Food Policy xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Food Policy

journal homepage: www.elsevier.com/locate/foodpol

A comparison of recall and diary food expenditure data

Matthew Brzozowski[a], Thomas F. Crossley[b,*], Joachim K. Winter[c]

[a] Department of Economics, York University, 4700 Keele Street, Toronto, Ontario, Canada
[b] Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom
[c] Department of Economics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany

ARTICLE INFO

Keywords:
Expenditure
Survey data
Measurement error
Recall bias

JEL classification:
C81
D12

ABSTRACT

Recall food expenditure data, which is the basis of a great deal of empirical work, is believed to suffer from considerable measurement error. Diary records are believed to be more accurate. We study an unusual data set that collects recall and diary data from the same households and so allows a direct comparison of the two methods of data collection. The diary data imply measurement errors in recall food expenditure data that are substantial, and which do not have the properties of classical measurement error. However, we also present evidence that the diary measures are themselves imperfect.

1. Introduction

Information on household food expenditure is crucial for a broad range of economic and policy research, including research on consumption and demand behaviour, and on living standards, poverty and inequality. This is in part because there is a long tradition of treating food consumption as a welfare measure, and because food expenditure feeds into nutrition and health. Additionally, and more practically, household surveys in developed countries that have a panel structure, or that collect other important information from households, often collect only limited expenditure information because of response load considerations. Such surveys usually do ask a recall food expenditure question. Well-known examples are the Panel Study of Income Dynamics (PSID) in the U.S.,[1] the British Household Panel Survey (BHPS), and longitudinal surveys of aging such as the English Longitudinal Study of Ageing (ELSA) and the Survey of Heath Ageing and Retirement in Europe (SHARE). Developing and middle-income countries are facing new social and economic challenges and those challenges make longitudinal and multiple-domain surveys critical inputs to good policy making. A good example is population aging, and the China Health and Retirement Longitudinal Study (CHARLS), first fielded in 2011, includes a simple recall question on expenditure for food consumed at home.

Measurement error in expenditure data has been an important concern of researchers who employ such data. Given the prominent role of food expenditure data, measurement error in food expenditure data is of particular interest. This paper provides new evidence on the extent and character of measurement error in food expenditure data. Our specific focus is a comparison of food expenditure measures obtained from simple recall questions and from expenditure diaries, as the latter have long been viewed as providing superior measures but come with high respondent load.

The literature on survey response behaviour noted early on that questions that require recalling quantities from memory are difficult to answer (Gray, 1955). There is now substantial evidence of 'forgetting': that memory declines with the length of the recall period, leading to under-estimation; see Sudman et al. (1996) for a review. The situation is complicated by the fact that forgetting does not occur at random but might be differential across respondents and types of questions. The existing evidence on the measurement of consumption expenditure, and on sources of measurement error, is summarized by Browning et al. (2014) and Crossley and Winter (2015).

Interestingly, despite the growing concern about the quality of recall data, there are few systematic comparisons of simple recall expenditure questions with diary measures. The Canadian Food Expenditure Survey (FoodEx) provides a unique opportunity to study how food expenditure measures constructed from simple recall questions compare to those obtained from expenditure diaries. The survey asks respondents to first estimate their household's food expenditure over the past four weeks, and then to record food expenditure in a diary for two weeks. Thus it allows for *within-subject* comparisons. Most existing studies of measurement error in expenditure survey use *between-*

* Corresponding author.
E-mail addresses: brzozows@yorku.ca (M. Brzozowski), tcross@essex.ac.uk (T.F. Crossley), winter@lmu.de (J.K. Winter).

[1] Though the PSID has been increasing the breadth of the expenditure information it collects. See Andreski et al. (2014).

http://dx.doi.org/10.1016/j.foodpol.2017.08.012

subject designs. For example, Battistin et al. (2003) and Browning et al. (2003) compare data from different surveys, so that corrections must be made for differences in sample design, coverage etc. Gibson (2002), Beegle et al. (2012) and Battistin and Padula (2013) compare multiple samples from a single survey. This allows for a direct estimate of difference in distributions, it does not allow for an examination of the distribution of differences between recall and diary records. In contrast, a within-subject design allows for calculation of a recall-diary difference for each household, and for an examination of the properties of those differences. Of course this advantage must be balanced against potential disadvantages of a within-subject design, and we discuss this further below.

In their *Handbook of Econometrics* survey, Bound et al. (2001) emphasize that while econometric methods for dealing with measurement error typically assume that measurement errors are "classical", much of the available empirical evidence contradicts this assumption. They also emphasize the usefulness of validation data in characterizing the joint distribution of error-ridden measures and their true values, and for testing the assumption of classical measurement error or other assumptions about measurement error. Bound et al. report evidence on measurement error in a variety of constructs (for example wages and earnings) but not food expenditure.

The FoodEx was not a designed validation study. However, because diary measures are widely considered the gold standard for collecting expenditure information, and because of the within-subject design, it is possible to use treat the FoodEx as an approximation to a validation study of the recall data, and to carry out analyses similar to those discussed by Bound et al. At the same time, how well the FoodEx approximates a genuine validation study depends on how well the diary measures capture true expenditure, and we also investigate this question.

The next section of this paper describes the Canadian Food Expenditure survey as well as a second, more widely used Canadian expenditure survey (the Family Expenditure Survey or FamEx), which also collects recall food consumption data. This section also provides a preliminary analysis of the different food expenditure measures available in the two surveys.

In Section 3, we calculate errors in recall food expendiure, using the diary measures to construct "true" food expenditure in a number of different ways. Under the assumption that true food expenditure can be constructed from the diary records, measurement errors in recall food expenditure data appear to be substantial, and they do not have the properties of classical measurement error. In particular, they are neither mean independent of true expenditure nor homoscedastic. They are also not well approximated by a normal distribution. However, we also show evidence that diary measures are themselves imperfect. This suggests alternative interpretations for the differences between recall and diary expenditure measures.

Finally, Section 4 offers some concluding remarks.

2. Canadian household expenditure surveys

The 1996 Canadian Food Expenditure Survey (FoodEx) was a large, nationally representative survey of Canadian households. Respondents were asked basic demographic questions and recall food expenditure questions. In addition, they were asked to record every food purchase in a diary, for two contiguous weeks. Conducting the survey involved three visits to each household. At the initial visit, demographic and recall food expenditure questions are asked. In addition, respondents were instructed on the proper technique for filling out the food expenditure diaries. After a week the first diary was collected and the household received another second blank diary in which to record purchases made in the following week. This second diary was collected during the third visit. During the second and third visits, the interviewers double-checked the diaries and verified the exactness and fullness of the responses. The survey was run continuously throughout the year so that the seasonality of purchases is not an issue. The initial response rate was 76 percent, and there were 10,898 responding households. Attrition between the first and second week was less than 2 percent. Statistics Canada provides household weights that take account of the survey design and non-response, but not of attrition between the two weeks. Further details can be found in Statistics Canada (1999).

