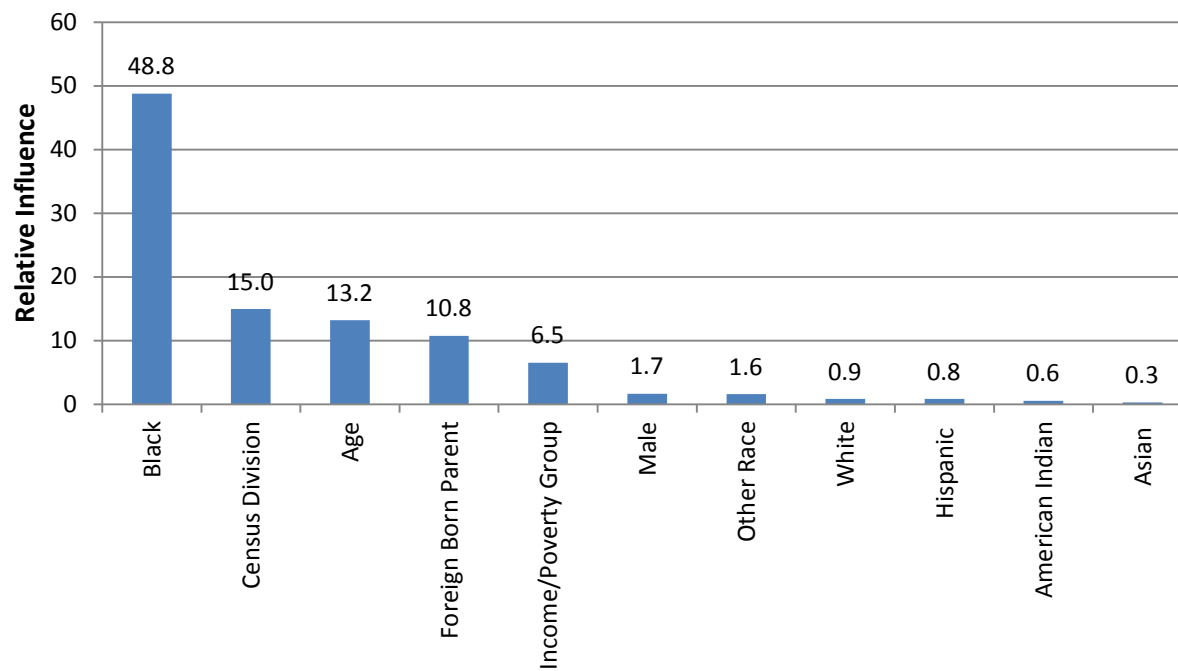
Variable	<b>Black</b>		<b>Hispanic</b>		<b>White</b>	
	<b>ACS</b>	<b><u>Very low prob. of appearing in PSID</u></b>	<b>ACS</b>	<b><u>Very low prob. of appearing in PSID</u></b>	<b>ACS</b>	<b><u>Very low prob. of appearing in PSID</u></b>
<b>Age</b> (years)	9.2	10.1	8.4	10.6	9.0	10.3
<b>Male</b>	50.9%	57.4%	51.0%	52.0%	51.5%	47.8%
<b>Foreign-born parent</b>	14.7%	100.0%	61.7%	57.0%	8.4%	100.0%
<b>Income:poverty</b>						
Quartile 1	45.7%	35.5%	41.3%	96.6%	16.6%	5.9%
Quartile 2	25.5%	15.3%	30.0%	1.2%	23.7%	20.4%
Quartile 3	17.5%	19.1%	17.6%	0.0%	28.6%	21.4%
Quartile 4	11.3%	30.1%	11.1%	2.2%	31.0%	52.4%
<b>Census Division</b>						
New England	2.1%	3.7%	2.4%	0.1%	5.6%	11.7%
Mid-Atlantic	12.4%	0.7%	8.7%	0.0%	14.0%	5.0%
East North Central	14.2%	39.4%	6.8%	81.5%	19.4%	18.0%
West North Central	3.1%	8.0%	2.1%	0.0%	9.2%	2.6%
South Atlantic	35.4%	22.7%	12.7%	10.6%	17.3%	14.8%
East South Central	10.6%	11.8%	1.1%	0.0%	6.9%	4.0%
West South Central	13.8%	3.8%	21.0%	0.0%	9.6%	30.1%
Mountain	2.0%	5.5%	11.4%	0.0%	7.3%	5.9%
Pacific	6.3%	4.3%	33.7%	7.8%	10.7%	8.0%
<b>Observations</b>	78,107	2,831	127,928	4,503	436,431	15,385



