Survey Instruments and the Reports of Consumption Expenditures: Evidence from the Consumer Expenditure Surveys*

Erich Battistin
University of Padova and IRVAPP

Mario Padula
University Ca' Foscari of Venice and CSEF

March 7, 2010

Abstract

This paper provides evidence on the relevance of the mode collection for the analysis of consumption data for the United States using complementary data sets from the Consumer Expenditure Surveys (CEX). We first show that population figures from consumption reports obtained with diaries markedly differ from those obtained using recall data. We then exploit multiple measurements of food expenditure available in the CEX to identify the effects of the mode collection on important features of the distribution of consumption. Finally, we show how to purge the expenditure measurements from most of the effects of mode collection and thus obtain a better measure of consumption. The paper concludes by suggesting some guidelines for empirical research that have important implications for the measurement of inequality and well being.

Keywords: Consumption; Data Collection Methods; Rank Invariance.

JEL Classification: C13, C21, D12.

*This is a largely revised version of a previous paper authored by Erich Battistin circulated under the title *Errors in Survey Reports of Consumption Expenditures* and whose first draft was presented at the NBER Summer Institute 2002. The current version benefited from useful discussions with Orazio Attanasio, Margherita Fort, David Johnson and Luigi Pistaferri, and from comments by audiences at Napoli, 3^{rd} ICEEE Conference, 2009 North American Meeting of the Econometric Society and NBER Summer Institute 2009. Email addresses for correspondence: erich.battistin@unipd.it and mpadula@unive.it.

1 Introduction

Data quality is an issue of longstanding concern among researchers interested in testing the implications of behavioural models of consumption. The empirical analysis of these models requires good micro-data on expenditures at household or individual level. This paper considers data quality issues arising from the analysis of expenditure data for the United States, characterizes the effect of the mode collection on the reports of expenditure categories widely considered in empirical studies, and devises appropriate remedies to measurement issues that are relevant to empirical research.

In many countries expenditure data are regularly collected either by diaries covering purchases made within a short period of time (typically one or two weeks) or by retrospective questions on the *usual* spending over a longer period (see Browning *et al.* (2003)). There is a consensus that the time consuming task required by diaries produces good quality expenditure data for small and frequently purchased items, while recall questions should be targeted to bulky items (major consumer durables: real property, automobiles and major appliances) or for those components either having regular periodic billing or involving major outlays (such as transportation or rent). Such an idea is not only intuitively clear, but it is also supported with evidence from cognitive studies (see, for example, Winter (2002)) and from the comparison of aggregated consumption measures to national account data (see Garner *et al.* (2009)).

The drawback of this idea is that *neither* diary *nor* recall data *alone* can provide a reliable aggregate measure of *total* expenditure. One might argue that for any practical purpose data collected using these alternative survey methodologies lead to same empirical conclusions, but unfortunately this is not the case. There is evidence that data from recall questions lead to potentially misleading results in the analysis of household saving behavior (see for example Battistin et al. (2003)). Battistin (2003) and Attanasio *et al.* (2007) show that data collected using diaries or retrospective questions imply nearly opposite policy conclusions about the evolution of consumption inequality over time in the United States. Other studies demonstrate how adjustments provide greater consistency concerning

the time series properties of consumption (Slesnick (1998)). As pointed out by Wilcox (1992), the imperfections of micro-data on consumption expenditures may be important enough to influence the conclusions of empirical work. Are data relevant to the theory? Is the economic model really in error? Should research be directed towards alternative models of economic behavior or are data themselves not suitable to validate existing models?

Ideally, the computation of aggregate expenditures at the *micro-level* would require detailed information on a variety of consumption categories obtained with the most appropriate methodology.[1] However, because of time constraints and survey practice, questionnaires cannot cover all the aspects of consumption with the same level of accuracy. Thus, learning about the effects of the collection mode is important for empirical research but, at the same time, represents a very difficult task. On the one hand one would need to compare figures obtained from diaries and recall questions for different expenditure groups and across several types of individuals in the population. As a matter of fact, cognitive studies designed to this end often refer to specific expenditure groups and typically don't involve a large sample of individuals from the population of interest. On the other hand complex phenomena such as forgetfulness and telescoping (see Neter and Waksberg (1964)) call for the analysis of the effects of the survey instrument on the full distribution of expenditures, not just its mean.

The potential of combining retrospectively collected information to diary information on household consumption using micro-level data from the Consumer Expenditure Survey (CEX in the following) has been first brought forward by Battistin (2003). This survey consists of *two* different components: a quarterly Interview Survey (IS) and a weekly Diary Survey (DS), each with its own questionnaire and sample. The most interesting feature that makes the CEX a unique and extremely appealing source of data is that the IS and the DS overlap for nearly all expenditure categories for which information is collected using different methodologies (recall questions and diaries, respectively). The two survey

[1]The Family Expenditure Survey for the United Kingdom represents a notable implementation of this strategy. It consists of a comprehensive household questionnaire which asks about regular household bills and expenditure on major but infrequent purchases and a diary of all personal expenditure kept by each household member (including children) for two weeks. According to the evidence for the United Kingdom, consumption measures obtained from such a design are comparable to aggregated values from national accounts (see, for example, Banks and Johnson (1998)).

components are explicitly designed to collect information on *different* types of expenditures. The IS is targeted to those types of expenditures that respondents can recall for a period of three months or longer; the DS is instead designed to obtain reliable data on frequently purchased smaller items. Neither survey alone is thus expected to represent all aspects of consumption. Accordingly, the Bureau of Labor Statistics (BLS in the following) publishes aggregate figures by combining data from the two components to provide a complete accounting of consumer expenditures which, by design, neither survey component alone is designed to do.

Given the overlap between the IS and the DS for many categories of consumption, and given that the IS is explicitly designed to collect good quality information only on a subset of these categories, the question then arises of whether DS and IS micro-data can be *jointly* exploited to derive a *superior* measure of household spending. Building upon the results in Battistin (2003), Attanasio *et al.* (2007) and Attanasio *et al.* (2008) pursue a number of strategies to combine information from the two survey components of the CEX to study the evolution of consumption inequality in the US over the last two decades. One of the most puzzling results that arises from these papers is that the overall picture regarding the evolution of inequality heavily depends upon the survey instrument exploited. This finding alone implies that data from at least one (and possibly either) survey are affected by *measurement error*, and that its extent might have important implications for empirical research.[2]

This paper contributes to this discussion by offering two new contributions of considerable policy and practical relevance, as well as of methodological interest. First, we show that for food expenditure the mode collection (recall questions vis-à-vis diaries) is roughly *rank preserving*, in the sense that the relative position of households in the IS and DS distributions is the same net of random slippages

[2]Inconsistencies between the CEX and national accounts have been pointed out in the literature by several authors (notably Slesnick (1998), (2001)) suggesting that the quality of these data might have deteriorated over the last decade. Battistin (2003) and Attanasio *et al.* (2007) reports a decline in the ratio of CEX to Personal Consumption Expenditures for non-durables and services over the 90s. This gap and its growth over time is even larger when attention is restricted to the IS. Slesnick (2001) reports similar findings for total consumption, shows that only part of this discrepancy can be explained by definitional differences and concludes that "*the remaining gap is a mystery that can be resolved only by further investigation*" (page 154). This evidence contrasts sharply with similar comparisons for the United Kingdom, where aggregating a time series of individual cross sectional data one obtains close to 95% of non-durable consumption, as documented in Banks and Johnson (1998).

that we are able to characterize.[3] We thus show that the ranking of households in the expenditure distribution is less affected by the survey instrument than the reporting of expenditures. To this end we make use of *multiple measurements* of food spending which follow from DS households being asked, before the beginning of the diary, about *usual* spending on food using the same retrospective questions as in the IS. This results in the same collection mode (i.e. recall questions) applied to independent samples of (similar) households, and in different collection modes (i.e. recall questions and diaries) applied to the same sample of households. To the best of our knowledge, this is the first paper that exploits this unique feature of the CEX.

Second, by assuming that the same *rank preserving* property holds across all expenditure groups (i.e. not just for food), we characterize the effect of varying the survey instrument on all components of non-durable spending. We find that the effect varies a great deal along observable dimensions such as age, ethnicity and education and, most notably, with the value of expenditure. Most importantly, this characterization allows us to back out *for each household* an alternative measure of total spending by following the BLS procedure that establishes, for each expenditure group, the most reliable survey measurement to use (DS or IS). Households reports are thus purged of the effects of the survey instrument by using a procedure that closely follows from the design of the two survey components of the CEX. We show how to identify *any* functional of the distribution of total expenditure by combining the most reliable information from the DS and the IS, and therefore generalize the procedure suggested by Attanasio *et al.* (2007).

*** CHECK THIS *** Our findings have important practical implications for empirical research. Knowing the distribution of *potential* reports of consumption expenditures resulting from diary and recall instruments may be informative about the extent of measurement error in the data. Most importantly, recent evidence by Kofi *et al.* (2006) and by Garner *et al.* (2009) show that PSID expenditure data align quite closely with the CEX. Thus understanding the effects of the survey

[3] Our approach closely follows similar ideas previously suggested by Heckman *et al.* (1997) and Chernozhukov and Hansen (2005).

instrument in the CEX also reveals measurement properties of expenditure records in the PSID. Since the CEX has been often used to impute consumption data to the PSID households (see, for example, Blundell, *et al.* (2008)), the validity of the imputation procedure rests upon assumptions on these properties.

*** CHECK THIS *** The remainder of the paper is organized as follows. Section 2 describes the two CEX surveys and the sample used in the analysis. Section 3 poses the general identification problem for inference about the effects of changing the survey instrument on the report of consumption expenditures. Section 4 clarifies how multiple measurements available in the CEX surveys can be used to get round this problem for the case of food expenditure. Section 5 derives the identifying restrictions and their testable implications. The estimation strategy is discussed in Section **??**. Results are presented in section 7. Implications and conclusions are drawn in Section 9.

2 Data

2.1 The consumer expenditure surveys

The CEX is currently the only micro-level data set reporting comprehensive measures of consumption expenditures for a large cross-section of households in the United States. The survey is run by the BLS and consists of two separate components based on a common sampling frame, each of them with its own questionnaire addressing a *different* sample.[4]

In the IS, households are interviewed about their expenditures every three months over five consecutive quarters. Information is collected using recall questions on the *usual* weekly or monthly spending, depending on the item. After the last interview, households are dropped and replaced by a new unit, so that, by design, 20 percent of the sample is tossed out every quarter. Expenditure information is collected in the second through the fifth interview; one month recall expenditures are asked in the first interview only for bounding purposes. The DS is instead a cross-section of consumer units

[4]Sample designs differ only in terms of frequency and over sampling of DS households during the peak shopping period of Christmas and New Year holidays.

asked to self-report their daily purchases for two consecutive one-week periods using product-oriented diaries. Each diary is organized by day of purchase and by broad classifications of goods and services (see, for example, **Silberstein and Scott 1991** for a detailed discussion of this survey).

2.2 The definition of expenditure categories

Throughout the analysis only figures for expenditure on *non-durable goods and services* will be considered. In particular, expenditures on durables, health, education and mortgage/rent payments are excluded. Given that we want to compare information from the two surveys, this is, we think, the safest choice to make. The DS presumably does a very bad job picking up expenditures on infrequently purchased items and most durables because of the short time horizon it refers to.[5]

The definition of non-durable expenditure is the one in Attanasio and Weber (1995) and is broadly consistent with the one considered by several studies in the literature. Nine expenditure categories are considered (see Table 1): food and non-alcoholic beverages (both at home and away from home), alcoholic beverages, tobacco and expenditures on other non-durable goods such as heating fuel, public and private transports (including gasoline), services and semi-durables (defined by clothing and footwear). Expenditure groups have been made comparable across surveys and consistent over time, focusing on non-durable items common to the two surveys.[6] Public use tapes permit to integrate data on non-durable consumption from both surveys only after 1986, since only selected expenditure and income data from the DS were published before then.

Only expenditure figures for the month preceding the interview are considered for the IS sample, thus leaving four observations for each household (one observation per interview/quarter). Monthly

[5] Attanasio *et al.* (2008) exploit information on expenditures for durables to come out with a complete picture on the evolution of consumption inequality in the United States from the early 80s, thus extending previous work by Attanasio *et al.* (2007).

[6] Two (apparently unavoidable) limitations to the full comparability of the CEX surveys are worth mentioning. First, although the bulk of the questionnaires and survey methodology were remarkably stable over time, some minor changes did occur. For example, new diaries with more cues were introduced in the DS in the early 90s and, for the IS, the food question changed in 1982 and 1988 (see Appendix B for more details). Second, the two surveys are *not* completely exhaustive for non-durable expenditure. For most items, we have a measure both for the households in the DS and for those in the IS. However, the IS excludes expenditures on housekeeping supplies (e.g. postage stamps), personal care products and non-prescription drugs, which are instead collected in the DS. On the other hand, the DS excludes expenditures incurred by members while away from home overnight or longer and information on reimbursements (such as for medical care costs or automobile repairs), which are collected in the IS.

expenditure in the DS is defined as $26/12 = 2.16$ times the expenditure observed over two weeks, assuming equally complete reporting. Family consumption is adjusted using the OECD equivalence scale (although our results are fairly robust to this choice) and real expenditures are obtained using the Current Price Index published by the BLS.[7]

2.3 The working sample

Sample information is used for the period 1982-2003. All consumer units satisfying at least one of the following criteria: (i) living in rural areas, (ii) with single females, (iii) residing in student housing, (iv) whose head is self-employed or (v) whose head is aged below 25 and above 65, are not considered in the final sample. Additionally, we dropped consumer units presenting null expenditure on total food (both at home and away from home), incomplete income response or not completing the diary. The importance of each selection step for the size of the final sample is documented in Appendix A (see Table A-1).

Throughout the analysis the family head will be conventionally fixed to be the male in all husband/wife families, representing the 53 percent and 56 percent of the whole sample for DS and IS data, respectively. In the IS, all available observations for the same household over the interview period will be used.

3 The identification problem

3.1 General setup

The general identification problem can be easily put across by considering the standard programme evaluation setting. Alternative collection modes can be seen as mutually exclusive states of the world that are potentially available to measure household expenditures. Households are assigned to different modes (retrospective questions or diaries) so that only measurements corresponding to the mode

[7] There are of course important comparability issues in combining bi-weekly data from one source (the DS) and monthly data from another (the IS) that need to be addressed. Low expenditures in one two-week period may be made up with higher expenditures over the next two-week period, and *viceversa*. There are basic economic reasons to expect that more smoothing will be done over a longer time period. However, a model-based approach to account for the frequency of purchasing is not exploited in this paper and is left to future research.

assigned are revealed. Let D be a dummy indicator which takes value one if recall questions are used (like in the IS) and zero if diaries are exploited (like in the DS). Two *potential* measurements corresponding to each survey instruments are logically defined. Let Y_1 and Y_0 be the two measurements corresponding to recall questions and diaries, respectively. They represent *potential outcomes* from using alternative survey instruments to collect information on household consumption. The measurement actually observed is instead $Y = Y_0 + D(Y_1 - Y_0)$, so that $Y = Y_0$ if $D = 0$ and $Y = Y_1$ if $D = 1$. Data allow identification of $F_{Y_1|D}[\eta|1]$ and $F_{Y_0|D}[\eta|0]$, that is of the expenditure distributions obtained by using recall questions and diaries, respectively.[8]

Let X be observable characteristics that affect expenditure, and assume the following condition.

Assumption 1 (strong ignorability). For all values x there is:

$$(Y_0, Y_1) \perp D | X = x,$$

$$e(X) \equiv P[D = 1 | X = x] \in (0, 1),$$

where $e(X)$ is the propensity score.

Using the balancing property of the propensity score (see Rosenbaum and Rubin, (1983)), the conditional independence condition stated in Assumption 1 holds also with respect to $e(X)$, implying that the following distributions are identified from observed data:[9]

$$\begin{aligned} F_{Y_0}[\eta] &= \int F_{Y_0|D,e(X)}[\eta|0, e] dF_{e(X)}[e], \\ F_{Y_1}[\eta] &= \int F_{Y_1|D,e(X)}[\eta|1, e] dF_{e(X)}[e]. \end{aligned}$$

The conditional independence condition in Assumption 1 implies that the marginal distributions of Y_1 and Y_0 can be recovered from the observed distributions in DS and IS by simply correcting for

[8] It also implies that all distributions that are conditional to any observable variable are identified. Here and in what follows, the notation $F_{A|B,C}[a|b, c]$ will indicate the conditional distribution of A given $B = b$ and $C = c$ calculated at a.