For the purposes of this paper, the key feature of the FoodEx is that each household is asked recall food expenditure questions as well as recording food expenditures in diaries. As noted above, this allows for a within-subject design. For a validation study, a within-subject design has the important advantage that the difference between the data being assessed (here the recall data) and the superior data (the diary) can be calculated for each responding unit. This allows for a direct analysis of these differences. If the superior data closely approximate the truth, these differences reveal the measurement errors in the data being assessed at the level of the responding unit. This in turn reveals key properties of the measurement error (such as whether the measurement error correlated with the true value).

Against this, there may be important disadvantages of a within-subject design. Perhaps the most important is the possibility of cross-contamination between the two measures. It may be that the expectation of completing a diary influences the effort that households put into their recall estimate of food expenditures or other aspects of the recall response. Equally, it may be that having offered a recall estimate affects diary behaviour. A between-subject design does not suffer from this possibility. Below we describe how we use a second Canadian expenditure survey to provide some evidence on cross-contamination.

A second possible concern with comparisons such as the one allowed by the FoodEx was raised by Gibson (2002).[2] He notes that in the FoodEx, the beginning of the recall period is not marked by a visit from an interviewer, whereas the diary period is. This may lead to "telescoping errors" in the recall data. We believe this is not a problem, for two reasons. First, most of the empirical evidence on telescoping is for larger, irregularly purchased items, like home repairs, and not for more regularly purchased expenditure categories like food.[3] Second since almost all simple recall expenditure questions longitudinal and multiple-domain surveys in developed countries share this possible problem, the FoodEx allows the appropriate comparison: between diary collection and recall information as usually collected in such surveys. A study of recall expenditure data from a survey in which the recall measure was marked by a visit from an interviewer would not be as informative about the recall expenditure data in the longitudinal surveys listed in the Introduction.

The exact wording of the key recall food expenditure questions is as follows:

In the last four weeks…

Q1. How much do you estimate this household spent on food and other groceries purchased from stores (including farmer stalls and home delivery)? Exclude periods away from home overnight or longer. Report bulk purchases of food for canning, freezing in question 3.
Q2. About how much of this amount was for non-food items such as paper products, household supplies, pet food, alcoholic beverages, etc.?

Surveys that ask simple sets of recall food expenditure questions do differ somewhat in their formulation. For example, the PSID refers to the amount the household "usually" spends on food at home, while the FoodEx refers particularly to the last four weeks.

[2] Gibson was responding to a very early version of Ahmed et al. (2010).

[3] A key development in the literature on recall expenditure questions was the identification of 'telescoping' as a significant problem by Neter and Waksberg (1964). This is the phenomena of respondents erroneously including in their response expenditures that occurred before the specified recall period, leading to an over-estimation of expenditure in the recall period. See Browning et al. (2014) and Crossley and Winter (2015) for further discussion.

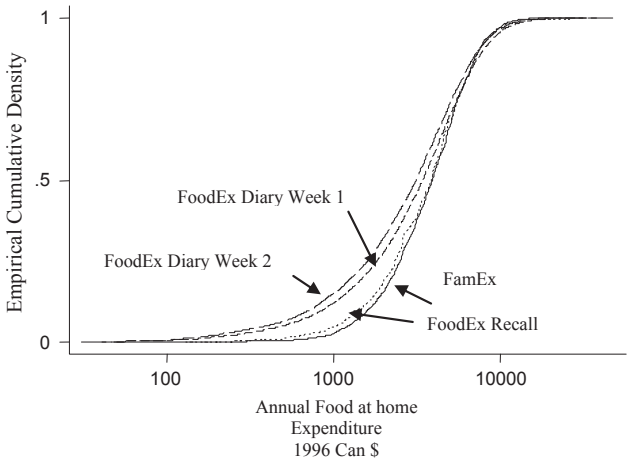

Fig. 1. Food expenditure, empirical CDFs.

We construct recall food expenditure as Q1 – Q2. From a total of 10,898 respondent households, this quantity is available for all but 220 households, a very low rate of item non-response (2 percent).

Although comparison of recall and diary data within the FoodEx is the main focus of our analysis, we can also compare the FoodEx data to data from a second large Canadian survey. The 1996 Family Expenditure Survey (FamEx) is a full household expenditure survey (collecting information on all categories of expenditure).[4] Unlike most national expenditure surveys, the FamEx does not have a diary component. Instead, face-to-face interviews are conducted in the first quarter to collect income and expenditure information for the previous year (thus the 1996 data were collected in January, February and March of 1997 but refer to the 1996 year calendar year). The FamEx is therefore an unusual kind of recall survey. For food expenditure, the FamEx asks a very simple two-question sequence almost identical to the FoodEx questions given above.

Considerable effort is made to ensure the quality of the FamEx data. Barrett et al. (2015) show that the main Canadian budget survey outperforms other similar national budget surveys in developed countries (particularly the UK, the US and Australia) with respect to the match of food totals to the national accounts. Statistics Canada also undertakes various checks of the data and the data are generally thought to be of very good quality.[5] There are 10,085 respondent households in the 1996 FamEx.[6]

Because the FamEx collects annual data and the FoodEx survey is run continuously over the year, they refer to the same time period. The surveys were based on the same (Labour Force Survey) sampling frame. Thus these two surveys readily lend themselves to comparison.[7]

Summary statistics comparing the two data sets are presented in a supplemental appendix (see Appendix Table A1). Our main interest in the FamEx follows from the fact that although the recall food expenditure questions are very similar to the FoodEx, respondents are not also asked to complete expenditure diaries. Thus comparisons of the FoodEx and FamEx recall responses allow us to assess the possibility that anticipation of completing a diary affects the recall responses in the FoodEx.

In summary then, we have four distinct data items that capture the distribution of food expenditure in Canada in 1996. These are:

(i) The “food at home” expenditure category in the FamEx.
(ii) The recall food expenditure measure we construct for the FoodEx (described above).
(iii) Food expenditures recorded in the first week diary of the FoodEx.
(iv) Food expenditures recorded in the second week diary of the FoodEx.

We have multiplied the second by 13 and the third and fourth by 52 so that all are annual measures.

Fig. 1 displays the empirical cumulative distribution of these four measures, while Table 1 reports the mean, median and coefficient of variation for these four measures as well as for budget shares and income in the two surveys.[8] Several features are notable. First, the recall responses in the FoodEx and FamEx are very similar. Those two distributions in Fig. 1 are difficult to distinguish and median and coefficient of variation is identical to two digits for these two measures. The mean is also close. Recall that the two sets of recall questions are almost identical but respondents to the FamEx are not asked to complete expenditure diaries. From this comparison we conclude that responses to the FoodEx recall question are not affected by anticipation of completing the diary. Unfortunately, this comparison is silent with regard to contamination in the other direction: it does not rule out the possibility that diary response behaviour is affected by having given a recall estimate.