**Table 7.** Comparison of ACS Children's Characteristics based on their inclusion in the PSID Sampling Frame

<b>Variable</b>	<b>Mean</b>		<b>Sig.</b>
	<b>Not In PSID</b>	<b>In PSID</b>	
<b>Age (years)</b>	7.15	8.85	***
<b>Race/ethnicity</b>			
American Indian	0.1%	0.9%	***
Asian	23.2%	3.7%	***
Black	7.1%	10.4%	***
Hispanic	50.8%	17.0%	***
Other	1.7%	3.4%	***
White	17.0%	64.6%	***
<b>Male</b>	51.6%	51.2%	
<b>Income:Poverty</b>			
Quartile 1	44.0%	23.5%	***
Quartile 2	25.2%	24.9%	
Quartile 3	15.2%	25.5%	***
Quartile 4	15.5%	26.0%	***
<b>Census Division</b>			
New England	4.3%	4.6%	*
Mid-Atlantic	12.4%	12.9%	**
East North Central	8.8%	16.1%	***
West North Central	3.1%	7.1%	***
South Atlantic	21.7%	18.0%	***
East South Central	2.3%	5.9%	***
West South Central	14.0%	12.0%	***
Mountain	8.1%	7.6%	***
Pacific	25.3%	16.0%	***
<b>Observations</b>	3.0%	97.0%	

**Figure 1.** Relative Influence of Variables in the Generalized Boosted Regression Model



## Appendix

### **Imputation of Immigration Status for ACS Children**

In order to identify children of post-1997 immigrant parents in the ACS, we need to determine the immigration status of both parents. This is a challenge because when both parents do not coreside with a child we are unable to determine the parents' immigration status. The percentage of children for whom we can identify 0, 1 or 2 parents is shown in Table A.1.

**Table A.1.** Number of Parents Identified for ACS Children

<b>Number of coresident parents in ACS</b>	<b>Number of children</b>	<b>Percent of children</b>
0	34,985	5.0%
1	188,575	26.9%
2	478,484	68.2%
	702,044	100.0%

If we restricted our analysis to only those children for were coresident with both parents, Table A.1 shows that the analysis sample would decline by about one-third. We would lose 5% of the sample because neither parent is coresident. For children not coresiding with either parent, Table A.2 shows that in most cases they are living with a grandparent, a more distant relative (e.g., an uncle or aunt), or a non-relative.

**Table A.2.** Relationship to Household Head for Children with No Coresident Parent

<b>Relationship to head</b>	<b>Number of children</b>	<b>Percent of Children</b>
Grandchild	15,493	44.28
Other relative	7783	22.25
Other non-relative	6381	18.24
Institutional inmate	2070	5.92
Sibling	1854	5.3
Sibling-in-law	562	1.61
Head/householder	378	1.08
Partner, friend, visitor	304	0.87
Spouse	114	0.33
Child-in-law	46	0.13
	34,985	100.0%

Compared to the full ACS sample, Table A.3 shows that the 35,000 children in the ACS with no coresident parent are more likely to be black or Hispanic and to foreign-born.

**Table A.3.** Racial and Ethnic Composition of ACS Children with No Coresident Parent

Variable	No coresident parent	Full ACS
American Indian	2.2%	0.9%
Asian	2.7%	4.2%
Black	26.7%	11.1%
Hispanic	22.7%	18.2%
Other	3.8%	3.3%
White	41.9%	62.2%
Foreign born	6.8%	4.5%