[9] It is worth noting that the estimated distributions in what follows are not conditional on survey membership D, that is they identify the effect of survey instrument for a randomly chosen household in the *population*. This we think is the most interesting object to consider, as it allows to extend our results to other surveys representative of the US population like the PSID. Alternatively, under the assumptions stated one could also identify *conditional* distributions for the DS ($D = 0$) and the IS ($D = 1$) populations. This difference closely resembles the difference between *treatment effects in the population* and *treatment effects on the treated* in the evaluation literature.

compositional differences which are entirely due to observables X. In other words, it amounts to saying that households with the same values of X in either survey component can be taken as random samples of the same population. Being the two surveys designed by the BLS to obtain representative figures for the population of the United States, this appears to be a credible assumption in the context of this paper as it accounts for residual imbalances that may result from non-response or from the sample selection criteria adopted.[10]

Clearly, the difference $Y_1 - Y_0$ is informative about the effect of the survey instrument on the reporting of expenditures. The comparison between $F_{Y_1}[\eta]$ and $F_{Y_0}[\eta]$ reveals certain features of this effect. For example, the difference in the means of the marginal distributions of Y_1 and Y_0 corresponds to the *mean* effect on the measurement of expenditures following to a change in the survey instrument.[11] Identification of other features of the distribution $F_{Y_1-Y_0}[\eta]$ from the marginals $F_{Y_1}[\eta]$ and $F_{Y_0}[\eta]$ is in general precluded without additional assumptions. In the next section sufficient conditions are provided to achieve identification.

3.2 Rank preserving property of survey instruments

First notice that the following representation holds:

$$Y_0 = F^{-1}_{Y_0|e(X)}[U_0|e], \qquad Y_1 = F^{-1}_{Y_1|e(X)}[U_1|e],$$

where U_0 and U_1 are uniform random variables that can be interpreted as ranks of the corresponding conditional distributions. The requirement that:

$$U \equiv U_0 = U_1, \tag{1}$$

for all values of $e(x)$, is sufficient to recover the joint distribution of Y_1 and Y_0 conditional on $e(x)$. The *rank invariance* condition (1) preserves perfect dependence in the ranks between the two distributions

[10]Note that the requirement of common support condition $e(X) \in (0, 1)$ is key to retrieve the marginal distributions of interest, as identification relies on knowledge of the conditional distributions of expenditure at common values of the propensity score.

[11]If for example one is willing to assume that some types of expenditure are measured more accurately by using diaries rather than retrospective questions, this mean difference can be approximately interpreted as the mean of the measurement error distribution $F_{Y_1-Y_0}[\eta]$. This is for example the approach suggested by Battistin (2003), and further developed by Attanasio *et al.* (2007) and Attanasio *et al.* (2008).

of potential measurements and may be motivated by the existence of a common *unobserved factor* U (say, preferences) that determines the ranking of a given household across distributions determined by different collection modes. The identification power of rank invariance has been discussed by Heckman, Smith and Clements (1997) and, more recently, by Chernozhukov and Hansen (2005) and (2006).

In what follows we will make use of rank invariance conditional on values of the propensity score $e(x)$. This amounts to invoking a rank preserving property of the survey instruments across the conditional distributions of potential outcomes $F_{Y_0|e(X)}[\eta|e]$ and $F_{Y_1|e(X)}[\eta|e]$. By using the balancing properties of the propensity score (see Rosenbaum and Rubin, (1983)), the informational content of (1) amounts to saying that rank invariance holds for groups of households sharing the same distribution of the characteristics X.

Rank invariance implies that the joint distribution of potential outcomes Y_1 and Y_0 is not truly bivariate. For example, it follows from (1) that:

$$U_0 = F_{Y_0|e(X)}[Y_0|e], \qquad Y_1 = F^{-1}_{Y_1|e(X)}[U_0], \tag{2}$$

so that knowledge of U_0 is sufficient to retrieve Y_1. Via Assumption 1, this in turn implies that $F_{Y_1-Y_0|e(X)}[\eta|e]$ can be fully recovered from the marginal distributions $F_{Y_0|e(X)}[\eta|e]$ and $F_{Y_1|e(X)}[\eta|e]$. Identification of the distribution of $Y_1 - Y_0$ thus straightforwardly follows.

4 Survey instruments and reports of food expenditure

The requirement (1) can not be tested against data in general. Assumption 1 allows one to retrieve the marginal distributions of expenditure but, since the two components of the CEX refer to *different* households, identification of any functional of the joint distribution is precluded. To get around this problem, Battistin (2003) proposes a bounding approach which for the problem at hand often leads to inconclusive inference, and calls for the use of parametric models. Attanasio *et al.* (2007) make assumptions on the stability over time of the effects of the survey instrument to study the evolution of consumption inequality. In what follows we take a different route, and frame the problem in a more

general setting which includes that of Attanasio *et al.* (2007) as a particular case. In particular, our approach provides a direct test for the validity of Assumption 1 as well as of the rank condition (1). The general idea builds upon Battistin (2003), who pointed out the potential of using multiple reports of food expenditure to study the effects of the survey instrument in CEX data.

The case of food spending is particularly suited to this end, as features of the CEX survey design define multiple measurements for this category. *Three* measurements of food spending are available in the data: one from diary records coming from the DS, and two from global questions coming from the DS and the IS. The set of global questions is the *same* for respondents of the DS and the IS, thus implying that two samples of *independent* households representative of the same population are interviewed with the *same* collection mode (recall questions).[12] The design also implies that two measurements of food expenditure collected using *different* survey modes (recall questions and diaries) are available for the *same* sample of DS households.

It follows that for food spending the *marginal* distributions:

$$F_{Y_0|D,e(X)}[\eta|0,e], \qquad F_{Y_1|D,e(X)}[\eta|0,e], \qquad F_{Y_1|D,e(X)}[\eta|1,e],$$

are identified in the data. Any detectable difference between the latter two distributions should be taken as evidence against the *ceteris paribus* condition implied by Assumption 1, either distribution being obtained from the same survey mode (recall questions). In practise this amounts to testing:

$$F_{Y_1|D,e(X)}[\eta|0,e] = F_{Y_1|D,e(X)}[\eta|1,e], \tag{3}$$

[12]*** CHECK THIS *** In its current format, the questionnaire design does not make use of recall question to ask respondents directly about their food spending. Rather, respondents are first asked a sequence of questions for the *usual weekly* spending at the grocery stores or supermarkets, then asked about how much of this amount was for non-food items, and finally asked about *usual weekly* expense on food items at places other than grocery stores. The reference period to recall this information is the three months preceding the interview. Food spending as derived in public use data files thus results from the difference between a question on *usual total* spending at grocery stores and a question about *usual* spending on *non-food* items at these places, to which is added *usual* spending of food items at places *other than* grocery stores. We report in Appendix B the exact wording of these questions as well as a detailed description of changes that occurred over time. As shown in Battistin (2003) and in Appendix B, the time series of food expenditure is heavily affected by these changes, which determine statistically significant breaks in 1988. For food at grocery, the median goes up and the median absolute deviation from the median down after 1988 in the DS, while in the IS they both increase. The pattern for food at home in the IS is similar, since it accounts for a substantial fraction (above 90 percent) of the food at grocery.

for all values of η across all conditional distributions defined by varying the propensity score $e(X)$.[13] Most importantly, by design it also follows that the *bivariate* distribution:

$$F_{Y_1,Y_0|D,e(X)}[\eta_1, \eta_0|0, e],$$

is identified in the data. This represents the joint distribution of diary and recall measurements on food spending for DS households at common values of the propensity score. Under rank invariance, this distribution can be completely recovered from the two marginals $F_{Y_0|D,e(X)}[\eta|0, e]$ and $F_{Y_1|D,e(X)}[\eta|0, e]$, which are both identified in the data. A direct test on the validity of (1) can thus be constructed by comparing empirical distributions for given values of the propensity score to theoretical distributions obtained under the null hypothesis of rank invariance.

5 Identifying restrictions and their testable implications

The aim of this section is threefold. First, we will test for the validity of (3) using data on food and show that this condition is not rejected in the data. This result, though limited to food expenditure, provides evidence on the validity of Assumption 1. Second, we will test the validity of rank invariance using food from the DS and show that, under this assumption, knowledge of the marginals is *not* enough to back out the joint distribution of the two measurements observed in the data. Nevertheless, we will show that the joint distribution can be recovered by weakening the assumption of rank invariance to allow for slippages that we are able to characterize. Finally, we will assume common distribution of slippages across expenditures and use rank invariance to characterize the effect of the survey instrument across all categories.

[13] It is worth noting that the survey instrument might play a different role in the way it affects these distributions, as the relative position of the global question on food is not the same in the DS and IS questionnaires. In the former survey the question is asked at the very beginning of the interview, before the household starts the diary. In the latter survey the question is asked at the end of the interview. In what follows we will ignore possible effects arising from this difference in the ordering.

5.1 Testing the strong ignorability condition

As a result of sample selection, the DS and IS samples present compositional differences along important dimensions that possibly reflect in differences in expenditure behavior. To study the extent of this problem we modeled the probability of belonging to the IS sample vis-à-vis to the DS sample as a function of a rich set of household characteristics that are common across the two surveys. In particular, we focused on proxies of family composition as well as for those factors that have proved relevant to data quality in previous analysis of CEX data (see Tucker (1992)). Modeling such probability amounts to modeling the propensity score $e(X)$.

Estimation was carried out separately by expenditure year, and the propensity score estimated from a probit regression in which the dependent variable is zero for DS and one for IS households, and the independent variables are a full set of family type, ethnicity and education dummies, as well as month of expenditure dummies. The full set of estimation results for the propensity score is reported in Appendix C, and points to differences in the composition of the two samples with different patterns across survey years.

Under the assumption that all sampling differences are adequately captured by the variables included in the propensity score (Assumption 1), households sharing the same value of the propensity score also share the same distribution of the characteristics X (see Rosenbaum and Rubin, (1983)). Thus after conditioning on $e(x)$ any difference observed in the distribution of expenditures should reflect solely the nature of the survey instrument exploited. Exploiting the availability of a recall measure for food expenditure in the DS, we can formally test whether this condition is not rejected in the data. In particular, we tested the condition (3) by first stratifying households in the two samples on the estimated propensity score, and then testing for the equality of the distribution of the two recall measurements within strata.[14]

[14]We stratified observations into 15 groups depending on the value of $e(X)$. This choice ensured enough sample size within each stratum on the one hand, and it also guaranteed the balancing property of the propensity score (see Rosenbaum and Rubin, (1983)) for all strata.

The p-values of a Wilcoxon-Mann-Whitney test for the independence between Y_1 and D fail to reject the null hypothesis at the conventional level, thus implying that the two distributions of food spending statistically line up once compositional differences are taken into account (the evidence reported in Garner *et al.* (2009) points to the same result). Though limited to spending on food, this proves a necessary condition for the validity of Assumption 1.[15]

5.2 Testing the rank invariance condition

The availability of multiple measurements allows one to assess the assumption that the survey instrument is rank preserving. We considered households in the DS and, within each stratum defined by the propensity score, we constructed the difference $U_1 - U_0$ between their ranks in the two distributions of food spending. In Figure 1 we report the distribution of these differences across households, for two groups of years and for the same cell of the propensity score (the informational content of all other distributions is similar to that in the figure). The evidence provided is clearly against rank invariance: though the distribution of the difference of ranks is centered at zero, there is a great deal of variation from the mean that can hardly be reconciled with the hypothesis of rank invariance.[16]

It might therefore be desirable to allow the rank to change across survey instruments reflecting some unobserved, unsystematic variation. This can be achieved by weakening (1) to get the following condition.

Assumption 2 (random slippages from rank invariance). For all values of $e(x)$ there is:

$$U_1 = U_0 + V,$$

[15] The details on the stratification adopted as well as on the testing procedure are fully documented in Appendix C. Table A-7 presents the p-values of the null hypothesis (3) for the non-parametric test statistics considered, which we derived taking into account that strata are defined from the estimated propensity score using a bootstrap procedure (see Appendix C for further details). We also considered a parametric procedure which we run separately for each stratum. First, we grouped the values of Y_1 into four categories defined by quartiles of its distribution, and regressed the resulting ordinal categorical variable on D and a polynomial in $e(X)$ to control further for within stratum heterogeneity. P-values from this procedure are reported in Table A-6.

[16] A pooled regression of U_1 on U_0, $e(X)$ and year dummies yielded coefficients for the intercept and U_0 very precisely estimated at 0.056 and 0.892, respectively. *** CHECK THIS *** See Appendix 9 for a detailed description of the procedure that we followed to test the rank invariance assumption.

where V is a random variable that describes slippages whose distribution is such that:

$$F_{V|U_0,D,e(X)}[\eta|u_0,0,e] = F_{V|U_0,e(X)}[\eta|u_0,e]. \tag{4}$$

The informational content of Assumption 2 can be summarized as follows. First, it implies that the distribution of potential outcomes is fully characterized by the joint distribution of U_0 and V. For example, the relationship in (2) can be modified as follows:

$$U_0 = F_{Y_0|e(X)}[Y_0|e], \qquad Y_1 = F^{-1}_{Y_1|e(X)}[U_0 + V].$$

Second, Assumption 2 holds conditional on $e(x)$ and thus the distribution of slippages V is left to vary with X through the propensity score. This is important, as we found that the distribution of slippages varies a great deal with values of the propensity score, which in our data implies some degree of heterogeneity along observable dimensions such as family type, ethnicity and education. Finally, being the distribution of ranks with bounded support, the distribution of slippages cannot be assumed independent of the distribution of ranks. Because of this, the distribution in (4) will vary with u_0. We make the assumption that the extent of this correlation is household specific and does not depend on whether the household is surveyed in the DS or IS sample.[17]

5.3 Restrictions on the distribution of slippages

Assumption 2 implies that the key ingredients to recover the joint distribution of potential outcomes are the marginal distributions plus the distribution of slippages in (4). For the case of food expenditure, these distributions are all non-parametrically identified in the data. For all remaining expenditure categories, however, only the marginal distributions of potential outcomes can be recovered from raw data. If one is willing to make the assumption that the distribution of slippages remains stable across expenditure groups, under Assumption 2 it is possible to characterize the effects of changing the survey

[17] Assumption 2 embodies the idea of the approach taken by Chernozhukov and Hansen (2005). The effect of the survey instrument is completely modeled by $U_1 = U_0 + V$, which represents a measurement error model for ranks. Note that, since the distribution of U_0 has bounded support, the distribution of V must depend on U_0 and thus measurement error cannot have classical form.

instrument on expenditures other than food.[18]

Assumption 3 (common distribution of slippages). For all values of $e(x)$ and u_0 the conditional distribution $F_{V|U_0,e(X)}[\eta|u_0,e]$ is stable across all expenditure categories.

6 Estimation

??

Marginal distributions for all expenditure categories are estimated by the empirical analogues of the quantities defined in Section 3. We first derived kernel estimates conditional on strata defined by the propensity score by expenditure year and separately for the IS and the DS samples ($\hat{F}_{Y|D,e(X)}[\eta|i,e]$, $i=0,1$). We then integrated the conditional distributions with respect to the observed propensity score distribution $\hat{F}_{e(X)}[e]$, thus obtaining:

$$\hat{F}_{Y_0}[\eta] \equiv \int \hat{F}_{Y|D,e(X)}[\eta|0,e]d\hat{F}_{e(X)}[e], \qquad \hat{F}_{Y_1}[\eta] \equiv \int \hat{F}_{Y|D,e(X)}[\eta|1,e]d\hat{F}_{e(X)}[e],$$

for DS and IS expenditures, respectively.[19] The distribution of the effects of using recall questions vis-*à*-vis diaries on the reporting of food expenditure was derived as follows:

$$\hat{F}_{Y_1-Y_0}[\eta] \equiv \int \hat{F}_{Y_1-Y_0|D,e(X)}[\eta|0,e]d\hat{F}_{e(X)}[e],$$

where again the integrand was obtained as a kernel estimate of the distribution of $Y_1 - Y_0$ in the DS.

Estimation for the other expenditure groups was computed according to the following steps. First, we modeled parametrically the distribution of slippages $F_{V|U_0,D,e(X)}[\eta|u_0,0,e]$ by fitting a flexible distribution using food measurements in the DS (see Appendix C for details). Second, we took 50 random draws from the fitted distribution of slippages and used the relationships:

$$U_0 = \hat{F}_{Y|D,e(X)}[Y_0|0,e], \qquad \hat{Y}_{1j} = \hat{F}^{-1}_{Y|D,e(X)}[U_0+V_j|1,e], \qquad j=1,\ldots,50$$

[18] *** CHECK THIS *** A necessary condition for the validity of Assumption 3 is that the joint distribution of potential outcomes recovered belongs to the set of distributions defined by the convex hull of the two observable marginals. Bounds on the joint distribution can always be obtained from knowledge of the marginals using classical probability theory (see Firpo and Ridder (2008)). If the distribution implied by Assumption 3 lies outside these bounds, then this should be taken as evidence against the validity of the identifying condition.