A second observation is that the diary records are considerably lower than the recall responses of the same individuals (in the FoodEx) or a second sample drawn from the same population (the FamEx). Gieseman (1987) and Bee et al. (2015) report that the recall measure of food US Consumer Expenditure Interview Survey is on average higher than the diary measure in the US Consumer Expenditure Diary survey, and that recall totals are closer to the PCE numbers from the National Income and Product Accounts (NIPA).

Third, the diary records are considerably more variable than the recall records. In household expenditure surveys, respondents are typically asked to keep diaries only for short periods, partly in recognition that careful completion of a diary implies significant respondent burden. For categories of expenditure that are purchased irregularly, or at regular intervals that exceed the duration of diary keeping, infrequency problems will arise. With modern freezing and refrigeration technologies, many types of food can be stored. Even in less developed economies, staples such as grains are often purchased in bulk.

Infrequency is a kind of measurement error: a household may over (or under) estimate their true rate of expenditure if the diary keeping period happens to include (or not include) a major purchase. While this may not affect estimates of average expenditure across households it certainly increases dispersion will therefore bias estimates of food poverty.

Fourth, there is a notable drop off, of more than 10 percent on average, between the first and second week of the diary. The drop off between the first and second week of the diary seems to be evidence of “diary fatigue” or “diary exhaustion”. Statistics Canada (1999)

[4] The FamEx (and its subsequent replacement, the Survey of Household Spending) are the surveys that are used to determine the weights for the Consumer Price Index in Canada. They have also been extensively used for demand analysis.

[5] Respondent households are asked to consult bills and receipts and income is carefully reconciled with expenditures and savings. In some cases, multiple visits to a household are made. Further details on the quality of this data are in Brzozowski and Crossley (2011).

[6] Statistics Canada reports that the response rate to the FamEx surveys is about 75%.

[7] The only significant obstacle to the direct comparison of the data stems from differences in the household income information included in the files. The FamEx file includes only net household income while the FoodEx file includes only gross household income (the FoodEx file does not contain personal income data). However, the FamEx also includes gross personal income for head and spouse, and thus the FamEx income variable used in our analysis below is constructed as the sum of these two items. This obviously is an imperfect match to the FoodEx income information when there are additional earners in the household. A second minor difference between the data sets concerns the top coding of numbers of different types of persons (children, young adults, adults, seniors) in the household. For the Foodex these are recorded as 0, 1 or (2 or more). In the FamEx, the top-coding is at 3. In both data sets total household size is top-coded at 6.

[8] Empirical cumulative distributions for income and budget shares are presented in Appendix Tables A1 and A2.

Table 1
Summary statistics: annual household food expenditures, income, and budget shares.

		FamEx	FoodEx		
			Diary Week 1	Diary Week 2	Recall measure
Sample size		10,085	10,876	10,719	10,678
Food at home Expenditure	Mean	4336	3854	3432	4156
	Median	3900	3261	2839	3911
	Coefficient of variation	0.58	0.82	0.88	0.58
Food at home Budget Share	Mean	0.15	0.12	0.11	0.12
	Median	0.10	0.08	0.07	0.10
	Coefficient of variation	2.70	1.57	2.69	2.22
Income Before Taxes	Mean	45,716	44,016		
	Median	38,500	37,200		
	Coefficient of variation	0.73	0.75		

Notes:
a. The 1996 FOODEX contains 10,898 observations (households). 22 did not submit a first week diary while 179 did not submit a second week diary. The attrition rate (from week 1 to week 2) was 1.6%. 220 households did not provide a recall food expenditure estimate.
b. Statistics are calculated using survey weights.

concludes that diary exhaustion was a significant factor affecting accuracy of the responses. They report that, in addition to the between week differences, within week responses tended to be significantly larger for the earlier days of either week. Such exhaustion effects in expenditure diaries have been known for a long time (Kemsley, 1961; Turner, 1961; Sudman and Ferber, 1971; McWhinney and Champion, 1974.) Similar phenomena have been reported in the diary sample of the U.S. Consumer Expenditure Survey (CEX) (Silberstein and Scott, 1991; Stephens, 2003) and in the U.K. Family Expenditure Survey (Tanner, 1998).

We have no way with these data to assess whether diary response behaviour is affected by having given a recall estimate. However, because the problems we identify in the diary data (infrequency, diary fatigue) are shared by other developed-country diary-based expenditure surveys (in the U.S and the U.K., for example), we do not believe that the broad features of diary reporting behaviour in the FoodEx are due to priming with the recall question.

Tables 2 and 3 and Fig. 2 provide some supplemental analysis of diary fatigue in the FoodEx. Table 2 reports a regression of week-on-week changes on observable characteristics of households. The covariates are coded as deviations from means, so that the constant in the regression is just the average week-on-week change in the sample (converted to an annual amount). Week-on-week changes in recorded food expenditure are largely unrelated to observable household characteristics. The one exception is that households from the Atlantic

Table 2
Regression analysis: week on week change in food expenditure diary. Dependent Variable: (Week 1 Diary − Week 2 Diary) × 52.

	Coef.	(Standard error)
ln pcy	54.45	(59.39)
$(ln\ pcy)^2$	−0.46	(0.58)
Log household size	−753.66	(687.31)
Presence of children (0–15)	137.26	(171.33)
Presence of youths (16–24)	−3.22	(126.61)
Presence of seniors (65+)	6.23	(102.91)
2nd Earner in Household	−108.77	(125.83)
Constant	−418.97*	(43.60)
R-squared	0.001	

Notes:
a. Regressors are all measured as deviations from means.
b. * indicates $p < 0.05$.

Table 3
Ratio of mean week 2 expenditure over mean week 1 expenditure (By Broad Food Categories and Store Types).