For our analysis to identify the percentage of children in the ACS who are not in the PSID sampling frame, we require the post-1997 immigration status of both parents. For children in the ACS coresiding with only one parent who was a post-1997 immigrant, we therefore need to impute the likelihood that the other parent was also a post-1997 immigrant. Our approach is as follows. First, for each post-1997 immigrant parent in a two-parent household, we calculate the probability that the other parent was a post-1997 immigrant by the first parent's age, sex, education, and ancestry category. We collapsed age, education, and ancestry categories to maintain a minimum cell size of approximately 100 observations. With the exception of parents with Mexican ancestry, we classified age into two groups, (1) less than 35 years or (2) 35 years or older, and classified education into two groups as well, (1) 12 years or less and (2) more than 12 years (i.e., some college). For parents of Mexican ancestry, which was a large group, we used a finer education classification. Because many parents of Mexican ancestry had low levels of education, we classified their education into the following three groups: (1)  $\leq 8$  years, (2) 9–12 years, and (3) more than 12 years (i.e., some college).

Even with the large categories we used for each variable, we had sparse cells that had to be combined. For example, we moved Indian, Japanese/Korean/Chinese, and Middle Eastern fathers with less than 12 years of education to the corresponding "Other Asian" ancestry category due to small cell sizes. Similarly, we collapsed the age groups for parents of African origin with less than 12 years of education. The final classification scheme for parents along with the probability of the second parent being a post-1997 immigrant (given that the first parent is a post-1997 immigrant) is presented in Table A.4.

**Table A.4.** Probabilities by Group that the Other Parent is Also a Post-1997 Immigrant

Ancestry	Age	Education	Sex	Probability that non-coresident parent is also a post-1997 immigrant	Obs.
African	<35 Years	$\leq 12$ Years	Female	63.7%	270
African	<35 Years	$\leq 12$ Years	Male	72.8%	169
African	<35 Years	Some College	Female	57.3%	206
African	<35 Years	Some College	Male	54.3%	140
African	$\geq 35$ Years	Some College	Female	69.8%	182
African	$\geq 35$ Years	Some College	Male	81.1%	280

Central American	<35 Years	≤12 Years	Female	54.5%	336
Central American	<35 Years	≤12 Years	Male	69.6%	237
Central American	≥35 Years	≤12 Years	Female	44.9%	98
Central American	≥35 Years	≤12 Years	Male	67.9%	81
Central American	≥35 Years	Some College	Female	40.7%	167
Central American	≥35 Years	Some College	Male	61.7%	128
Indian	<35 Years	Some College	Female	65.3%	943
Indian	<35 Years	Some College	Male	85.8%	520
Indian	≥35 Years	Some College	Female	76.2%	269
Indian	≥35 Years	Some College	Male	89.2%	445
Japanese/Korean/Chinese	<35 Years	Some College	Female	57.6%	505
Japanese/Korean/Chinese	<35 Years	Some College	Male	87.3%	229
Japanese/Korean/Chinese	≥35 Years	Some College	Female	71.1%	650
Japanese/Korean/Chinese	≥35 Years	Some College	Male	90.9%	624
Mexican	<35 Years	≤8 Years	Female	57.1%	1050
Mexican	<35 Years	≤8 Years	Male	72.0%	785
Mexican	<35 Years	>8 and ≤12 Years	Female	51.1%	2015
Mexican	<35 Years	>8 and ≤12 Years	Male	64.6%	1370
Mexican	<35 Years	Some College	Female	47.2%	477
Mexican	<35 Years	Some College	Male	66.9%	284
Mexican	≥35 Years	≤8 Years	Female	41.5%	593
Mexican	≥35 Years	≤8 Years	Male	80.5%	421
Mexican	≥35 Years	>8 and ≤12 Years	Female	51.0%	620
Mexican	≥35 Years	>8 and ≤12 Years	Male	73.4%	512
Mexican	≥35 Years	Some College	Female	45.3%	311
Mexican	≥35 Years	Some College	Male	75.3%	271
Middle Eastern	<35 Years	≤12 Years	Female	34.8%	141
Middle Eastern	<35 Years	Some College	Female	41.9%	222
Middle Eastern	<35 Years	Some College	Male	53.0%	115
Middle Eastern	≥35 Years	≤12 Years	Female	61.8%	102
Middle Eastern	≥35 Years	Some College	Female	58.8%	131
Middle Eastern	≥35 Years	Some College	Male	78.7%	188
Other	<35 Years	≤12 Years	Female	52.5%	162
Other	<35 Years	≤12 Years	Male	45.7%	138
Other	<35 Years	Some College	Female	47.1%	310
Other	<35 Years	Some College	Male	55.5%	182
Other	≥35 Years	≤12 Years	Female	40.0%	135
Other	≥35 Years	≤12 Years	Male	62.7%	126
Other	≥35 Years	Some College	Female	52.5%	141
Other	≥35 Years	Some College	Male	61.8%	136
Other Asian	<35 Years	≤12 Years	Female	31.9%	517
Other Asian	<35 Years	≤12 Years	Male	59.4%	219
Other Asian	<35 Years	Some College	Female	42.5%	659
Other Asian	<35 Years	Some College	Male	77.4%	235
Other Asian	≥35 Years	≤12 Years	Female	61.0%	589
Other Asian	≥35 Years	≤12 Years	Male	86.4%	450
Other Asian	≥35 Years	Some College	Female	53.0%	615
Other Asian	≥35 Years	Some College	Male	81.9%	529
Other Caribbean	<35 Years	≤12 Years	Female	40.6%	143
Other Caribbean	<35 Years	≤12 Years	Male	48.6%	111
Other Caribbean	≥35 Years	≤12 Years	Female	46.5%	155