[19] For food expenditure, (3) makes the conditioning on D in $\hat{F}_{Y|D,e(X)}[\eta|1,e]$ redundant, so that DS or IS data (or both) could be used to compute the integrand in the right hand side expression.

to impute recall measurements $\hat{Y}_{1j}$ onto the DS sample, and the relationships:

$$U_1 = \hat{F}_{Y|D,e(X)}[Y_1|1,e], \qquad \hat{Y}_{0j} = \hat{F}^{-1}_{Y|D,e(X)}[U_1 - V_j|0,e], \qquad j = 1,\ldots,50$$

to impute diary measurements $\hat{Y}_{0j}$ onto the IS sample. Finally, by defining:

$$\Delta_j \equiv (Y_1 - \hat{Y}_{0j})D + (\hat{Y}_{1j} - Y_0)(1-D), \qquad j = 1,\ldots,50$$

we computed stratum specific kernel estimates of the distributions of Δ_j, $\hat{F}_{\Delta_j|e(X)}[\eta|e]$, which were used to compute:

$$\hat{F}_{Y_1-Y_0}[\eta] \equiv \int \frac{1}{50}\sum_{j=1}^{50}\left(\hat{F}_{\Delta_j|e(X)}[\eta|e]\right) d\hat{F}_{e(X)}[e].$$

7 Results

*** LOGS OR LEVELS? *** Were the collection mode irrelevant, one should not find appreciable differences in expenditure reports whether the marginal distribution is estimated with recall or diary data, and the distribution $F_{Y_1-Y_0}[\eta]$ should be nearly degenerate. In addition to expenditure on food, in what follows we will also consider two other broader categories of expenditure derived from the aggregation rule used by the BLS for the publication of aggregate totals (see Garner, McClelland and Passero (2009) for details): one comprising $\mathcal{D}$ goods (food at home, food away from home, alcohol, tobacco, housekeeping services, personal care and entertainment services), and one comprising $\mathcal{R}$ goods (housing and public services, heating fuel, light and power, transportation, clothing and footwear and services). The classification reflects the idea of distinguishing between components having regular periodic billing or involving major outlays, and expenditures on smaller, frequently purchased items and services.[20] Assessing the effects of the survey instrument on the reporting of $\mathcal{D}$ goods is an interesting exercise in itself, as it sheds light on the effects of using the alternative collection mode when diaries are presumably most suited. A similar interpretation applies to the reporting of $\mathcal{R}$ goods, which are presumably easier to recall.

[20]Evidence in favor of this classification, which is rather conventional across all statistical offices, is reviewed by Battistin (2003). In the case of the CEX, this is exactly what motivates the existence of the Diary and the Interview components.

We will present results for the mean, the median and the inter-quantile range of $F_{Y_1-Y_0}[\eta]$ for food, $\mathcal{D}$ and $\mathcal{R}$ expenditures over the period $1988-2003$ (see Figure 2).[21] The survey effects for food expenditure (see the top panel of Figure 2) are on average positive, thus implying that records collected using recall questions on food overstate diaries of about 18 to 25 percentage points depending on the year considered. This result is likely to depend on the sequence of recall questions used to ask respondents about their spending on food (see Appendix B), or to problems associated with diaries (such as insufficient attention given by the respondent to recording the purchase, or proxy reporting for all individuals in the household; see Silberstein and Scott (1991)) that typically result in under-reporting. The median is also positive, but about half the value of the mean: the distribution is skewed towards large, positive values, and the skewness does not change much over time. The $25^{th}-75^{th}$ range steadily increases over time, pointing to a change of about 5 percentage points over the period considered. Overall these results are indicative of sizable effects of the survey instrument that become more disperse over time across households in the populations.

The distribution $F_{Y_1-Y_0}[\eta]$ for $\mathcal{D}$ goods is characterized by decreasing mean and median over time (see the central panel of Figure 2). The median survey effect is negative across all years, implying that recall questions lead to understate expenditures compared to diaries. The same result applies to the distribution mean after 1998, painting sizeable negative effects of about 5 to 10 percentage points in 2003 depending on the location indicator considered. The difference between mean and median again remains relatively constant over time, and the $25^{th}-75^{th}$ range steadily increases by about 10 percentage points. Finally, the distribution $F_{Y_1-Y_0}[\eta]$ for $\mathcal{R}$ goods is characterized by large positive values for both mean and median (see the bottom panel of Figure 2). The two location indicators generally decrease over time, but remain steadily apart. *** CHECK THIS *** The extent to which recall data overstate the diaries is actually decreasing by about 8 percentage points from 1998, thus

[21] Public use data allow to retrieve information on $\mathcal{R}$ goods only after 1986 and the survey designed has changed after 1987 (see Appendix B). Appendix D shows the estimated full distribution of consumption reports and survey impacts. The results reveal a great deal of heterogeneity of the survey impacts: the survey instrument has sizeable effects on the available measures of consumption. Measuring consumption with recall or diary report is not indifferent, which leaves open the question of whether one should use the recall, the diary or possibly both data.

suggesting that the consumption measures elicited with the two survey instruments are becoming increasingly similar, which is consistent with the overall decrease of the interquartile range. As for the other expenditure categories considered, heterogeneity in these effects plays an important role. The evidence reported in the central and bottom panels of Figure 2 clarifies the rationale for publishing totals using integrated data from the two survey components of the CEX. The idea that households under-report expenditures when interviewed using the least appropriate collection mode is generally accepted (see for example, the discussion in Silberstein and Scott (1991)), so that aggregation of expenditures on $\mathcal{D}$ and $\mathcal{R}$ goods yields larger consumption totals.

Overall, the evidence suggests that the survey effects for the D and the R goods decrease on average, which implies that the two survey methods yield increasingly similar data. However, the average effect hides important differences between goods and survey instruments, which are revealed by the pattern of the interquartile range. For the D goods the heterogeneity of the survey effects increases over time, for the R goods in increases until 1997 and then it decreases.

*** CHECK FROM NOW ON ***

To shed further light on the pattern of heterogeneity, we correlate the effects of the survey instrument to households characteristics, such as age, ethnicity, education and family type, expenditure and family income before taxes and check if the correlation has changed over time.

We pool the data for the years $1988-1990$, $1991-1995$, $1996-2000$ and $2001-2003$. The results are shown in the Tables **??**, **??**, and **??**, for food, $\mathcal{D}$ and $\mathcal{R}$ goods expenditure, respectively.

The effect of the survey instruments on food expenditure reports varies a great deal with the respondent characteristics. Recall tend to overstate diaries: the extent of overstatement decreases with age and is lower for the black, but increases with education. [22] The results points to a great deal of heterogeneity across households observable characteristics and imply that survey impacts are

[22]Recall questions overstate food expenditure diaries by about 34 percent in years 1988-1990 at the average values of food expenditure and income before taxes, for the single households, with 55+, white and high-school drop out heads. This is obtained adding to the constant term the sum of two terms: the product between the diary measure coefficient and the average of diary measure of food expenditure and the product between the income in log coefficient and the average of income in log.

negatively correlated to the diary measure of food expenditure. The correlation with family income before taxes is positive and statistically significant, thus suggesting that the survey impacts change across the income distribution. Over time, heterogeneity increases: the more educated overstate by more and more, the less than 35 and the black overstate by less and less, and, except for the husband and wife only families, the estimated coefficients on the type of family dummies increase in absolute value. The correlation between income and the survey impacts is instead stable over time.

The results for $\mathcal{D}$ goods convey the same basic message. The heterogeneity of impacts across socio-economic groups is large and statistically detectable. While for the more educated recall overstate diaries, for the all the other groups the opposite happens: recall understate diaries. Compared to the baseline category, recall understate diaries for all other family types, except for Husband and Wife only households, for the black, and for households with other ethnicity heads after 1996 and overstate for the more educated. The correlation with diary expenditure is negative, that with income positive. Both these correlations are stable over time. Conversely, the effect of the households observable characteristics changes over-time: on the one hand, holding fixed the other characteristics, for households whose head is aged less than 35 and between 35 and 45 the recall understate the diaries at an over-time increasing rate; on the other, the impact for the husband and wife only families looks years after years increasingly similar to that for the singles.

The results for the $\mathcal{R}$ goods confirms that the survey impacts varies across socio-economic group. The correlation with education is positive and larger for the more educated. The impacts decrease with age, are lower for the black and increase with income. Over time, the impacts for the the young (less than 45) are increasingly smaller that those for the 55+, those for the high school graduate larger and larger. The estimated coefficients change over time also for the black, the correlation with income attenuates. The other coefficients are instead more stable over time.

Knowledge of the distribution of survey effects, and not just of its mean, allows to identify the characteristics of the households for whom recall questionnaire are more likely to overstate consumption

reports.

Tables ??, ??, and ?? refer to food, $\mathcal{D}$ and $\mathcal{R}$ goods expenditure, respectively, and show the results from estimating the probability that the recall overstate the diary records as a function of the respondent age, education, ethnicity, family-type. We also control for the the expenditure as measured by the diary records and for income.

For the $\mathcal{D}$ goods, diaries are more likely overstate recall for the young and the black , and to understate for the more educated. For husband and wife only households diaries are more likely to overstate recall, and this is increasingly so for families with children. Income has a positive effect on the probability of recall overstating diaries, which instead decreases with consumption. The coefficients are quite stable over time, thus suggesting that the decrease in the average effect of the survey instrument is not due to the changing proportion of those for whom recall understate (or overstate) diaries.

The time pattern for the $\mathcal{R}$ goods is different. The probability of recall overstating diaries is related to age, family type, ethnicity, education, income before taxes and expenditure form the diaries, but changes at different rates for the various socio-economic groups. The black are decreasingly less likely to overstate, the high school graduate more likely, the correlation with income decreases. Overall, the evidence thus suggest that the performance of diaries versus recall changes across different socio-economic group in a way that depends on the group of goods considered.

Our results have three implications. First, the effects of the survey instruments on consumption reports are sizeable. The time evolution of the median effect implies that for food and $\mathcal{R}$ expenditure, recall overstate diaries by about 10 and 15 percent over the period considered, for $\mathcal{D}$ goods recall understate diaries by up to 11 percent in 2003. The distribution of these effects fans out between 1988 and 2003 for food and $\mathcal{D}$ goods expenditure, while it spreads out up to 1997 and shrinks afterwards for $\mathcal{R}$ goods expenditure: the interquartile range increases by 10 and by almost 20 percentage points for food and $\mathcal{D}$ goods expenditure, respectively.

Second, recall and diary reports are multiple measures of expenditure that cannot be treated as

substitute. At least one measure is error ridden, with the error varying a great deal with households observable characteristics. For instance, the average effect of the survey instrument on measured $\mathcal{D}$ goods expenditure can be made close to zero for single households with 55+, black, and high-school drop out head. Similarly, the proportion of households for whom recall questions overstate diaries varies considerably across households. For food expenditure, that proportion decreases with age and is lower for the black, for $\mathcal{R}$ goods the proportion increases with education and age.

Third, the impact of the survey instruments differs within the income distribution, with the family income before taxes affecting the average impact of the survey instrument and the probability of recall overstating diaries.

Forth, recall and diaries are two noisy measures of expenditure with different statistical properties, which change between expenditure groups and across the household population. This suggests that combining recall and diaries has the potential to provide us with a superior measure of household spending, which is what we turn to in the next section.

8 Implications for empirical research

That the survey instruments matter for consumption reports raises the question of which data one should use to investigate model of consumers behavior. Repeated measures are often used to deal with the measurement error that inevitably plagues many survey data. In a recent paper, Browning and Crossley (2009) argue that two noisy measures might be better than one expensive, accurate one. The argument is that the joint behavior of two measures might help at understanding the statistical properties of the underlying, latent consumption at a fair cost. The focus is on the estimation of the variance of consumption, since the comparison between the time-evolution of consumption and income inequality allows to discriminate across competing models of consumer behavior. The question is quantitative in nature, and is often one of how much the income inequality has increased compared to the consumption inequality (see Krueger and Perri(n.d.) and Blundell *et al.* (2008)).

To investigate the implications of our results for empirical research, we also focus on the evolution on the consumption inequality in the US. The available evidence is not conclusive. While Krueger and Perri (n.d.) show that consumption inequality has a flat time-series profile, Attanasio *et al.* (2007), Attanasio *et al.* (2008) and more recently Meyer and Sullivan (2009) document a steady increase of consumption inequality over the years 1980-2003. The issue is likely to be one of measurement. Whether one uses consumption reports from diaries or recall surveys, it makes a great difference for the evolution of the consumption distribution.

Our method allows to make the best use of available data. Since for the $\mathcal{D}$ goods recall understate diaries, and for the $\mathcal{R}$ goods diaries understate recall, our measure of total consumption will draw on diaries for the expenditure on the $\mathcal{D}$ and on recall for the expenditure on $\mathcal{R}$ goods. Since we are able to identify the joint distribution of diaries and recall for all expenditure categories, we can estimate any functional of the data. Here we focus on two indicators of consumption inequality: the $25^{th} - 75^{th}$ range and the median absolute deviation from the median. The two indicators are robust to extreme outliers and are therefore less sensitive to sample selection choices. The upper panel of Figure 3 focuses on the $25^{th} - 75^{th}$ range, the lower on the median absolute deviation from the median. Each panel features three lines, the time-series profile of the inequality indexes in the DS and in the IS and for our combined measure of total consumption. The increase of the $25^{th} - 75^{th}$ range is mild in the Interview Survey: from 70 in 1982 to 74 percent in 2003. Therefore, the trend in the $25^{th} - 75^{th}$ range disclose the same basic message of a moderate increase of consumption inequality found by Krueger and Perri (n.d.). [23] The trend in the Diary Survey is instead different, thus confirming that the instruments matters for consumption reports: the $25^{th} - 75^{th}$ changes from 74 percent to 98, implying a much more pronounced increase in consumption inequality. The $25^{th} - 75^{th}$ range for the combined measure of consumption lies in between and increases from 72 to 80 percent. Consumption inequality thus

[23] There are however differences in the sample selection. Krueger and Perri (n.d.) select out respondent who completed less than 5 interviews, respondent who report only food expenditures for the quarter, those who report positive labor income but no hours worked, and those with negative or zero after-tax labor earnings plus transfers. Moreover, we exclude household whose head is less than 25 and more that 65, while Krueger and Perri exclude household whose head is less than 21 and more that 64.

appear to increase by above 10 percent. The patterns in the lower panel are similar. The median absolute deviation from the median is quite stable in the Interview survey, and increases in the Diary Survey, and the combined measure ranging from 36 to above 40 percent.

9 Conclusions

Diary surveys are purposively designed to record information soon after the expenditure has occurred, thus potentially eliminating recall problems typical of interview surveys. In this paper we have shown that the collection mode for expenditure data significantly exacerbates the relative importance of survey errors, resulting in inconsistent population figures obtained from the two survey components of the Consumer Expenditure Survey (CEX). The collection mode matters for consumption reports: the effects are sizeable, and imply that data drawn from recall and diaries are not perfect substitute.

Several studies have already discussed various sources of errors associated with alternative collection modes (see, for example, Lyberg and Kasprzyk (1991)). In the diary-interview CEX comparison a complex picture emerges, in which a surprising number of items have roughly equivalent results in terms average expenditure in the population (see Silberstein and Scott (1991) and Garner, McClelland and Passero (2009)). However, diaries appear to improve the reporting of smaller, less salient purchases, whereas recall interviews yield better data on less frequent and more salient purchases. This simple idea is reflected in the practice followed by several statistical offices, including the Bureau of Labor Statistics (BLS), to publish figures for totals that integrate information on single expenditure items using the most reliable source of data. In the case of the CEX, this motivates the existence of the Diary or the Interview components. The choice of the most reliable source involves the computation of estimates from the two survey components, considers the frequency of reports and compares raw figures with those on personal consumption from the National Accounts (see Garner, McClelland and Passero (2009)).

Despite the extensive literature assessing the quality of CEX information (see, for example, Slesnik

(2001)), this paper marks something of a departure by making a fairly simple point. Since the estimation of intertemporal models of consumption requires reliable micro-data on household expenditures (Attanasio and Weber (1993) and (1995)), the rationale for the existence of the Diary and Interview components is somehow at odds with the fact that all empirical studies have used data only from one (and, typically, the Interview) component. This is particularly worrying, as recent evidence by Battistin (2003) and Attanasio, Battistin and Ichimura (2007) suggests that diary or interview data may lead to quite different conclusions with respect to the evolution of consumption inequality and the definition of consumption poverty. As the effects of the collection mode are far from negligible, integrating diary and interview micro-data consistently with the rule followed by the BLS would represent a first step towards reconstructing a *superior* measure of expenditure for samples of households representative of the US population on a continuous basis since the early 1980s. This would also have important practical implications for empirical research.