All food at home	0.91
By category	
Meat	0.91
Fish and other marine products	0.94
Dairy products and eggs	0.91
Bakery and cereal products	0.91
Fruits and nuts	0.91
Vegetables	0.92
Condiments spices and vinegar	0.92
Sugar and sugar preparations	0.86
Coffee and tea	0.88
Fats and oils	0.92
Other food	0.93
Non alcoholic beverages	0.84
By Store Type	
Food from specialty stores	0.83
Food from convenience stores	0.75
Food from supermarkets	0.93
Food from other stores	0.83

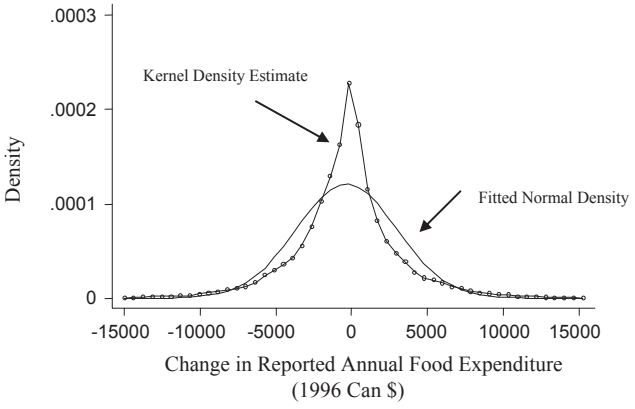

Fig. 2. Changes in reported food expenditure diary Week 1 to Week 2.

Provinces exhibit (on average) less diary fatigue. Table 3 examines the week-on-week change in recorded outlay by expenditure category and by store type. The results suggest that records of small items (coffee and tea, non-alcoholic beverages, sugar), and especially purchases from convenience stores decline from week one to week two. Fig. 2 illustrates that week-on-week changes in recorded expenditures are both positive and negative, are highly variable, and roughly symmetric around the (negative) mean.

Because diary records are usually thought to be quite accurate, the usual interpretation of the gap between the diary and recall measures might be that the latter suffer from significant over-reporting. However, the significant diary fatigue evident in the diary records suggests the possibility that the diary records (and even the first week diary records) suffer from significant under-reporting. This is in fact the conclusion reached by Statistics Canada which routinely inflates the diary information in publicly released data by the factor necessary to match the recall information.[9] (We have undone this adjustment for the purposes of our analysis.)

Fig. 3 displays histograms of the four food expenditure measures (note that, in this figure only, amounts are weekly rather than annual). These suggest that both diary and recall data may suffer from their own particular problems. In particular, the diary data exhibit significant numbers of zeros (as much as 10% of the sample each week). Since it is implausible that this large a fraction of the sample is fasting, a natural

[9] The factor that Statistics Canada inflates by is 15.8%.

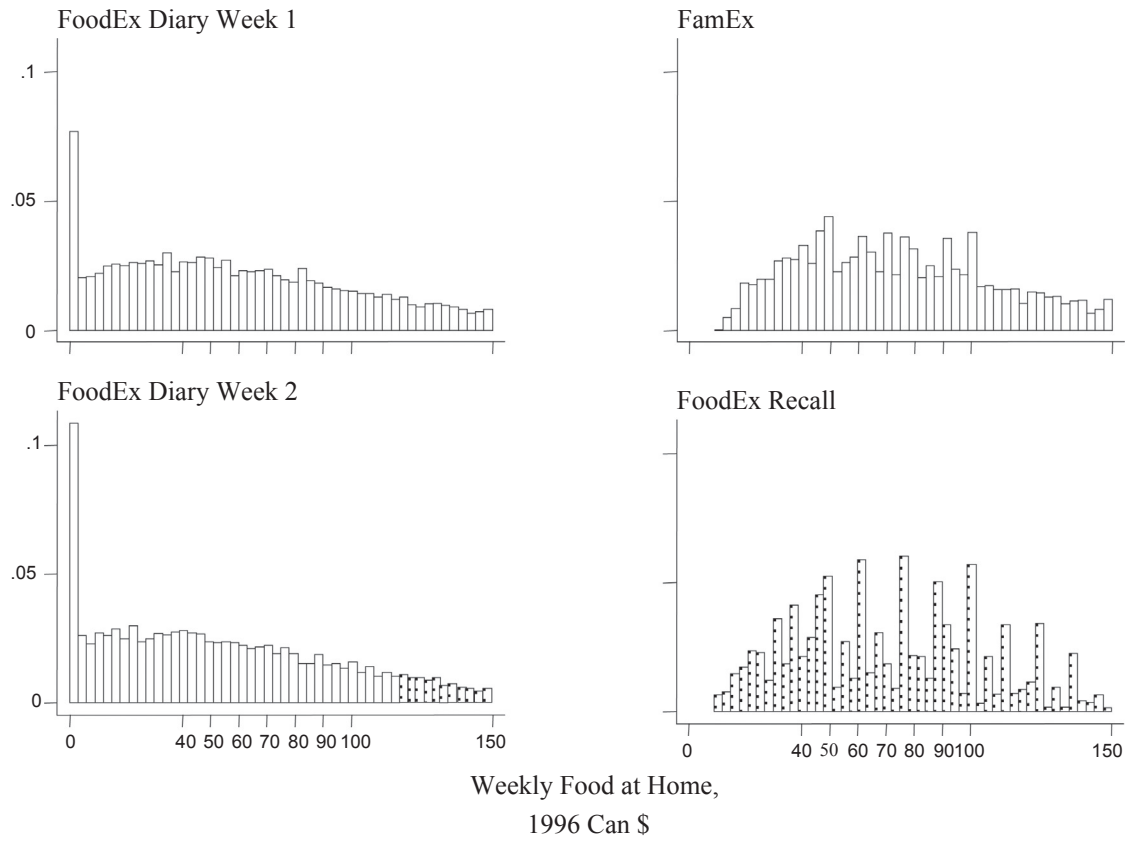

Fig. 3. Food expenditure, histograms.

interpretation is that the diary data suffer from purchase infrequency. There is a small literature on methods for dealing with purchase infrequency, including Keen (1986), Pudney (1988, 1989) and Meghir and Robin (1992). Note that this problem is not entirely resolved by combining the two weeks of diary data: the combined data still exhibit a significant spike at zero (of about 7%).[10]

On the other hand, Fig. 3 also suggests that the recall data suffer from considerable heaping and rounding (note the "spikes" in the empirical distribution at round figures such as \$50 and \$100). The consequences of such heaping and rounding, and methods for dealing with it, are discussed in Battistin et al. (2003), Heitjan and Rubin (1990) and Hoderlein et al. (2015). We now turn to a more detailed analysis of the differences between the recall and diary data.

3. Measurement errors in recall food expenditures

Let c^* be true food expenditure and c be an imperfect measure of that quantity. Define $\varepsilon = c - c^*$ so that:

$$c = c^* + \varepsilon$$

In order to work with c, it is common to make assumptions about the characteristics of ε. Typical assumptions include those that characterize "classical" measurement error (Bound et al., 2001): that the errors are mean zero and independent of the true level of expenditure and all other variables in the model. In our notation:

(i) ε is mean zero: $E[\varepsilon] = 0$,
(ii) ε is mean independent of (or uncorrelated with) c^*: $E[\varepsilon|c^*] = E[e]$. Note that a testable implication of this assumption is that a regression of c on c^* should give a coefficient (on c^*) of 1.
(iii) ε is mean independent of other variables, X: $E[\varepsilon|X] = E[\varepsilon]$.
(iv) ε is independent of c^*. This of course implies that higher moments of ε are not related to c^*: $E[\varepsilon^k|c^*] = E[\varepsilon^k]$, $k = 2,3...$, starting with conditional homoscedasticity: $E[\varepsilon^2|c^*] = E[\varepsilon^2]$.