Other Caribbean	≥35 Years	≤12 Years	Male	60.8%	143
Other Caribbean	≥35 Years	Some College	Female	50.8%	124
Other Caribbean	≥35 Years	Some College	Male	67.7%	124
Other Eastern European	<35 Years	≤12 Years	Female	70.9%	189
Other Eastern European	<35 Years	≤12 Years	Male	77.2%	114
Other Eastern European	<35 Years	Some College	Female	48.0%	275
Other Eastern European	<35 Years	Some College	Male	77.5%	129
Other Eastern European	≥35 Years	≤12 Years	Female	85.1%	148
Other Eastern European	≥35 Years	≤12 Years	Male	90.6%	191
Other Eastern European	≥35 Years	Some College	Female	67.8%	199
Other Eastern European	≥35 Years	Some College	Male	86.0%	207
Other Western European	<35 Years	≤12 Years	Female	43.7%	206
Other Western European	<35 Years	≤12 Years	Male	41.5%	130
Other Western European	<35 Years	Some College	Female	42.6%	340
Other Western European	<35 Years	Some College	Male	46.5%	260
Other Western European	≥35 Years	≤12 Years	Female	57.8%	187
Other Western European	≥35 Years	≤12 Years	Male	58.1%	160
Other Western European	≥35 Years	Some College	Female	64.0%	542
Other Western European	≥35 Years	Some College	Male	69.8%	648
South American	<35 Years	≤12 Years	Female	51.6%	213
South American	<35 Years	≤12 Years	Male	69.2%	107
South American	<35 Years	Some College	Female	43.5%	230
South American	<35 Years	Some College	Male	62.0%	108
South American	≥35 Years	≤12 Years	Female	62.6%	203
South American	≥35 Years	≤12 Years	Male	87.3%	126
South American	≥35 Years	Some College	Female	64.2%	288
South American	≥35 Years	Some College	Male	87.3%	283

We used the probabilities in Table A.4 to estimate the likelihood that a child coresiding with just one parent was within the PSID sampling frame. The estimated probabilities of a child not being within the PSID sampling frame based on status whether they were coresiding with one or both parents is presented in Table A.5.

**Table A.5.** Summary of ACS Children’s Inclusion in the PSID Sampling Frame Based on Coresidence with Parents

Child’s coresidence status	Not in PSID frame	In PSID frame	Total	Non-coverage rate
Coresident with one parent	4,046	184,529	188,575	2.1%
Coresident with both parents	15,665	462,828	478,484	3.3%
	19,711	647,357	667,059	3.0%

Children coresiding with only one parent are more likely to be in the PSID sampling frame than those coresiding with both parents. This is because the latter group of children is more likely to have parents who are older, have higher levels educational attainment, and be of Indian, Japanese/Chinese/Korean, and other Eastern European ancestry; these children are both more likely to live with both parents and it is more likely that both parents are post-1997 immigrants.