This is the problem that we have tried to address. The two survey components of the CEX not only have different methodologies, but also have different samples (though sharing the same design). As far as the computation of population totals or means is concerned, which is the scope pursed by the BLS, the existence of independent samples is not a problem. However, it makes integration impossible at the micro-level without additional assumptions: a straightforward application of the rule followed by the BLS reveals the most reliable measurement of the marginal distributions for certain expenditure items, while in fact we are interested in their joint distribution to compute total consumption. We have used multiple measurements of food spending available in the CEX to shed light on the effects of the collection mode on this expenditure category. In particular we have shown that the diary and interview instruments are roughly *rank preserving*, in the sense that the relative position of households in the expenditure distribution is unaffected by the collection mode. This is an important regularity which is interesting in itself, as food expenditure represents a sensible share of total spending for a large proportion of households in the population and is collected in other general purpose surveys

(such as the PSID) where CEX totals have often been imputed (see, for example, **Blundell**).

We have shown that the assumption of rank invariance is sufficient to impute interview expenditures onto the Diary sample and diary expenditures onto the Interview sample (see, for example, **Heckman**), and is thus sufficient to retrieve micro-data on all expenditure items presumably purged of most survey errors using the same method of integration followed by the BLS. Thus we have proposed a way to make the best use of available data and combine the information from either component of the CEX; in particular, our procedure allows identification of any functional of total consumption, including various inequality measures. Also, the procedure that we have proposed allows to characterize the *distribution* of the effect of the collection mode, which is also an interesting exercise in itself. Explain why.

*** CHECK THIS *** We have shown that the effects change in a predictable fashion with households characteristics and differ across expenditure groups. Diaries overstate recall questions by up to 10 percent in 2003 for food away from home, alcohol, tobacco, housekeeping services, personal care and entertainment services, while for housing and public services, heating fuel, light and power, transportation, clothing and footwear and service recall overstate diaries, at a rate that decreases over time from 25 to 7 percent. The chances of diaries overstating recall decreases with age and with education. The effect of the survey instrument is also positively correlated with income for all expenditure groups considered. To the extent that the effect of the survey instrument convolutes the measurement error in diaries and recall, this cannot be compatible with a classical measurement error in consumption unless measurement error in income and consumption are correlated. The distribution of the effects shows also important differences across expenditure categories. Over time the distribution of the survey impacts on frequently purchased good fans out, that on regular billing items shrinks.

TABLE 1. Definitions of expenditure categories *** ADD IN NOTE THE DEFINITION OF D AND R GOODS? ***

Food and Non-Alcoholic Beverages at Home
Food and Non-Alcoholic Beverages Away from Home
Alcoholic Beverages (at home and away from home)
Non-Durable Goods and Services
Newspapers and Magazines
Non-durable Entertainment Expenses
Housekeeping Services (DS only)
Personal Care (DS only)
Housing and Public Services
Home Maintenance Services
Public Utilities
Miscellaneous Home Services
Tobacco and Smoking Accessories
Clothing, Footwear and Services
Clothing, Footwear
Services
Heating Fuel, Light and Power
Transportation (including gasoline)
Fuel for Transportation
Transportation Equipment Maintenance and Repair
Public Transportation
Vehicle Rental and Misc. Transportation Expenses

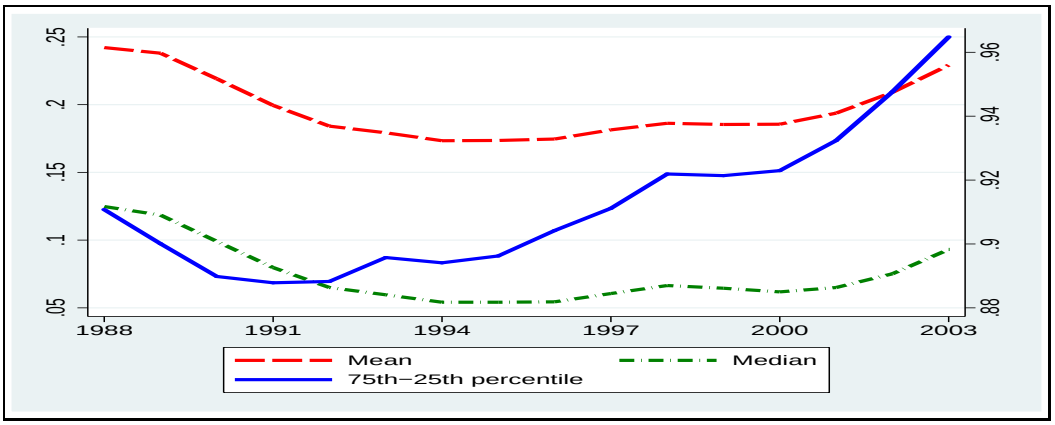
Note. The table reports in bold face the macro-categories that group the categories reported in normal face.

FIGURE 1. Distribution of slippages from rank invariance

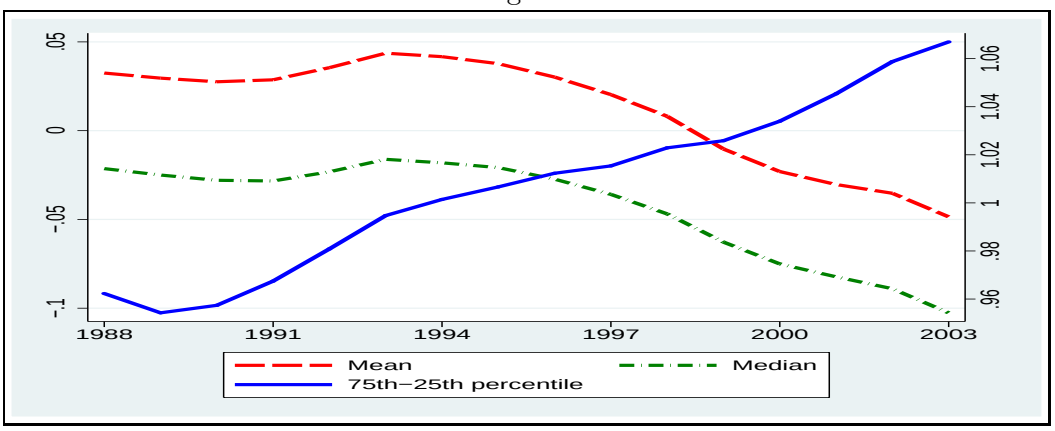
Note. Empirical distribution of the rank difference $U_1 - U_0$ for selected groups of years using Diary Survey data and conditional on selected values of the propensity score. The informational content of all other conditional distributions is similar to that in figure.

FIGURE 2. Mean, Median and Interquartile Range of the Effect of the Survey Instrument

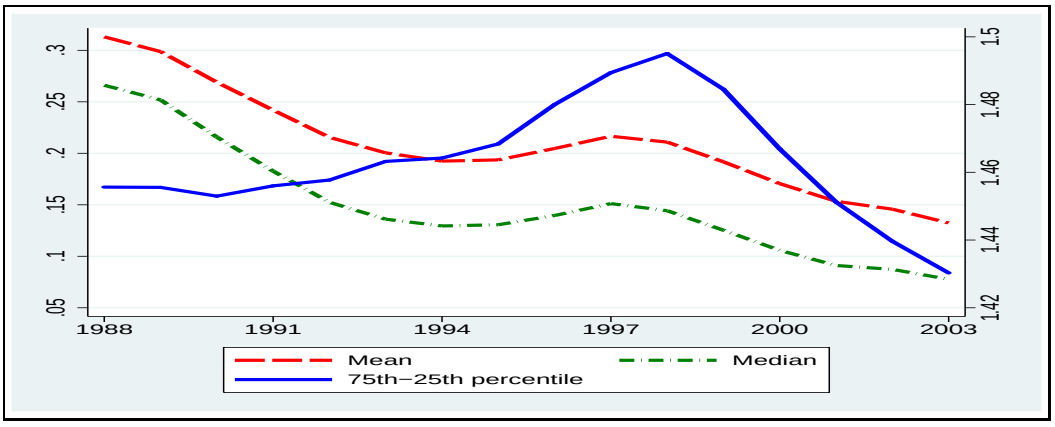
Food at Home



$\mathcal{D}$ goods



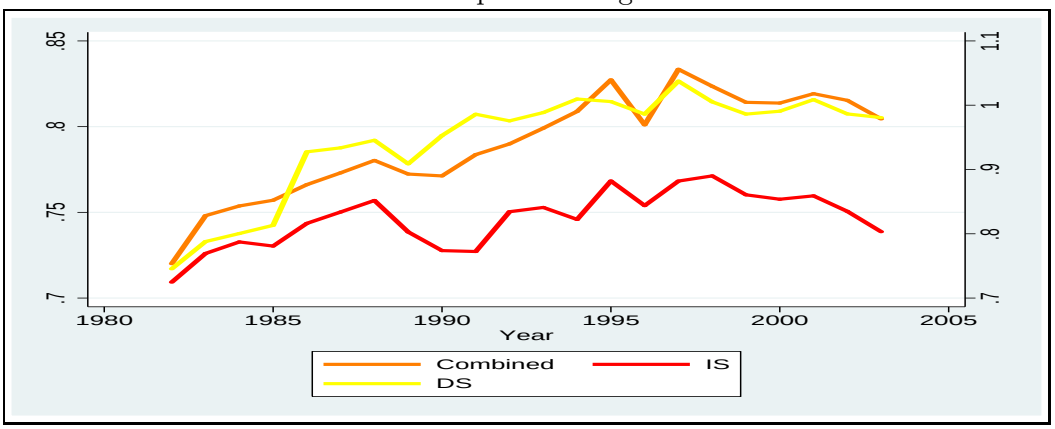
$\mathcal{R}$ goods



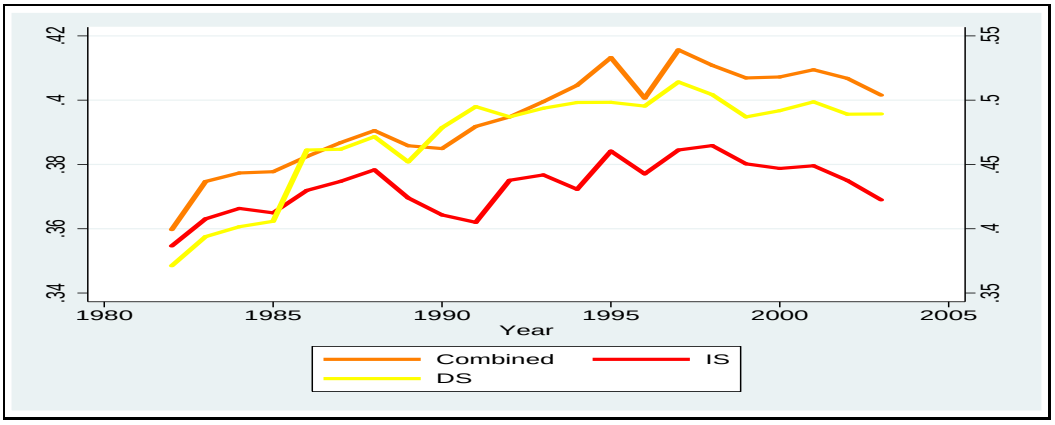
Note. Estimates of the mean, the median and the interquartile range of $F_{Y_1-Y_0}[\eta]$ for expenditure on food, $\mathcal{D}$ and $\mathcal{R}$ goods (see Table 1 for the definition of $\mathcal{D}$ and $\mathcal{R}$ goods).

FIGURE 3. Two measures of consumption inequality

Interquartile range



Median absolute deviation from the median



Note. Estimates of the interquartile range and the median absolute deviation from the median.

TABLE 2. Food at home: heterogeneity in the effect of the survey instrument

	1988-1990	1991-1995	1996-2000	2001-2003
Age				
≤ 35	-0.123	-0.112	-0.099	-0.067
	(0.019)***	(0.016)***	(0.016)***	(0.019)***
$(35-45]$	-0.047	-0.014	-0.046	-0.014
	(0.020)*	(0.016)	(0.016)**	(0.019)
$(45-55]$	-0.001	0.027	-0.018	-0.016
	(0.021)	(0.016)	(0.016)	(0.019)
Family type				
H/W only	-0.021	0.019	-0.004	-0.014
	(0.017)	(0.014)	(0.014)	(0.017)
H/W, oldest child 6-	-0.081	-0.086	-0.127	-0.144
	(0.020)***	(0.016)***	(0.017)***	(0.022)***
H/W, oldest child 6-17	-0.038	-0.044	-0.071	-0.105
	(0.016)*	(0.012)***	(0.012)***	(0.015)***
H/W, oldest child 18+	-0.064	-0.098	-0.138	-0.152
	(0.023)**	(0.018)***	(0.018)***	(0.021)***
All other H/W	-0.160	-0.174	-0.186	-0.205
	(0.028)***	(0.022)***	(0.022)***	(0.026)***
Ethnicity				
Black	-0.057	-0.059	-0.087	-0.081
	(0.019)**	(0.014)***	(0.014)***	(0.017)***
Other	-0.062	-0.008	-0.011	-0.015
	(0.029)*	(0.021)	(0.019)	(0.024)
Education				
High school graduate	-0.015	0.019	0.001	0.052
	(0.019)	(0.015)	(0.016)	(0.020)*
College dropout	0.010	0.031	0.002	0.056
	(0.020)	(0.016)	(0.017)	(0.021)**
At least college graduate	0.017	0.056	0.040	0.053
	(0.020)	(0.016)***	(0.016)*	(0.020)**
Income before taxes	0.060	0.058	0.060	0.066
	(0.008)***	(0.007)***	(0.006)***	(0.007)***
Z_0	-0.827	-0.843	-0.842	-0.845
	(0.007)***	(0.006)***	(0.006)***	(0.007)***
Constant	3.256	3.226	3.227	3.133
	(0.082)***	(0.066)***	(0.064)***	(0.075)***

Note. Results from a regression of $Y_1 - Y_0$ on demographics (age, education, family type, ethnicity) and expenditure as measured from diaries. Three stars indicate statistically significance at the 0.1% confidence level; two stars at 1% level; one star at the 5% level. The baseline groups for age, family type, ethnicity and education are, respectively, households with head aged more than 55, single households, households with white head, households with high-school drop out head.

TABLE 3. $\mathcal{D}$ goods: heterogeneity in the effect of the survey instrument

	1988-1990	1991-1995	1996-2000	2001-2003
Age				
≤ 35	-0.072	-0.073	-0.099	-0.096
	(0.006)***	(0.005)***	(0.005)***	(0.006)***
$(35-45]$	-0.026	-0.033	-0.040	-0.054
	(0.007)***	(0.005)***	(0.005)***	(0.006)***
$(45-55]$	-0.003	0.000	-0.011	-0.030
	(0.007)	(0.005)	(0.005)*	(0.006)***
Family type				
H/W only	-0.078	-0.051	-0.048	-0.043
	(0.006)***	(0.004)***	(0.004)***	(0.005)***
H/W, oldest child 6-	-0.289	-0.301	-0.278	-0.285
	(0.007)***	(0.006)***	(0.006)***	(0.007)***
H/W, oldest child 6-17	-0.278	-0.249	-0.247	-0.217
	(0.005)***	(0.004)***	(0.004)***	(0.005)***
H/W, oldest child 18+	-0.301	-0.301	-0.317	-0.278
	(0.007)***	(0.006)***	(0.005)***	(0.006)***
All other H/W	-0.400	-0.403	-0.411	-0.441
	(0.009)***	(0.007)***	(0.007)***	(0.008)***
Ethnicity				
Black	-0.199	-0.211	-0.198	-0.214
	(0.006)***	(0.005)***	(0.004)***	(0.005)***
Other	-0.105	-0.131	-0.121	-0.130
	(0.009)***	(0.007)***	(0.006)***	(0.007)***
Education				
High school graduate	0.071	0.072	0.100	0.086
	(0.006)***	(0.005)***	(0.005)***	(0.006)***
College dropout	0.131	0.134	0.154	0.139
	(0.006)***	(0.005)***	(0.005)***	(0.006)***
At least college graduate	0.203	0.228	0.245	0.220
	(0.006)***	(0.005)***	(0.005)***	(0.006)***
Income before taxes	0.213	0.209	0.194	0.196
	(0.003)***	(0.002)***	(0.002)***	(0.002)***
Y_0	-0.803	-0.810	-0.810	-0.813
	(0.007)***	(0.007)***	(0.007)***	(0.006)***
Constant	2.152	2.151	2.250	2.198
	(0.039)***	(0.036)***	(0.033)***	(0.032)***

Note. Results from a regression of $Y_1 - Y_0$ on demographics (age, education, family type, ethnicity) and expenditure as measured from diaries. The figures are obtained using information on Δ_j from pooled ($D = 0$ and $D = 1$) data, for $j = 1, \ldots, 50$ (as described in Section 6); reported are values of the coefficients averaged across the 50 draws. Standard errors of estimate are equal to the square root of the main diagonal elements of the matrix $T = (1 + 1/50)B + W$, where B is the variance of the estimated parameters between draws , and W is the average of the estimated variances for each draws. Three stars indicate statistically significance at the 0.1% confidence level; two stars at 1% level; one star at the 5% level. The baseline groups for age, family type, ethnicity and education are, respectively, households with head aged more than 55, single households, households with white head, households with high-school drop out head.