Sometimes a distributional assumption is added, in particular, that the measurement error is normally distributed:

(v) $\varepsilon \sim N(0,\sigma^2)$,

Finally, it is useful to have a measure of the relative size of ε. A common measure is the signal-to-noise ratio of c, which is calculated as $R^2/1-R^2$ from a regression of c on c^*.

If c^* is observable, these things are all amenable to empirical investigation. On first thought, the FoodEx would seem to offer such a possibility. In particular, diary records of food expenditure are thought to be very accurate (see Battistin and Padula, 2013). Thus, a natural approach is to take the diary information in the FoodEx as true expenditure. However, the analysis of the previous section suggests that the diary measures are not prefect. Nevertheless, it is still very informative to compare the recall data to a superior measure. As Bound et al. (2001) note, most validation studies do not have a "perfect" or true measure to which to compare survey responses as even administrative records contain some errors. Moreover, it is common in the literature to assume that (i) diary records are very accurate, and (ii) the measurement errors in recall measures have particular (eg. classical) properties. It is certainly possible to test those hypotheses jointly with these data.

[10] It is worth noting that while the evidence of diary fatigue does suggest under-reporting in the diary measure, infrequency does not necessarily imply under-reporting (at least on average). To see this, suppose that households have usual food consumption (per period) of c^*. Suppose further that in any period they go food shopping probability p, so that when they do, they must purchase c^*/p to maintain usual consumption. Then the mean of observed purchases, c, (*including zeros*) is: $E[c] = pE[c^*/p] + p(0) = E[c^*]$. Thus the mean may not be affected by some kinds of infrequency (the key assumption here is that purchase frequency, p, is independent of consumption level, c^*.) See Pudney (1989) for further discussion.

The question, then, is how to best use the diary information. What we do is to construct, from the diary records, three alternative measures of "true" food expenditure, c^*:

(A) The first week diary;
(B) The average of 1st and 2nd week diaries;
(C) The linear projection of the recall measure onto the two diary measures.

Arguments can be made for each of these measures. (A) has the virtue that it minimizes the effects of diary exhaustion. On the other hand, it will be affected more by infrequency than (B). To construct (C) we regress the recall measure on the diary week records and take the predicted values from this regression as true expenditure (and hence the regression error is interpreted as measurement error in the recall measure). (C) is a weighted average of the first and second week of the diary (plus a constant), where the weights are chosen in a way that assumes the "best case" for the recall measure: note that this procedure imposes the assumptions that measurement error is mean zero and uncorrelated with the true value.

Table 4 presents summary statistics for the measurement error in recall food expenditures. Each column corresponds to one of the assumptions outlined above (A, B and C) regarding the true value. The first panel shows that the measurement errors have a positive mean if we take either the first week of the diaries or the average of the two weeks as c^* ($301 and $512 respectively). In either case, the errors have negative skew (−0.71 and −0.14 respectively), and have much thicker tales than the normal distribution (with measures of kurtosis of 10.0 and 12.1 respectively, where the normal distribution would be 3). Our third procedure (C), which imposes a mean of zero on the measurement errors, results in a distribution of measurement errors that is positively skewed, but again with thick tails. Kernel density estimates of all three distributions are presented in Fig. 4.

Table 4
Errors in recall food expenditure - descriptive statistics (1996 Can $ per year).

		A	B	C
Mean		301	512	0
Variance		9,198,159	6,057,782	4,297,449
Skewness		−0.71	−0.14	1.30
Kurtosis		9.97	12.07	9.50
Percentiles	5%	−4431	−3071	−2572
	10%	−2998	−2007	−2101
	25%	−1117	−720	−1360
	50%	367	428	−307
	75%	1913	1741	1024
	90%	3560	3223	2490
	95%	4797	4390	3696
Test of Mean Independence(β = 1)				
	β−1	−0.67	−0.52	β = 1
	[t-stat]	[−53.9]	[−32.8]	by construction
Test of Conditional Homoscedasticity B-P test,				
	Chi2	194	558	566
	dfProb > Chi2	2	2	2
		< 0.01	< 0.01	< 0.01
K-S test for Normality, p-value		< 0.01	< 0.01	< 0.01
R^2		0.19	0.27	0.27
Signal to Noise Ratio		0.23	0.36	0.36

Notes:
a. (A) Assumes first week diary measures "true" food expenditure. (B) assumes the average of 1st and 2nd week diaries measures "true" food expenditure. (C) Assumes the linear projection of the recall measure onto the two diaries measures "true" expenditure.
b. Signal to Noise Ratio is calculated as $R^2/1 - R^2$ from a regression of the recall measure on the assumed "true" measure.
c. Linear Regression of the recall measure on the two diary week records yields: Recall = 2391.6 + 0.239 Week1 + 0.245 Week2 + error. (0.012) (0.015)

The third and fourth panel of Table 4 present tests for mean independence and homoscedasticity of the error terms. These tests are implemented by regressing c on c^*. If the measurement errors are mean independent (uncorrelated with c^*), then the coefficient, β, on c^* should be 1. We present a t-test of this hypothesis. We then use a standard Breusch-Pagan test to test whether the second moment of the measurement errors is independent of c^* (that is, to test for heteroscedasticity in the measurement errors).

If we use the first week of the diary or the average of the two weeks as true food expenditure, then the measurement errors in the recall measure of food expenditure are strongly and negatively correlated with the true value. Mean independence is rejected with t-statistics of −55.8 and −32.2 respectively. Recall that true measure (C) assumes mean independence. By any measure of true food expenditure, homoscedasticity is strongly rejected, with p-values for the Breusch-Pagan test less than 0.01. Thus even if we impose mean independence (as in (C)), we reject independence.

The Breusch-Pagan tests uses residuals from the regression of recall expenditure on c^*. Squares of those residuals are regressed on c^* and c^* squared. Regardless of the choice of c^*, the linear term is always strongly negative and the quadratic term positive; in each case the estimated elasticity of the squared measurement error with respect to c^* is negative at the mean of c^*.[11] The nature of the heteroscedasticity seems to be that the measurement error variance falls with value of "true" expenditure, but at a decreasing rate.

In the next (5th) panel of Table 4 we present Kolmogorov-Smirnov tests of normality of the implied measurement errors. In all three cases, normality is strongly rejected, with p-values less than 0.001.