TABLE 4. $\mathcal{R}$ goods: heterogeneity in the effect of the survey instrument

	1988-1990	1991-1995	1996-2000	2001-2003
Age				
≤ 35	-0.117	-0.161	-0.165	-0.172
	(0.009)***	(0.007)***	(0.007)***	(0.008)***
$(35-45]$	-0.076	-0.124	-0.121	-0.146
	(0.009)***	(0.007)***	(0.007)***	(0.008)***
$(45-55]$	0.000	-0.037	-0.024	-0.055
	(0.009)	(0.007)***	(0.007)***	(0.008)***
Family type				
H/W only	-0.012	-0.006	0.018	0.025
	(0.008)	(0.006)	(0.006)**	(0.007)***
H/W, oldest child 6-	-0.291	-0.276	-0.246	-0.221
	(0.010)***	(0.008)***	(0.008)***	(0.010)***
H/W, oldest child 6-17	-0.345	-0.322	-0.286	-0.241
	(0.007)***	(0.006)***	(0.006)***	(0.007)***
H/W, oldest child 18+	-0.277	-0.270	-0.250	-0.208
	(0.010)***	(0.008)***	(0.008)***	(0.009)***
All other H/W	-0.430	-0.430	-0.390	-0.391
	(0.013)***	(0.010)***	(0.010)***	(0.011)***
Ethnicity				
Black	-0.125	-0.096	-0.053	-0.054
	(0.008)***	(0.006)***	(0.006)***	(0.007)***
Other	-0.217	-0.180	-0.139	-0.181
	(0.013)***	(0.010)***	(0.009)***	(0.010)***
Education				
High school graduate	0.183	0.197	0.220	0.226
	(0.009)***	(0.007)***	(0.007)***	(0.009)***
College dropout	0.298	0.298	0.321	0.306
	(0.009)***	(0.007)***	(0.008)***	(0.009)***
At least college graduate	0.395	0.404	0.405	0.382
	(0.009)***	(0.007)***	(0.007)***	(0.009)***
Income before taxes	0.306	0.282	0.260	0.261
	(0.004)***	(0.003)***	(0.003)***	(0.003)***
Y_0	-0.923	-0.932	-0.923	-0.931
	(0.004)***	(0.004)***	(0.005)***	(0.004)***
Constant	1.827	2.068	2.147	2.155
	(0.039)***	(0.032)***	(0.033)***	(0.033)***

Note. Results from a regression of $Y_1 - Y_0$ on demographics (age, education, family type, ethnicity) and expenditure as measured from diaries. The figures are obtained using information on Δ_j from pooled ($D = 0$ and $D = 1$) data, for $j = 1, \ldots, 50$ (as described in Section 6); reported are values of the coefficients averaged across the 50 draws. Standard errors of estimate are equal to the square root of the main diagonal elements of the matrix $T = (1 + 1/50)B + W$, where B is the variance of the estimated parameters between draws , and W is the average of the estimated variances for each draws. Three stars indicate statistically significance at the 0.1% confidence level; two stars at 1% level; one star at the 5% level. The baseline groups for age, family type, ethnicity and education are, respectively, households with head aged more than 55, single households, households with white head, households with high-school drop out head.

TABLE 5. Food at home: probability of recall overstating diaries

	1988-1990	1991-1995	1996-2000	2001-2003
Age				
≤ 35	-0.038	-0.068	-0.054	-0.022
	(0.017)*	(0.014)***	(0.014)***	(0.017)
$(35-45]$	0.009	-0.021	-0.024	-0.013
	(0.018)	(0.014)	(0.014)	(0.017)
$(45-55]$	0.019	-0.007	-0.019	-0.003
	(0.018)	(0.015)	(0.014)	(0.017)
Family type				
H/W only	-0.010	0.011	-0.012	0.012
	(0.015)	(0.012)	(0.012)	(0.015)
H/W, oldest child 6-	-0.045	-0.046	-0.060	-0.074
	(0.018)*	(0.015)**	(0.015)***	(0.019)***
H/W, oldest child 6-17	-0.003	-0.006	-0.011	-0.029
	(0.014)	(0.011)	(0.011)	(0.013)*
H/W, oldest child 18+	-0.005	-0.042	-0.067	-0.053
	(0.020)	(0.016)**	(0.016)***	(0.019)**
All other H/W	-0.046	-0.050	-0.101	-0.099
	(0.025)	(0.020)*	(0.019)***	(0.023)***
Ethnicity				
Black	-0.043	-0.030	-0.050	-0.031
	(0.017)**	(0.013)*	(0.013)***	(0.015)*
Other	-0.019	0.017	-0.003	0.011
	(0.025)	(0.019)	(0.017)	(0.021)
Education				
High school graduate	-0.033	0.006	-0.028	0.008
	(0.017)	(0.014)	(0.014)	(0.018)
College dropout	-0.020	-0.001	-0.016	0.020
	(0.017)	(0.014)	(0.015)	(0.019)
At least college graduate	-0.019	0.019	-0.007	0.014
	(0.018)	(0.015)	(0.014)	(0.018)
Income before taxes	0.028	0.021	0.026	0.031
	(0.007)***	(0.006)***	(0.005)***	(0.006)***
Z_0	-0.325	-0.346	-0.332	-0.329
	(0.006)***	(0.005)***	(0.005)***	(0.006)***
Constant	1.735	1.839	1.754	1.636
	(0.073)***	(0.060)***	(0.055)***	(0.065)***

Note. Results from a regression of $\mathbf{1}\{Y_1 - Y_0 > 0\}$ on demographics (age, education, family type, ethnicity) and expenditure as measured from diaries. Three stars indicate statistically significance at the 0.1% confidence level; two stars at 1% level; one star at the 5% level. The baseline groups for age, family type, ethnicity and education are, respectively, households with head aged more than 55, single households, households with white head, households with high-school drop out head.

TABLE 6. $\mathcal{D}$ goods: probability of recall overstating diaries

	1988-1990	1991-1995	1996-2000	2001-2003
Age				
≤ 35	-0.039	-0.038	-0.050	-0.046
	(0.006)***	(0.005)***	(0.005)***	(0.005)***
$(35-45]$	-0.013	-0.016	-0.019	-0.024
	(0.007)*	(0.005)***	(0.005)***	(0.005)***
$(45-55]$	-0.001	0.002	-0.004	-0.010
	(0.007)	(0.005)	(0.005)	(0.005)*
Family type				
H/W only	-0.046	-0.028	-0.027	-0.024
	(0.006)***	(0.004)***	(0.004)***	(0.004)***
H/W, oldest child 6-	-0.162	-0.160	-0.147	-0.146
	(0.007)***	(0.005)***	(0.005)***	(0.006)***
H/W, oldest child 6-17	-0.158	-0.136	-0.130	-0.108
	(0.006)***	(0.004)***	(0.004)***	(0.004)***
H/W, oldest child 18+	-0.168	-0.162	-0.163	-0.142
	(0.007)***	(0.006)***	(0.005)***	(0.006)***
All other H/W	-0.222	-0.208	-0.208	-0.212
	(0.010)***	(0.007)***	(0.006)***	(0.007)***
Ethinicity				
Black	-0.104	-0.106	-0.098	-0.105
	(0.006)***	(0.004)***	(0.004)***	(0.005)***
Other	-0.055	-0.064	-0.054	-0.064
	(0.009)***	(0.006)***	(0.006)***	(0.007)***
Education				
High school graduate	0.035	0.035	0.043	0.039
	(0.006)***	(0.005)***	(0.005)***	(0.006)***
College dropout	0.066	0.066	0.073	0.065
	(0.006)***	(0.005)***	(0.005)***	(0.006)***
At least college graduate	0.106	0.115	0.119	0.105
	(0.006)***	(0.005)***	(0.005)***	(0.006)***
Income before taxes	0.115	0.109	0.098	0.094
	(0.003)***	(0.003)***	(0.002)***	(0.002)***
Y_0	-0.393	-0.381	-0.376	-0.370
	(0.007)***	(0.006)***	(0.006)***	(0.005)***
Constant	1.442	1.393	1.455	1.438
	(0.041)***	(0.035)***	(0.033)***	(0.033)***

Note. Results from a regression of $\mathbf{1}\{Y_1 - Y_0 > 0\}$ on demographics (age, education, family type, ethnicity) and expenditure as measured from diaries. The figures are obtained using information on Δ_j from pooled ($D = 0$ and $D = 1$) data, for $j = 1, \ldots, 50$ (as described in Section 6); reported are values of the coefficients averaged across the 50 draws. Standard errors of estimate are equal to the square root of the main diagonal elements of the matrix $T = (1+1/50)B+W$, where B is the variance of the estimated parameters between draws , and W is the average of the estimated variances for each draws. Three stars indicate statistically significance at the 0.1% confidence level; two stars at 1% level; one star at the 5% level. The baseline groups for age, family type, ethnicity and education are, respectively, households with head aged more than 55, single households, households with white head, households with high-school drop out head.

TABLE 7. $\mathcal{R}$ goods: probability of recall overstating diaries

	1988-1990	1991-1995	1996-2000	2001-2003
Age				
≤ 35	-0.030	-0.038	-0.043	-0.042
	(0.007)***	(0.006)***	(0.005)***	(0.006)***
$(35-45]$	-0.020	-0.029	-0.033	-0.036
	(0.008)**	(0.006)***	(0.006)***	(0.006)***
$(45-55]$	-0.001	-0.007	-0.006	-0.012
	(0.008)	(0.006)	(0.006)	(0.006)*
Family type				
H/W only	-0.003	-0.006	0.005	0.004
	(0.006)	(0.005)	(0.005)	(0.005)
H/W, oldest child 6-	-0.076	-0.079	-0.070	-0.062
	(0.008)***	(0.007)***	(0.006)***	(0.008)***
H/W, oldest child 6-17	-0.090	-0.090	-0.079	-0.066
	(0.006)***	(0.005)***	(0.004)***	(0.005)***
H/W, oldest child 18+	-0.076	-0.075	-0.063	-0.049
	(0.008)***	(0.006)***	(0.006)***	(0.007)***
All other H/W	-0.111	-0.112	-0.104	-0.102
	(0.011)***	(0.009)***	(0.008)***	(0.008)***
Etnicity				
Black	-0.040	-0.023	-0.007	-0.010
	(0.007)***	(0.005)***	(0.005)	(0.005)*
Other	-0.049	-0.047	-0.030	-0.047
	(0.010)***	(0.008)***	(0.007)***	(0.008)***
Education				
High school graduate	0.044	0.042	0.052	0.059
	(0.007)***	(0.005)***	(0.006)***	(0.006)***
College dropout	0.070	0.068	0.074	0.078
	(0.007)***	(0.006)***	(0.006)***	(0.007)***
At least college graduate	0.098	0.098	0.098	0.098
	(0.007)***	(0.006)***	(0.006)***	(0.006)***
Income before taxes	0.079	0.071	0.067	0.065
	(0.004)***	(0.003)***	(0.002)***	(0.002)***
Y_0	-0.102	-0.098	-0.106	-0.095
	(0.006)***	(0.006)***	(0.006)***	(0.005)***
Constant	0.306	0.323	0.390	0.331
	(0.035)***	(0.031)***	(0.029)***	(0.030)***

Note. Results from a regression of $\mathbf{1}\{Y_1 - Y_0 > 0\}$ on demographics (age, education, family type, ethnicity) and expenditure as measured from diaries. The figures are obtained using information on Δ_j from pooled ($D = 0$ and $D = 1$) data, for $j = 1, \ldots, 50$ (as described in Section 6); reported are values of the coefficients averaged across the 50 draws. Standard errors of estimate are equal to the square root of the main diagonal elements of the matrix $T = (1+1/50)B+W$, where B is the variance of the estimated parameters between draws , and W is the average of the estimated variances for each draws. Three stars indicate statistically significance at the 0.1% confidence level; two stars at 1% level; one star at the 5% level. The baseline groups for age, family type, ethnicity and education are, respectively, households with head aged more than 55, single households, households with white head, households with high-school drop out head.

References

Attanasio, Orazio and Guglielmo Weber, "Consumption Growth, the Interest Rate and Aggregation," *Review of Economic Studies*, 1993, *60*, 631–649.

___ and ___, "Is Consumption Growth Consistent with Intertemporal Optimization? Evidence from the Consumer Expenditure Survey," *Journal of Political Economy*, 1995, *103*, 1121–1157.

Attanasio, Orazio P., Erich Battistin, and Hidehiko Ichimura, "What Really Happened To Consumption Inequality in the US," in E. Berndt and C. Hulten, eds., *Hard-to-Measure Goods and Services: Essays in Honor of Zvi Griliches*, University of Chicago Press, 2007.

___, ___, and Mario Padula, "Inequality in Living Standards since 1980:Evidence from Expenditure Data," 2008. Mimeo.

Banks, James and Paul Johnson, *How Reliable is the Family Expenditure Survey? Trends in Incomes and Expenditures over Time*, The Institute for Fiscal Studies, London, 1998.

Battistin, Erich, "Errors in Survey Reports of Consumption Expenditures," 2003. Working Paper W03/07, Institute for Fiscal Studies, London.

___, Raffaele Miniaci, and Guglielmo Weber, "What Can We Learn From Recall Consumption Data?," *Journal of Human Resources*, 2003, *38* (2), 354–385.

Blundell, Richard, Luigi Pistaferri, and Ian Preston, "Consumption inequality and partial insurance," *American Economic Review*, December 2008, *98* (5), 1887–1921.

Browning, Martin and Thomas F. Crossley, "Are Two Cheap, Noisy Measures Better Than One Expensive, Accurate One?," *American Economic Review: Papers and Proceedings*, 2009, *99* (2), 1–9.

____ , ____ , and Guglielmo Weber, "Asking consumption questions in general purpose surveys," *Economic Journal*, November 2003, *113* (491), F540–F567.

Chernozhukov, Victor and Christian B. Hansen, "An IV model of quantile treatment effects," *Econometrica*, 2005, *73* (1), 245–261.

____ and ____ , "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 2006, *132* (2), 491–525.

Firpo, Sergio and Geert Ridder, "Bounds on Functionals of the Distribution of Treatment Effects," IEPR Working Paper 08-09, Institute for Economic Policy Research 2008.

Garner, Thesia, Robert McClelland, and William Passero, "Strengths and Weaknesses of the Consumer Expenditure Survey from a BLS Perspective," 2009. NBER/CRIW presentation.

Heckman, James J., Jeffrey Smith, and Nancy Clements, "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, October 1997, *64* (4), 487–535.

Kerwin, Kofi Charles, Sheldon Danziger, Geng Li, and Robert F Schoeni, "Studying Consumption with the Panel Study of Income Dynamics: Comparison with the Consmer Expenditure Survey and an Application to the Intergenerational Trasmission of Well-being," 2006. Population Studies center, Report 06 - 590.

Krueger, Dirk and Fabrizio Perri.

Lyberg, Lars E. and Daniel Kasprzyk, *Measurement Errors in Surveys*, John Wiley and Sons: New York, 1991.

Meyer, Bruce D. and James X. Sullivan, "Five Decades of Consumption and Income Poverty," NBER Working Papers 14827, National Bureau of Economic Research, Inc March 2009.

Neter, John and Joseph Waksberg, "A Study of Response Errors in Expenditures Data from Household Interviews," *Journal of the American Statistical Association*, 1964, *59* (305), 18–55.

Pesarin, Fortunato, *Multivariate Permutation Tests. With Applications in Biostatistics*, John Wiley & Sons: Chichester, 2001.

Rosenbaum, Paul and Donald Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 1983, *70* (1), 41–55.

Silberstein, Adriana R. and Stuart Scott, *Expenditure Diary Surveys and Their Associated Errors*, John Wiley and Sons: New York, 1991.

Slesnick, Daniel T., "Empirical Approaches to the Measurement of Welfare," *Journal of Economic Literature*, December 1998, *36* (4), 2108–2165.

____ , *Consumption and Social Welfare: Living Standards and Their Distribution in the United States*, Cambridge University Press, 2001.

Tucker, Clyde, "The Estimation of Instrument Effects on Data Quality in the Consumer Expenditure Survey," *Journal of Official Statistics*, 1992, *8*, 41–61.

Wilcox, David W., "The Construction of U.S. Consumption Data: Some Facts and Their Implications for Empirical work," *American Economic Review*, 1992, *82* (4), 922–941.

Winter, Joachim K., "Design effects in survey-based measures of household consumption," 2002. Discussion Paper No. 02-34, University of Mannheim.

Appendix A: Sample selection

Table A-1 shows the observations lost at each selection step for DS and IS data, while the final size of the two samples in each survey year is reported in Table A-2. After selection, the DS and IS align well along various dimensions. Table A-3 shows the age, family type, ethnicity and education frequencies in the DS and IS for various years. In the DS the percentage of households with heads aged less than 35 ranges from 42 (in the years 1982-1987) to 29 percent (2001-2003); in the IS from 41 (1982-1987) to 29 percent (1001-2003). The percentage of households with heads aged between 35 and 45, 45 and 55, and older than 55 is also very similar across the two surveys. Differences in the sample composition show up for the type of family, ethnicity and education. Singles and Husband and Wife households with oldest child aged 18+ are more frequent in the IS than in the DS, and Husband and Wife households more in the DS than in the IS in the years 1982-1987 and 1988-1990. As for the ethnicity, black heads are more prevalent in the IS in the years 1988-1990 and 1991-1995 (from 11 to 12 percent) than in the DS (10 percent). The percentage of households with high school dropout heads is higher in the IS that of at least college graduate in the DS for the years 1988-1990 and 1991-1995.

Appendix B: Food expenditure in the CEX

The aim of this section is to document changes in the DS and IS questionnaires that occurred over the period $1982 - 2003$ and are relevant to the measurement of food expenditure.[24]

Summary

The main conclusions from this analysis can be summarized as follows. First, for the IS we can distinguish two time periods in our data characterized by homogeneous recall questions on food: 1982-1987 and 1988-2003. The difference between these two groups of years is in the recall period: *usual monthly* expenditure in the former group, and *usual weekly* expenditure in the latter group. Second, for the DS we can again distinguish between 1982-1987 and 1988-2003, though the difference

[24] We thank Thesia Garner, David Johnson and Bill Passero who generously helped us clarify this point.

between the two groups of years is not simply limited to the period recalled. Third, the wording of the recall questions after 1988 is common for DS and IS respondents, thus implying that in our data households in the two surveys were interviewed using the same survey instrument for the years 1988-2003.

The effects of these changes on the reporting of food expenditure were first documented by Battistin (2003), and can be seen in Figure A-1 and Figure A-2. The break in 1988 affects both the location and the scale of the food at grocery distribution. Using the median and the median absolute deviation as robust measures of location and scale, respectively, Figure A-1 shows that the former increases in both the DS and IS after the change, while the latter decreases in the DS, and increases in the IS. Figure A-2 shows a similar pattern for food at home in the IS (right-hand side of the panel), since it accounts for a substantial fraction (above 90 percent) of the food at grocery (left-hand side of the panel).

The measurement of food expenditure in the CEX

The first questionnaire of the Interview continuing surveys (1979) included two set of questions. The first set comprised the following questions: (1a) Since the 1st of (month, 3 months ago), how often have you and other members of your CU shopped at the grocery store? (1b) What was the *usual* amount of your purchase *per visit*? (1c) About how much of this amount was for food and nonalcoholic beverages? The second set of questions referred to places other than grocery stores and also asks the *usual amount spent* for these foods and beverages *per visit*.

The next questionnaire (1982) instituted many changes in either set of questions: (1a) Since the 1st of (month, 3 months ago), what has been your *usual monthly* expense at the grocery store or supermarket? (1b) About how much of this amount was for *non food* items, such as paper products, detergents, home cleaning supplies, pet foods and alcoholic beverages? The second set of questions referred to places other than grocery stores and also asked the *usual monthly* expense at these places.

The next version of the questionnaire that made changes relevant to the reports of food expenditure

was in 1988. Changes referred to the recall period mentioned in either set of questions, asking for *usual weekly* instead of *usual monthly* expenditures. In particular, *usual weekly* expense at the grocery store or supermarket was asked in the first set of questions, and *usual weekly* expense at places other than grocery stores in the second set of questions. The survey questionnaire remained unchanged since then.

As for the Diary surveys, other than reporting all expenditures for food items occurred in the two-week diary, respondents are asked two sets of recall questions that are almost identical to those in the interview questionnaire. As for the 1980 questionnaire, the set of questions was as follows: (1a) Since the 1st of (month, 3 months ago), have you and other members of your CU shopped at the grocery store? (Monthly, Weekly, Never) followed by: How many times per (week, month) did you shop at the grocery store? (1b) What was the *usual* amount of your purchase *per visit*? (1c) About how much of this amount was for food and nonalcoholic beverages?

The questionnaire in April 1982 added a few changes: (1b) About how much of this amount was for *non food* items, such as paper products, detergents, home cleaning supplies, pet foods and alcoholic beverages?

These questions remain unchanged until January 1988 when the following changes were made: (1a) (now 3a) Since the 1st of (month, 3 months ago), what was your *usual weekly* expense at the grocery store or supermarket? (1b) (now 3b) About how much of this amount was for *non food* items, such as paper products, detergents, home cleaning supplies, pet foods and alcoholic beverages? (1c) (now 3c) Have you (or any members of your CU) purchased any food or nonalcoholic beverages from places other than grocery stores, such as home delivery, specialty stores, bakeries, convenience stores, dairy stores, vegetable stands, or farmers markets? (1d) What was your usual *weekly* expense at these places? There have been no changes since then.

The definition of food expenditure used in the analysis

*** CHECK THIS *** We use three measures of food expenditure, a diary and two recall measures from the DS and the IS. The diary measure refers to all food expenditure items bought over two weeks. The recall question in the DS refers to food items bought at grocery stores, and excludes those bought at convenience stores, specialty stores, bakeries, home delivery, vegetable stands, or farmers markets. These account for between 6 and 8 percent of food expenditure at home in the IS, as shown in the left-hand panel of Figure A-2. The percentage is rather stable over time, but it has a break in 1988.

Appendix C: Estimation

Estimation of the propensity score

We balanced the distribution of household characteristics that are observable in the two surveys of the CEX exploiting the properties of the propensity score $e(X)$. The estimation of $e(X)$ was carried out separately by expenditure year specifying a probit regression in which the dependent variable is zero for DS and one for IS households, and the independent variables are a full set of family type, ethnicity, education dummies as well as a full set of month dummies. The value of the propensity score was calculated by using predictions from these regressions.[25]

Predictions were then stratified into groups defined by expenditure year and percentiles of the distribution of the score for DS observations in each year. We made this choice mainly for convenience, as this allowed us to guarantee a reasonable number of observations from the DS sample across all strata. The number of strata was selected so to ensure the same distribution of the X's across surveys within each stratum, that is by ensuring that the balancing property of the propensity score (see Rosenbaum and Rubin, 1983) was satisfied in the data. We tested for this condition by running the same probit regression considered above within each stratum, and by looking at the F statistic for

[25] We experimented with different specifications of the propensity score, including as additional regressors region of residence dummies, and quadratic polynomials in age, family size and number of dependent children. However, the latter set of variables turned out not statistically significant in all sets of regressions considered, and we found that their inclusion in the specification worsened considerably the balancing properties within strata that are discussed in what follows.

the joint significance of all X's included.

By doing so, we found that 15 strata in each expenditure year (22 overall, from 1982 to 2003) were a reasonable compromise between sample size and values of the F statistics. We found p-values associated to the F statistics smaller than 5% only in 16 out of the $15 \times 22 = 330$ strata considered, the average p-value being around 50%. The average number of households across strata was 130 and $2,018$ for the DS sample and the IS sample, respectively. Table A-4 and Table A-5 report the size of the strata for each survey year in the DS and the IS. Moreover, we did not find common support problems in the two surveys, in the sense that the distributions of predicted values of $e(X)$ overlapped considerably in all survey years. We found the propensity score estimates rather stable over survey years, revealing differences between samples as for type of family (in the survey years 1982, 1985-2003), ethnicity (1985-1988, 1990, 1998) and education (1984-1985, 1987-1988, 1992, 1995, 2003).

Testing the strong ignorability condition (Section 5.1)

The aim of this section is to describe how we tested for:

$$H_0 : \ F_{Y_1|D,e(X)}[\eta|0,e] = F_{Y_1|D,e(X)}[\eta|1,e]. \tag{5}$$

This amounts to testing the hypothesis that two samples defined from independent surveys are from populations with the same distribution. We tested this condition separately for each of the 15 strata defined from predicted values of propensity score obtained as described in the previous section. Thus, within each stratum we tested for the independence between D and Y_1. To this end, we used two test statistics that were defined as follows.[26]

1. First, we grouped observations into 4 categories defined by quartiles of the distribution of Y_1 and regressed the resulting categorical variable on the survey dummy D using a ordered probit regression. As additional regressors, we included polynomial terms in the propensity score

[26] We also experimented with variants of propensity score matching, by matching DS to IS households on the estimated value of $e(X)$. We considered the sample means of Y_1 and $\mathbb{1}(Y_1 \leq k)$, with k suitably chosen grid points on the support of Y_1, before and after matching, and found that these were statistically the same in DS and IS in almost all cases after matching.

to adjust for residual within stratum heterogeneity. We then took the regression coefficient associated to D as a test statistic for H_0.

2. Second, we considered a Mann-Whitney test for H_0.

We obtained the p-values of either statistic calculating its distribution under the null through re-sampling methods (see Pesarin (2001)). In particular, since (5) implies exchangeability of observed expenditures with respect to D, we randomly permuted values of the variable D within each stratum and calculated the p-value as:

$$p \equiv Pr_{H_0}\{|T^*| \geq |T_{obs}|\},$$

where T_{obs} is the observed value of the statistic and T^* are pseudo-values calculated from 100 permutations.

However, these p-values do not take into account the fact that stratification is based on the estimated propensity score. We thus repeated the procedure to compute p by bootstrapping the original working sample 100 times, and computed the values p_j, with $j = 1, \ldots, 100$, corresponding to each bootstrapped sample. Table A-6 and Table A-7 in the text report the average of these bootstrapped p-values:

$$\frac{1}{100}\sum_{j=1}^{100} p_j,$$

separately for the two test statistics considered. For both statistics, the average p-values never fall below standard significance level. Table A-8 and Table A-9 provide the 25th, the 50th, and 75th percentiles for the distribution of the p_j's in the years 1982-1987, 1988-1990, 1991-1995, 1996-2000 and 2001-2003. For both statistics, rejection is more likely in the years 1982-1987, in line with changes occurred from January 1988 to the recall questions in the DS.

Testing the rank invariance condition (Section 5.2)

*** CHECK THIS *** The aim of this section is to describe how we tested for:

$$H_0 : U_1 = U_0, \tag{6}$$

within strata defined from $e(X)$. Information on food expenditure available in the DS allowed us to identify $F_{Y_0|D,e(X)}[\eta|0,e]$ and $F_{Y_1|D,e(X)}[\eta|0,e]$, that is the distributions of either potential measurement corresponding to diaries (Y_0) and recall questions (Y_1) for DS households. We defined ranks from these two distributions as:

$$U_0 \equiv F_{Y_0|D,e(X)}[Y_0|0,e], \qquad U_1 \equiv F_{Y_1|D,e(X)}[Y_1|0,e],$$

and computed their difference $U_1 - U_0$ across strata. Under the null hypothesis (6), the distribution of this difference should be degenerate at zero. *** CHECK THIS *** In the left hand side panels of Figure 1 we report the distribution of $U_1 - U_0$ pooling observations from all strata for the years $1982 - 1987$ and $2001 - 2003$, while in the right hand side panels of the same figure we report the distribution of slippages given U_0. The two year groups have been chosen as they present differences in the wording of the survey questions for food (see Appendix B), though the results depicted are also representative of those for the years excluded. Figures referring to the same distributions within propensity score strata provided qualitatively similar information, and are not reported for brevity. The evidence provided clearly points to the presence of slippages from rank invariance, though the mode of the distribution is centered at zero (this value implying that the household has the same rank in the distributions of Y_0 and Y_1).

Testing (6) amounts to testing the hypothesis that the bivariate distribution (Y_0, Y_1) can be fully retrieved from knowledge of the marginal distributions $F_{Y_0|D,e(X)}[\eta|0,e]$ and $F_{Y_1|D,e(X)}[\eta|0,e]$. For example, one may obtain values of Y_1 under the null using:

$$Y_{1,H_0} \equiv F^{-1}_{Y_1|D,e(X)}[F_{Y_0|D,e(X)}[Y_0|0,e]|0,e],$$

and then look at the distance between the raw distribution (Y_0, Y_1) and the implied distribution (Y_0, Y_{1,H_0}). Instead of running a test that involves bivariate distributions, we tested for the whether the correlation between Y_1 and Y_0 is equal to that between Y_{1,H_0} and Y_0. If rejected in the data, this would also be sufficient to reject (6).

Fitting the distribution of slippages (Section 5.2)

The distribution of slippages $V \equiv U_1 - U_0$ can not be independent of that of U_0. This follows from the fact that the random variables describing ranks have bounded support, and values of U_0 close to zero (one) must imply that the distribution of V has a heavy right tail (left tail, respectively). It therefore follows that, under Assumption 2, knowledge of the distributions:

$$F_{Y_0|D,e(X)}[\eta|0,e], \qquad F_{Y_1|D,e(X)}[\eta|0,e], \qquad F_{V|U_0,D,e(X)}[\eta|u_0,0,e],$$

allows to retrieve the joint distribution of Y_0 and Y_1 given $e(X)$. As we have already discussed, all these distributions are (non-parametrically) identified in the data using repeated measurements of food expenditure.

The aim of this section is to describe how we estimated $F_{V|U_0,D,e(X)}[\eta|u_0,0,e]$. First, notice that this distribution can be derived from that of U_1 and U_0, and that standard calculations yield:

$$F_{V|U_0,D,e(X)}[\eta|u_0,0,e] = F_{U_1|U_0,D,e(X)}[\eta + u_0|u_0,0,e],$$

with $\eta \in (-u_0, 1-u_0)$. To ease computation of the estimation steps, we modeled the distribution of U_1 given U_0 parametrically by fitting different Beta distributions across strata defined from values of the propensity score $e(X)$. We allow both the shape parameters to depend linearly on U_0.[27] *** CHECK THIS *** Figure A-3 shows the estimated distribution of slippages given U_0 that resulted from this procedure. Mean and location of the estimated conditional distribution of slippages change with U_0. When U_0 is equal to zero, the distribution of slippages ranges form 0 to 1; when U_0 is equal to

[27]Results using second and third order polynomials are very similar, but less precise, and so we opted for the most parsimonious model. As an alternative to the Beta distribution, we experimented with the Skew Normal family of distributions coming out with qualitatively similar results.

one, from -1 to 0. The shape parameters are precisely estimated for most survey years and propensity score stratum (results are available upon request).

Appendix D: Additional results

Figures A-4, A-5 and A-6 show that the marginal distribution of food changes with the survey instrument and so do the marginal distributions of $\mathcal{D}$ and $\mathcal{R}$ goods (left-hand side panels). This calls for analyzing the effect of the survey instrument on consumption report, which we do in the right-hand side panels of the same figures. In line with previous finding (see Attanasio *et al.* (2007) and Attanasio *et al.* (2008)), the figures points to sizable effects of the survey instrument for all years.