Finally, we calculate the signal-to-noise ratio for c under each of our assumptions about c^*. These suggest that the measurement errors in c are very substantial. If we take the first week diary record to be c^*, the signal-to-noise ratio in c is only 0.22. With either of the other two measures of c^* the signal to noise ratio in c rises to 0.36 (differing only beyond the fourth decimal place.) Equivalently, 70–80% of the cross-sectional variance in recall expenditure is measurement error. This is a very large number, but it is not unprecedented. For example, on the basis of serial correlation in the errors in consumption growth equations, Runkle (1991; pg 86) concludes "that approximately 76 percent of that portion of the variance in the growth rate of consumption unexplained by family-specific real interest rates is the result of measurement error" (where consumption is food expenditure as measured in the PSID).[12]

Table 5 presents the results of regressing the implied measurement errors on variables typically used in the modelling of expenditure: income, and demographic variables. If we take either the first week diary measure (A) or the un-weighted average of the two weeks (B) as true expenditure, then these income and demographic variables do not seem to be significant determinants of the implied measurement errors, except for the presence of youths in the household. The measurement errors implied by our third procedure (C) appear to be more strongly related to variables such as income, household size and the presence of children and youths.[13]

Table 6 presents the results of regressing the squares of the implied measurement errors on the same set of variables, in order to further investigate heteroscedasticity in those errors. Again, the results seem sensitive to the measure of c^* used to construct the measurement errors. The variance of the measurement errors constructed by either

[11] The elasticity of a variable, y, with respect to another variable, x, is the percentage change in y associated with a percentage change in x: $(\Delta y/y)/(\Delta x/x)$.

[12] Note though that first differencing usually removes signal, so that typically one would expect measurement error to be a smaller fraction of the variance in *levels*.

[13] The analyses reported in Tables 4–6 were repeated but with observations for which "true" food consumption was zero deleted from the sample. The results do not differ significantly differ from those reported in Tables 4–6. In particular, it is *not* the case that the rejection of normality is driven by these zeros.

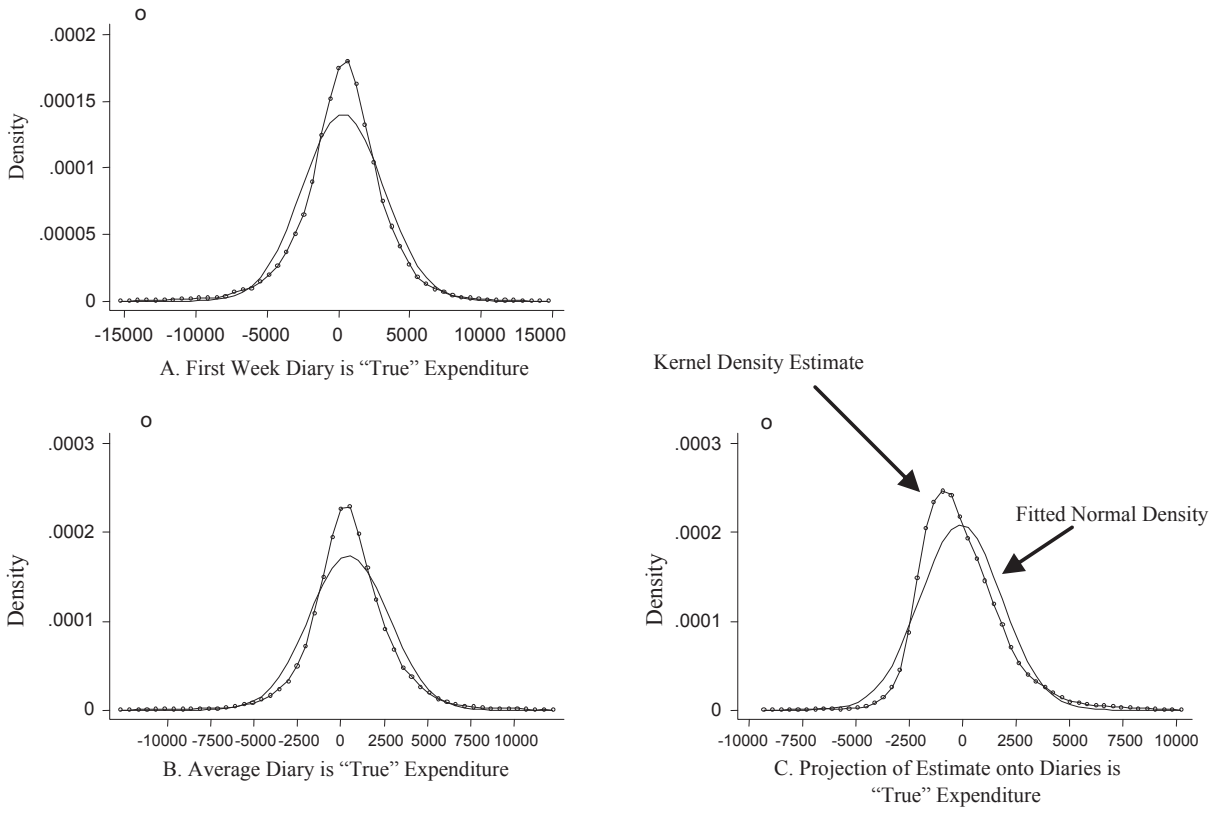

Fig. 4. Errors in recall food expenditure.

Table 5
Errors in recall food expenditure – regression on covariates (1996 Can $ per year).

	A		B		C	
	Coef	(Std Err)	Coef	(Std Err)	Coef	(Std Err)
ln pcy	1.64	(55.29)	−25.59	(40.63)	*139.41	(31.42)
(ln pcy)^2	< 0.01	(0.54)	0.24	(0.38)	*−0.82	(0.29)
Log household size	−181.58	(635.54)	195.25	(475.72)	*−900.01	(363.99)
Presence of children (0–15)	214.70	(160.08)	146.06	(120.68)	*−198.22	(89.99)
Presence of youths (16–24)	*373.79	(114.29)	*375.40	(92.38)	*181.72	(71.86)
Presence of seniors (65+)	−142.65	(97.89)	−145.76	(76.84)	−48.11	(60.69)
2nd Earner in Household	−91.88	(119.51)	−37.50	(94.61)	−51.16	(74.54)
Constant	*291.03	(40.01)	*500.51	(31.85)	−7.12	(24.79)

Notes:
a. (A) Assumes first week diary measures "true" food expenditure. (B) assumes the average of 1st and 2nd week diaries measures "true" food expenditure. (C) Assumes the linear projection of the recall measure onto the two diaries measures "true" expenditure.
b. All explanatory variables have been mean differenced.
c. * indicates p < 0.05.

procedure (B) or (C) is significantly related to household demographics.