FIGURE A-1. Effect of survey changes on measured food at grocery

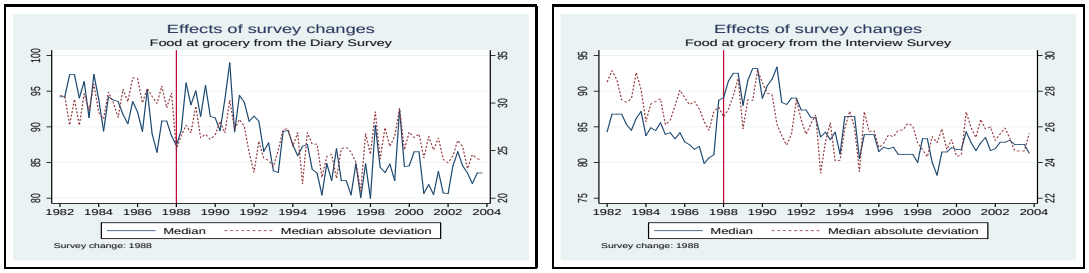


FIGURE A-2. Effect of survey changes on measured food at home

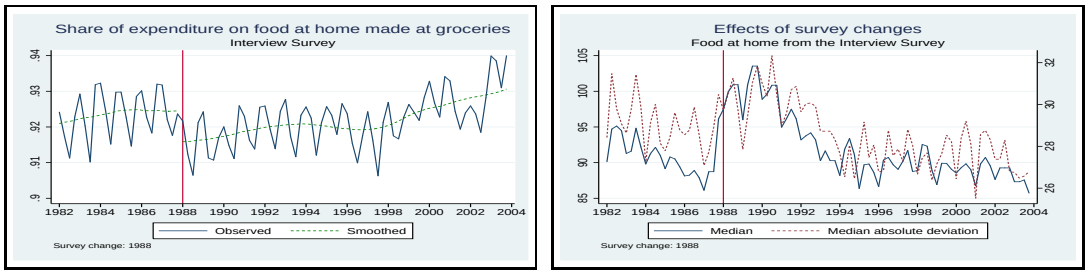


FIGURE A-3. The estimated distribution of slippages conditional U_0 *** CHECK LABELS IN FIGURES ***

Years 1982-1987 Years 2001-2003

TABLE A-1. Sample selection gradient by survey instrument

Sample size before selecting out	Diary sample	Interview sample
Households with incomplete income response	141,061	1,529,483
Non-urban households	109,166	1,274,674
Household heads aged less than 25 and more than 65	98,380	1,150,827
Self-employed household head	71,486	835,453
Final sample	57,608	670,292

Note. The table reports the size of the diary and interview samples before each selection criterium is applied to the data. The last row reports the final sample size after all selections. The first row reports the size of samples after dropping the observations with missing expenditure for food at home and food away from home.

TABLE A-2. Sample size by survey year

Year	Diary	Interview
1982	2,450	26,197
1983	2,469	26,895
1984	2,481	28,532
1985	2,478	30,687
1986	2,809	33,567
1987	2,897	29,904
1988	2,513	27,627
1989	2,673	27,932
1990	2,792	27,939
1991	2,709	27,278
1992	2,664	27,530
1993	2,533	27,899
1994	2,298	27,154
1995	2,111	26,571
1996	2,421	27,356
1997	2,468	27,617
1998	2,153	28,071
1999	2,857	37,058
2000	2,942	37,176
2001	2,900	39,408
2002	3,005	41,359
2003	2,985	36,535

Note. The table reports the size of the sample by survey year and instrument.

TABLE A-3. Summmary statistics

	1982-1987		1988-1990		1991-1995		1996-2000		2001-2003	
	Diary	Interview	Diary	Interview	Diary	Interview	Diary	Interview	Diary	Interview
Age										
≤ 35	0.419	0.405	0.388	0.378	0.360	0.348	0.321	0.324	0.294	0.292
$(35-45]$	0.271	0.276	0.306	0.310	0.329	0.323	0.329	0.327	0.323	0.315
$(45-55]$	0.176	0.183	0.188	0.197	0.202	0.215	0.238	0.240	0.257	0.260
> 55	0.134	0.136	0.118	0.116	0.108	0.114	0.112	0.109	0.125	0.133
Family type										
H/W only	0.172	0.157	0.174	0.157	0.163	0.156	0.158	0.158	0.160	0.157
H/W, oldest child 6-	0.106	0.095	0.102	0.093	0.096	0.087	0.082	0.077	0.076	0.073
H/W, oldest child 6-17	0.218	0.216	0.215	0.211	0.222	0.210	0.217	0.208	0.209	0.198
H/W, oldest child 18+	0.090	0.107	0.079	0.098	0.081	0.088	0.078	0.084	0.082	0.087
All other H/W	0.043	0.047	0.044	0.046	0.045	0.049	0.049	0.048	0.050	0.051
Other households	0.371	0.377	0.384	0.396	0.393	0.411	0.415	0.425	0.422	0.432
Ethnicity										
White	0.859	0.858	0.860	0.850	0.851	0.840	0.830	0.830	0.824	0.821
Black	0.104	0.103	0.101	0.110	0.106	0.117	0.113	0.118	0.122	0.120
Other	0.037	0.039	0.038	0.040	0.043	0.043	0.057	0.052	0.055	0.059
Education										
High school dropout	0.153	0.156	0.126	0.136	0.108	0.120	0.103	0.109	0.099	0.105
High school graduate	0.311	0.308	0.308	0.309	0.301	0.300	0.264	0.267	0.250	0.260
College dropout	0.248	0.240	0.257	0.260	0.262	0.258	0.197	0.198	0.197	0.191
At least college graduate	0.288	0.296	0.309	0.295	0.328	0.322	0.435	0.425	0.454	0.444

Note. The table reports the sample means by survey year and instrument for four age dummies defined on the basis of whether the age of the head is less than 35, between 35 and 45, between 45 and 55 and more than 55; six family type dummies defined on the basis of whether the respondent household is husband and wife household, is a husband and wife household with oldest child less than 6 years old, is a husband and wife household with oldest child between 6 and 17 years old, is a husband and wife household with oldest child elder than 18 years, is any other husband and wife household, and a residual category containing all other households; three ethnicity dummies, for black, white and other; four education dummies defined on the basis of whether the head of the household is a high-school drop-out, a high-school graduate, a college drop-out and at least a college graduate.

TABLE A-4. Sample size by propensity score stratum (Diary Survey)

	Stratum														
Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1982	127	120	123	128	125	122	118	127	118	125	123	123	122	122	123
1983	130	119	123	132	118	129	119	128	121	129	117	123	124	126	122
1984	129	137	120	129	128	130	128	129	127	133	124	129	129	143	113
1985	123	134	111	123	123	121	129	118	119	124	122	128	115	126	118
1986	152	139	141	164	129	144	146	137	144	147	141	147	147	145	137
1987	149	149	150	144	149	146	152	149	148	143	154	143	163	150	128
1988	128	125	136	116	126	126	139	114	133	125	120	135	130	119	120
1989	139	136	137	133	136	137	141	131	136	139	135	139	134	154	116
1990	144	144	141	143	147	145	137	139	147	138	155	129	144	143	140
1991	141	138	137	141	140	146	128	140	136	145	149	122	145	138	132
1992	137	141	125	135	135	137	134	130	134	136	141	134	125	142	126
1993	129	130	126	129	133	133	118	130	134	124	127	126	128	134	122
1994	120	122	105	120	111	120	114	116	118	111	115	118	115	120	109
1995	105	105	107	102	107	103	105	104	106	104	107	104	104	104	104
1996	127	132	116	129	118	122	128	127	120	126	131	120	133	121	109
1997	127	126	127	131	125	123	124	127	127	125	133	119	126	126	126
1998	109	109	111	96	111	106	104	104	109	105	105	114	98	117	94
1999	140	145	125	136	149	124	137	136	151	123	134	146	126	135	136
2000	156	130	139	144	142	141	154	127	150	138	150	130	149	132	141
2001	157	139	140	138	144	142	146	154	132	142	142	151	138	142	142
2002	150	133	143	140	143	140	150	140	137	138	146	141	140	143	137
2003	134	133	133	133	138	130	135	135	129	135	147	122	128	137	129

Note. Sample size across strata defined from the estimated propensity score (see Appendix C).

TABLE A-5. Sample size by propensity score stratum (Interview Survey)

	Stratum														
Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1982	4424	2893	2327	2155	1789	1589	1510	1559	1216	1494	1351	1105	931	804	817
1983	4935	3075	2501	1994	1520	1763	1628	1628	1383	1082	1237	968	1099	1008	861
1984	5513	3426	2125	2110	1846	1528	1602	1534	1480	1617	1283	1277	1163	949	858
1985	5512	3530	2903	2220	1955	1785	1856	1587	1408	1335	1455	1402	1321	1203	974
1986	6380	4103	3234	2793	1874	2224	1878	1693	1780	1535	1313	1401	1293	1058	828
1987	5035	3922	2510	2289	2041	1828	1724	1646	1659	1440	1422	1309	1219	958	713
1988	4700	3119	2962	2072	1926	1853	1761	1266	1599	1225	1353	1233	895	737	740
1989	4740	4210	2597	1847	1715	1648	1697	1655	1332	1509	1100	1136	1031	893	651
1990	4991	3225	2621	1989	1912	1496	1621	1573	1577	1307	1537	1153	1016	983	790
1991	5070	3666	2170	1696	1633	1784	1838	1522	1329	1389	1346	1008	1108	789	753
1992	5668	3286	2305	2023	1761	1550	1559	1277	1662	1159	1308	1102	938	929	839
1993	4872	3411	2953	1804	1878	1717	1356	1622	1485	1287	1222	1344	1046	959	765
1994	4285	3944	2657	1978	1796	1908	1304	1463	1417	1103	1161	1138	996	1113	730
1995	4834	3594	3016	2110	1722	1502	1139	1231	1222	1020	1215	1062	953	1012	791
1996	5200	3244	2625	2223	1603	1662	1506	1479	1462	1296	1215	1193	1060	780	634
1997	5206	3274	2973	2036	1718	1694	1338	1551	1411	1021	1296	1175	831	1035	902
1998	4412	3795	2941	2002	1695	1500	1349	1642	1458	1567	1121	1499	1147	1099	717
1999	5983	4214	3427	3118	2385	1935	1867	1911	1953	1907	1884	1736	1572	1608	1352
2000	5788	4146	3310	3295	2204	1874	2199	1848	1947	1805	1750	1689	1854	1651	1609
2001	6990	4548	3738	2952	2322	2055	2134	1889	1695	2093	1895	1770	1720	1780	1547
2002	6006	4854	4535	3272	2353	2434	2471	2120	1898	1919	1908	1864	1927	1939	1633
2003	6285	4743	4163	2587	1958	2054	1844	2248	1959	1915	1921	1223	1517	1207	663

Note. Sample size across strata defined from the estimated propensity score (see Appendix C).

TABLE A-6. Mean of $p \equiv Pr_{H_0}\{|T^*| \geq |T_{obs}|\}$, Ordered Probit

Stratum	1982-1987	1988-1990	1991-1995	1996-2000	2001-2003
1	0.165	0.286	0.380	0.368	0.298
2	0.144	0.345	0.363	0.341	0.340
3	0.119	0.313	0.390	0.340	0.370
4	0.131	0.332	0.380	0.376	0.346
5	0.141	0.320	0.382	0.330	0.365
6	0.146	0.396	0.340	0.341	0.358
7	0.148	0.383	0.365	0.323	0.363
8	0.142	0.399	0.407	0.353	0.350
9	0.151	0.406	0.372	0.312	0.360
10	0.150	0.396	0.357	0.334	0.367
11	0.143	0.383	0.346	0.319	0.413
12	0.144	0.385	0.343	0.317	0.381
13	0.160	0.432	0.329	0.351	0.292
14	0.200	0.387	0.326	0.345	0.373
15	0.164	0.407	0.343	0.384	0.357

Note. The table reports p-values for the null hypothesis of strong ignorability using the statistic based on the estimated coefficient of the survey dummy in the ordered probit regression with dependent variable equal to a quartile indicator of the distribution of Y_1. See Appendix B for details about the computation of p-values.

TABLE A-7. Mean of $p \equiv Pr_{H_0}\{|T^*| \geq |T_{obs}|\}$, Wilcoxon-Mann-Whitney

Stratum	1982-1987	1988-1990	1991-1995	1996-2000	2001-2003
1	0.144	0.267	0.370	0.358	0.284
2	0.115	0.331	0.376	0.334	0.314
3	0.095	0.323	0.400	0.337	0.352
4	0.105	0.337	0.385	0.374	0.358
5	0.117	0.318	0.379	0.330	0.365
6	0.125	0.392	0.351	0.347	0.345
7	0.121	0.392	0.374	0.324	0.348
8	0.123	0.412	0.399	0.347	0.367
9	0.135	0.428	0.397	0.311	0.371
10	0.133	0.401	0.363	0.336	0.371
11	0.125	0.380	0.335	0.314	0.414
12	0.127	0.386	0.349	0.326	0.394
13	0.149	0.441	0.319	0.351	0.276
14	0.196	0.374	0.331	0.344	0.351
15	0.150	0.411	0.368	0.396	0.350

Note. The table reports p-values for the null hypothesis of strong ignorability using the Wilcoxon-Mann-Whitney statistic. See Appendix B for further details about their computation.

TABLE A-8. Other summary statistics of $p \equiv Pr_{H_0}\{|T^*| \geq |T_{obs}|\}$, Ordered Probit

Stratum	1982-1987	1988-1990	1991-1995	1996-2000	2001-2003
			25th percentile		
1	0.000	0.030	0.090	0.060	0.030
2	0.000	0.060	0.080	0.060	0.050
3	0.000	0.050	0.090	0.050	0.050
4	0.000	0.040	0.060	0.080	0.075
5	0.000	0.060	0.080	0.050	0.055
6	0.000	0.090	0.065	0.070	0.060
7	0.000	0.100	0.085	0.060	0.080
8	0.000	0.120	0.105	0.050	0.060
9	0.000	0.090	0.070	0.035	0.090
10	0.000	0.110	0.080	0.040	0.070
11	0.000	0.110	0.040	0.050	0.080
12	0.000	0.100	0.050	0.030	0.100
13	0.000	0.160	0.050	0.060	0.020
14	0.000	0.090	0.045	0.040	0.080
15	0.000	0.125	0.060	0.080	0.065
			Median		
1	0.030	0.130	0.310	0.275	0.190
2	0.020	0.260	0.300	0.240	0.250
3	0.010	0.220	0.330	0.280	0.320
4	0.010	0.245	0.335	0.305	0.250
5	0.020	0.220	0.315	0.220	0.320
6	0.020	0.330	0.280	0.245	0.260
7	0.020	0.330	0.305	0.230	0.315
8	0.020	0.360	0.380	0.290	0.270
9	0.020	0.380	0.315	0.220	0.290
10	0.030	0.350	0.295	0.245	0.290
11	0.020	0.305	0.260	0.240	0.380
12	0.010	0.330	0.245	0.210	0.330
13	0.030	0.390	0.220	0.295	0.170
14	0.060	0.305	0.220	0.240	0.295
15	0.040	0.405	0.250	0.350	0.265
			75th percentile		
1	0.220	0.540	0.640	0.665	0.495
2	0.170	0.585	0.615	0.575	0.565
3	0.130	0.540	0.640	0.580	0.655
4	0.140	0.565	0.640	0.635	0.585
5	0.140	0.535	0.660	0.585	0.630
6	0.170	0.670	0.565	0.600	0.630
7	0.190	0.655	0.610	0.545	0.590
8	0.160	0.660	0.670	0.610	0.630
9	0.170	0.675	0.635	0.540	0.590
10	0.240	0.650	0.585	0.585	0.615
11	0.170	0.640	0.620	0.550	0.715
12	0.180	0.630	0.600	0.560	0.620
13	0.200	0.705	0.565	0.600	0.515
14	0.290	0.645	0.570	0.630	0.635
15	0.220	0.645	0.605	0.640	0.615

Note. The table reports the 25th, the 50th, and 75th percentiles for the distribution of the bootstrapped p-values for the null hypothesis of strong ignorability using the statistic based on the estimated coefficient of the survey dummy in the ordered probit regression with dependent variable equal to a quartile indicator of the distribution of Y_1. See Appendix B for further details.

TABLE A-9. Other summary statistics of $p \equiv Pr_{H_0}\{|T^*| \geq |T_{obs}|\}$, Wilcoxon-Mann-Whitney

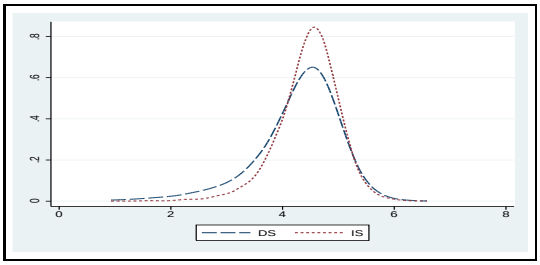
Stratum	1982-1987	1988-1990	1991-1995	1996-2000	2001-2003
			25th percentile		
1	0.000	0.020	0.080	0.050	0.020
2	0.000	0.050	0.100	0.070	0.050
3	0.000	0.040	0.100	0.055	0.050
4	0.000	0.050	0.080	0.080	0.060
5	0.000	0.055	0.090	0.050	0.050
6	0.000	0.095	0.060	0.060	0.080
7	0.000	0.110	0.080	0.050	0.070
8	0.000	0.095	0.110	0.050	0.070
9	0.000	0.120	0.090	0.030	0.090
10	0.000	0.130	0.080	0.040	0.090
11	0.000	0.110	0.040	0.050	0.095
12	0.000	0.125	0.050	0.050	0.115
13	0.000	0.150	0.040	0.060	0.010
14	0.000	0.090	0.050	0.040	0.070
15	0.000	0.130	0.080	0.100	0.055
			Median		
1	0.020	0.130	0.300	0.280	0.180
2	0.010	0.260	0.320	0.260	0.195
3	0.010	0.205	0.340	0.240	0.285
4	0.010	0.270	0.335	0.305	0.275
5	0.010	0.240	0.330	0.230	0.310
6	0.010	0.340	0.260	0.265	0.230
7	0.010	0.340	0.310	0.240	0.275
8	0.010	0.360	0.370	0.275	0.300
9	0.020	0.400	0.360	0.200	0.290
10	0.010	0.355	0.280	0.245	0.305
11	0.010	0.340	0.240	0.225	0.415
12	0.010	0.315	0.275	0.220	0.350
13	0.020	0.410	0.210	0.280	0.130
14	0.060	0.340	0.240	0.265	0.230
15	0.030	0.370	0.300	0.370	0.250
			75th percentile		
1	0.180	0.425	0.640	0.640	0.450
2	0.130	0.555	0.600	0.545	0.525
3	0.090	0.575	0.700	0.580	0.620
4	0.110	0.595	0.660	0.640	0.610
5	0.100	0.515	0.635	0.600	0.655
6	0.120	0.670	0.600	0.580	0.610
7	0.130	0.650	0.625	0.550	0.575
8	0.130	0.690	0.650	0.590	0.635
9	0.140	0.695	0.680	0.550	0.645
10	0.170	0.660	0.600	0.590	0.625
11	0.130	0.650	0.610	0.530	0.690
12	0.140	0.640	0.585	0.575	0.635
13	0.170	0.750	0.560	0.590	0.490
14	0.300	0.610	0.590	0.590	0.630
15	0.190	0.680	0.630	0.640	0.595

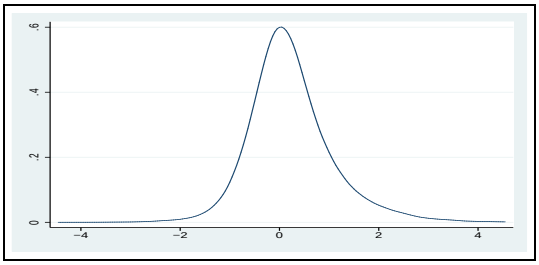
Note. The table reports the 25th, the 50th, and 75th percentiles for the distribution of the bootstrapped p-values for the null hypothesis of strong ignorability using the Wilcoxon-Mann-Whitney statistic. See Appendix B for further details.

FIGURE A-4. Food at Home

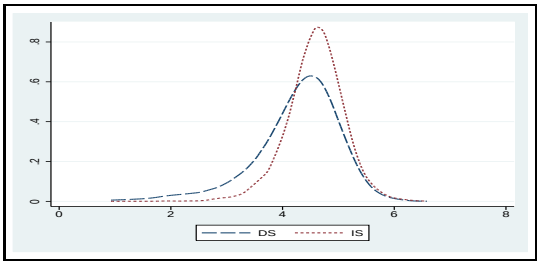
Distribution: $f_{Y_0}[\eta]$ (DS) and $f_{Y_1}[\eta]$ (IS) Effect of the survey instrument: $f_{Y_1-Y_0}[\eta](DS)$

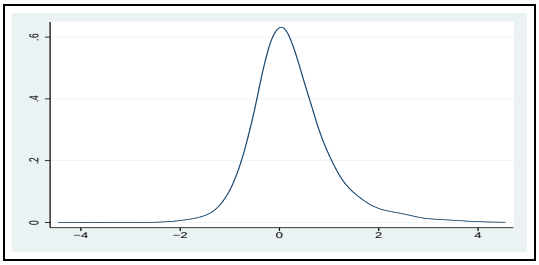
Years 1982-1987



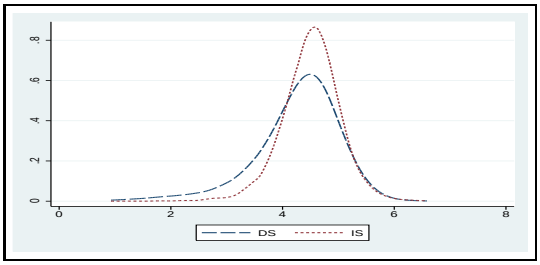


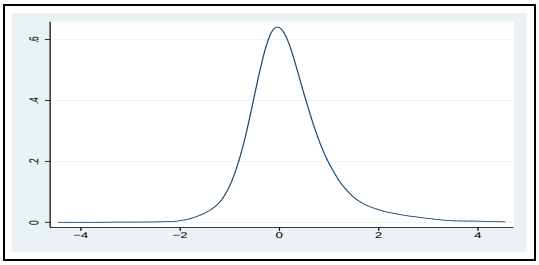
Years 1988-1990



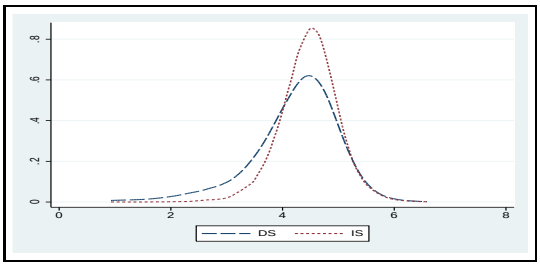


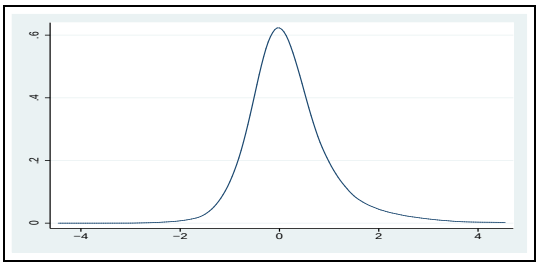
Years 1991-1995



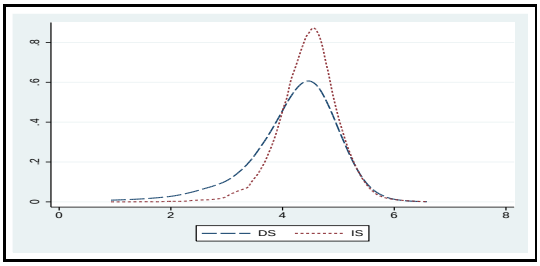


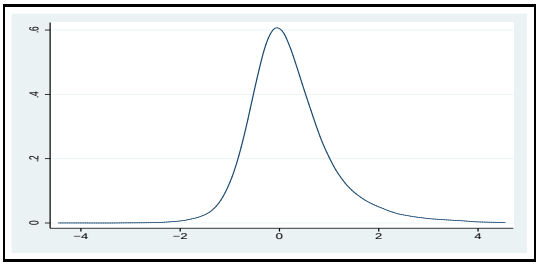
Years 1996-2000





Years 2001-2003



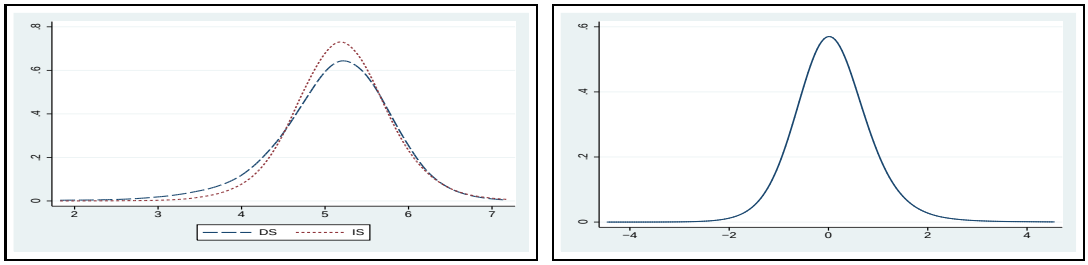


Note. Kernel density estimates of $f_{Y_0}[\eta]$, $f_{Y_1}[\eta]$ and $f_{Y_1-Y_0}[\eta]$ (see Section 6).

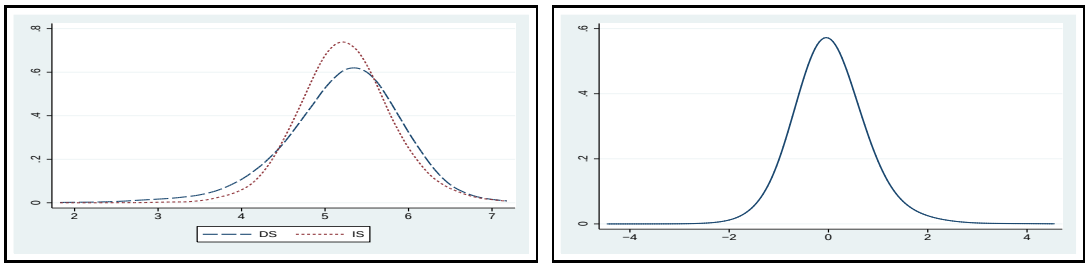
FIGURE A-5. $\mathcal{D}$ goods

Distribution: $f_{Y_0}[\eta]$ (DS) and $f_{Y_1}[\eta]$ (IS) Effect of the survey instrument: $f_{Y_1-Y_0}[\eta]$

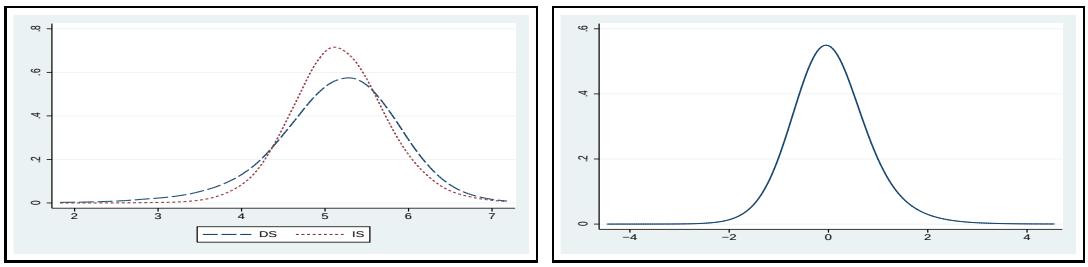
Years 1982-1987



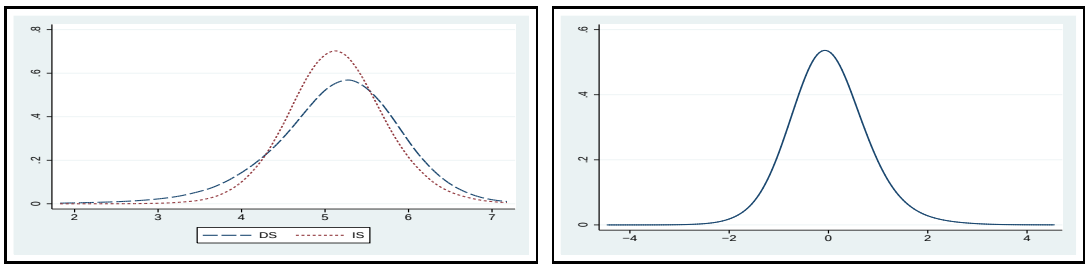
Years 1988-1990



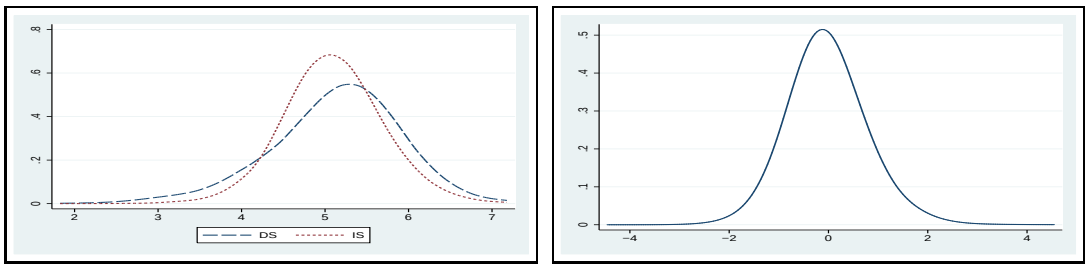
Years 1991-1995



Years 1996-2000



Years 2001-2003

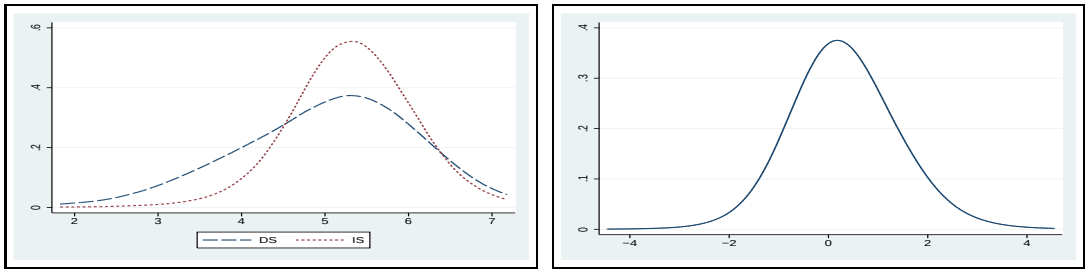


Note. Kernel density estimates of $f_{Y_0}[\eta]$, $f_{Y_1}[\eta]$ and $f_{Y_1-Y_0}[\eta]$ (see Table 1 for the definition of $\mathcal{D}$ goods).

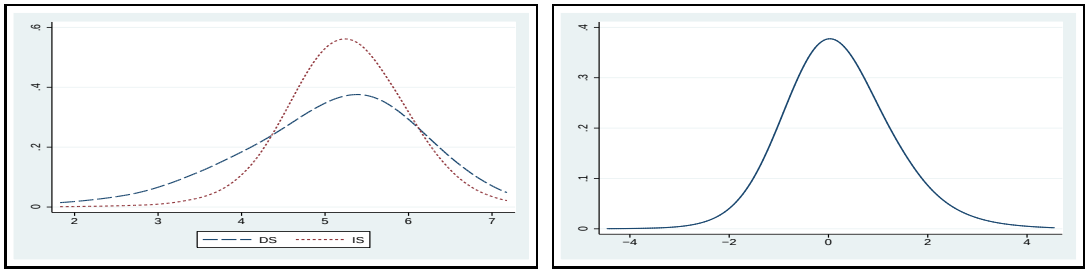
FIGURE A-6. $\mathcal{R}$ goods

Distribution: $f_{Y_0}[\eta]$ (DS) and $f_{Y_1}[\eta]$ (IS) Effect of the survey instrument: $f_{Y_1-Y_0}[\eta]$

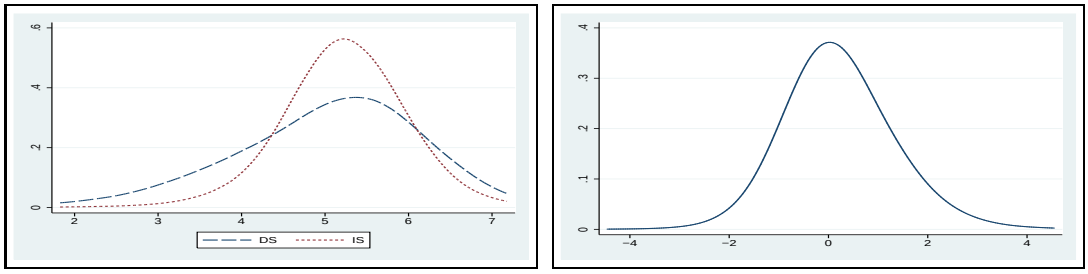
Years 1988-1990



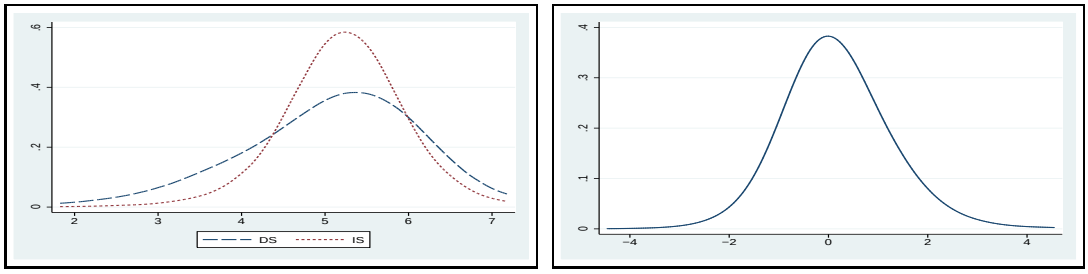
Years 1991-1995



Years 1996-2000



Years 2001-2003



Note. Kernel density estimates of $f_{Y_0}[\eta]$, $f_{Y_1}[\eta]$ and $f_{Y_1-Y_0}[\eta]$ (see Table 1 for the definition of $\mathcal{R}$ goods).