To summarize, this analysis suggests that, if the diary measures are accurate, the measurement errors in recall food expenditure are large, do not satisfy the "classical" measurement error assumptions, and are not normally distributed.

In the inter-temporal consumption literature it is common to work with the logarithm of expenditure and to model the measurement error as multiplicative rather than additive. In this case assumption i. is replaced by $E[e^{\varepsilon}] = E[c/c^*] = 1$ and e^{ε} is typically assumed to be log-normally distributed. Thus ε, which is now the difference between $\ln c$ and $\ln c^*$ is normally distributed (but not with mean 0): $\varepsilon \sim N(-\sigma^2/2, \sigma^2)$. The assumption of independence of c^* (and hence $\ln c^*$) is maintained. Accordingly, we repeated the analysis described above, but working in logarithms, rather than levels, of food expenditure. The results are available in a supplemental appendix.[14] The results for logarithms are quite similar to those for levels. We find evidence of negative correlation between the measurement errors and true values, except where it is zero by construction. We also reject homoscedasticity, and normality of the errors. The signal-to-noise ratios are again quite low. The coefficients of the linear (in $\ln c^*$) terms of the Breusch-Pagan regressions are again strongly negative and their absolute value is larger by orders of magnitude than the positive coefficients on the quadratic terms. Thus the elasticity of the measurement error variance with respect to $\ln c^*$ is estimated to be negative at the mean of

[14] See Appendix Tables A2, A3 and A4, which parallel the format of Tables 4–6 respectively, and Appendix Fig. A3.

Table 6
Squared errors in recall food expenditure –regression on covariates (1996 Can $ per year).

	A		B		C	
	Coef	(Std Err)	Coef	(Std Err)	Coef	(Std Err)
ln pcy	358,423	(671,381)	*794,247	(283,436)	*406,701	(232,427)
(ln pcy)^2	1707	(7184)	−4293	(2608)	−431	(2257)
Log household size	−1,920,796	(7,401,677)	*−7,366,240	(3,419,272)	*−6,403,234	(2,654,562)
Presence of children (0–15)	−2,854,017	(1,851,392)	*−2,022,081	(762,104)	*−1,206,404	(462,913)
Presence of youths (16–24)	−395,027	(941,823)	663,679	(547,176)	*942,777	(368,182)
Presence of seniors (65+)	−1,141,612	(858,817)	−376,535	(479,492)	−150,472	(351,927)
2nd Earner in Household	−1,154,663	(1,009,227)	*−1,573,653	(637,308)	*−941,136	(507,728)
Constant	*9,245,786	(348,872)	*6,121,834	(208,103)	*4,160,967	(151,546)

Notes:
a. (A) Assumes first week diary measures "true" food expenditure. (B) assumes the average of 1st and 2nd week diaries measures "true" food expenditure. (C) Assumes the linear projection of the recall measure onto the two diaries measures "true" expenditure.
b. All explanatory variables have been mean differenced.
c. * indicates $p < 0.05$.

$\ln c^*$ (and, indeed, even at its 99th percentile). We find more evidence in logarithms (than in levels) that the mean of the measurement errors is systematically related to income and demographics. There is also considerable evidence that the measurement error variance is related to household demographics as well.

4. Conclusion

The Canadian FoodEx survey allows for the comparison of simple recall food expenditure questions with diary methods of collecting food expenditure data using a within-subject design. Simple recall food expenditure questions are widely used in multi-domain and longitudinal surveys in developed countries. For a validation study, a within-subject design has the important advantage that the difference between the data being assessed (here the recall data) and the superior data (the diary) can be calculated for each responding unit. If the superior data closely approximate the truth, these differences reveal the measurement errors in the data at the level of the responding unit. This in turn reveals key properties of the measurement error (such as whether the measurement error is correlated with the true value).

If we follow the literature in assuming that the diary information on food expenditure is very accurate (at least much more so than the recall data), the FoodEx data suggest that simple recall food expenditure questions generate large measurement errors, and that those measurement errors do not have the properties of "classical" measurement error. In particular, our analysis suggests that the measurement errors are negatively correlated with the true values. Put another way, the data strongly reject two pairs of joint hypotheses: that the diary records and recall estimates are both accurate, or, that the diary records are accurate and the recall estimates suffers from classical measurement error.

However, our analysis also documented evidence of several kinds of problems with the diary data (including infrequency and diary exhaustion). There is evidence of similar problems in other diary based expenditure surveys. For example, in a recent study Bee et al. (2015) report a series of analyses of data from separate diary and recall surveys conducted in the U.S. that seem to suggest the recall data is of better quality. If one is open to the possibility that the diary data contain substantial measurement error, or even that they measure expenditure well but over the period usually covered by diaries (one to two weeks) there can be substantial deviation from a longer run average of expenditure, then our results are subject to alternative interpretations. In that case, what we have studied is the difference (at the household level) of the measurement errors in the recall estimates and diary records. Some of the error properties we have documented might be attributable to the diary records. For example, significant purchase infrequency in the diary records would generate the (negative) mean dependence we observe.

Overall, our analysis of the FoodEx data suggests that superiority of diary data may not be as obvious as the literature suggests. This is an issue that could bear further scrutiny. Indeed, the relative merits of recall and diary methods for expenditure data have figured prominently in the recent debate about the redesign of the Consumer Expenditure surveys in the U.S. The main report of the National Academic of Science panel commissioned to consider redesign options recommended that in the future the survey should rely on supported self-completion. This means diaries, but diaries employing new technologies (Natl. Res. Counc., 2013). A dissent to that report argued that there was considerable evidence to suggest recall expenditure questions produced superior data to diaries, referring to the work of Bee, Meyer and Sullivan cited above, as well as an earlier version of the analysis reported in this paper.[15] There is a clear need for more robust evidence on the best way to collect expenditure information broadly, and food expenditure information in particular.

The FoodEx allows a comparison between a simple sequence of recall food expenditure questions that is typical of multi-domain and longitudinal surveys in more developed countries, and expenditure diaries as typically implemented in budget surveys in more developed countries. In both cases there may important differences with food expenditure surveys in developing countries. Recall-based expenditure surveys in developing countries may collect much more disaggregate information. Diary-based collection of expenditures in poor populations may not suffer from the fatigue and compliance problems evident in the FoodEx (and other developed-country diary-based surveys). Instead, key obstacles to diary completion in developing countries may revolve around respondent literacy and numeracy. It may also be possible to administer diaries in lower income countries in ways that cost and respondent burden might preclude in high income countries. For example, Beegle et al. (2012) report good success with intensively-supervised, personal expenditure diaries in Tanzania.

While these caveats are important, it is also true that many of the world's poor now live in middle income countries (Kanbur and Sumner, 2012). As such countries face new social and economic challenges, such as population ageing, their data needs will come to more closely resemble the data needs of developed countries. Multi-domain surveys that collect data on expenditure alongside information from other domains (work, health, demographics) allow researchers to study the context and causes of food spending patterns and food security. Longitudinal surveys with food expenditure questions allow for the study of

[15] In terms of new technologies, the main report of the commission argued that web diaries were not the right technology going forward, advocating instead providing respondent households with tablet computers with a diary application. However, the eventual redesign proposal of the BLS rejected tablets in favour of web diaries. http://www.bls.gov/cex/geminiproject.htm#news (last accessed: 07/12/2016).

the dynamics of food spending patterns and food security. Thus, evidence from Canada and other developed countries on the efficacy of different survey methods can usefully feed into forward-looking data strategies for low and middle income countries.

Acknowledgements

Author order is alphabetical. This paper draws on material that previously circulated in unpublished form as Ahmed et al. (2010), which in turn drew on material from Brzozowski's PhD. Dissertation at McMaster University. The analysis was carried out in the Statistics Canada Research Data Centre at McMaster University and the authors thank James Chowan for his assistance in accessing the data. The authors thank two anonymous referees, Alberto Zezza and Gero Carletto for very useful comments; Sule Alan, Chris Carroll, Martin Dooley, Sonia Laszlo and Michael Veall provided helpful comments on early versions of the work. The authors gratefully acknowledge financial support from the Social and Economic Dimensions of an Aging Population (SEDAP) Research Program at McMaster University, the Arts Research Board of McMaster University, and the Social Sciences and Humanities Research Council of Canada (SSHRC). Crossley also gratefully acknowledges support from the ESRC-funded Centre for Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies (Grant number RES-544-28-5001).

Appendix A. Supplementary materials

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.foodpol.2017.08.012.

References

Ahmed N., Brzozowski, M., Crossley, T.F., 2010. Measurement errors in recall food consumption data. Institute for Fiscal Studies Working Paper 06/21, London.

Andreski, P., Li, G., Samancioglu, M., Schoeni, R., 2014. "Estimates of annual consumption expenditures and its major components in the PSID in comparison to the CE" *American Economic Review*. Am. Econ. Assoc. 104 (5), 132–135. http://dx.doi.org/10.1257/aer.104.5.132.

Battistin, E., Padula, M., 2013. Survey instruments and the reports of consumption expenditures: evidence from the Consumer Expenditure Surveys. Unpublished manuscript, Univ. Padova, Univ. Venice http://dx.doi.org/10.1111/rssa.12127.

Battistin, E., Miniaci, R., Weber, G., 2003. What do we learn from recall consumption data? J. Human Resour. 38 (2), 354–385. http://dx.doi.org/10.2307/1558748.

Barrett, G., Levell, P., Milligan, K., 2015. A comparison of micro and macro expenditure measures across countries using differing survey methods. In: Carroll, C., Crossley, T. F., Sabelhaus, J., (Eds.), Improving the Measurement of Consumer Expenditures. NBER Studies in Income and Wealth, Volume 74. University of Chicago Press, Chicago, http://dx.doi.org/10.3386/w19544.

Bee, A., Meyer, B.D., Sullivan, J.X., 2015. The validity of consumption data: Are the consumer expenditure interview and diary surveys informative? In: Carroll, C., Crossley, T.F., Sabelhaus, J. (Eds.), Improving the Measurement of Consumer Expenditures. NBER Studies in Income and Wealth, Volume 74. University of Chicago Press, Chicago, http://dx.doi.org/10.3386/w18308.

Beegle, K., De Weerdt, J., Friedman, J., Gibson, J., 2012. Methods of household consumption measurement through surveys: experimental results from Tanzania. J. Dev. Econ. 98 (1), 3–18. http://dx.doi.org/10.1016/j.jdeveco.2011.11.001.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement Error in Survey Data. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, Volume 5. Elsevier Science, http://dx.doi.org/10.1016/S1573-4412(01)05012-7.

Browning, M., Crossley, T.F., Weber, G., 2003. Asking consumption questions in general purpose surveys. Econ. J. 113, F540–F567. http://dx.doi.org/10.1046/j.0013-0133.2003.00168.x.

Browning, M., Crossley, T.F., Winter, J.K., 2014. The measurement of household consumption expenditures. Ann. Rev. Econ. 6, 475–501. http://dx.doi.org/10.1146/annurev-economics-080213-041247.

Brzozowski, M., Crossley, T.F., 2011. Measuring the well-being of the poor with income or consumption: a canadian perspective. Can. J. Econ. 44 (1), 88–106. http://dx.doi.org/10.1111/j.1540-5982.2010.01624.x.

Crossley, T.F., Winter, J.K., 2015. Asking households about expenditures: what have we learned? In: In: Carroll, C., Crossley, T.F., Sabelhaus, J. (Eds.), Improving the Measurement of Consumer Expenditures, Studies in Income and Wealth, vol. 74 University of Chicago Press, Chicago.

Gibson, J., 2002. Why does the Engel method work? Food demand, economies of size and household survey methods. Oxford Bull. Econ. Stat. 64 (4), 341–359. http://dx.doi.org/10.1111/1468-0084.00023.

Gieseman, R., 1987. The consumer expenditure survey: quality control by comparative analysis. Monthly Labor Rev. 8–14.

Gray, P.G., 1955. The memory factor in social surveys. J. Am. Statist. Assoc. 50, 344–363. http://dx.doi.org/10.1080/01621459.1955.10501269.

Heitjan, D.F., Rubin, D.B., 1990. Inference from coarse data via multiple imputation with application to age heaping. J. Am. Statist. Assoc. 85, 304–314. http://dx.doi.org/10.1080/01621459.1990.10476202.

Hoderlein, S., Siflinger, B., Winter, J., 2015. Identification of Structural Models in the Presence of Measurement Error Due to Rounding in Survey Responses. Working Paper No. 869, Department of Economics, Boston College.

Kanbur, R., Sumner, A., 2012. Poor countries or poor people? Development assistance and the new geography of global poverty. J. Int. Dev. 24 (6), 686–695. http://dx.doi.org/10.1002/jid.2861.

Keen, M., 1986. Zero expenditures and the estimation of Engel curves. J. Appl. Econom. 1, 277–286. http://dx.doi.org/10.1002/jae.3950010305.

Kemsley, W.F.F., 1961. The household expenditure enquiry of the ministry of labour: variability in the 1953–1954 enquiry. J. Roy. Statist. Soc., Series C, 10(3) pp. 117–135. http://dx.doi.org/10.2307/2985204.

McWhinney, I., Champion, H.E., 1974. The Canadian experience with recall and diary methods in consumer expenditure surveys. Ann. Econ. Soc. Meas. 3 (2), 411–435.

Meghir, C., Robin, J.M., 1992. Frequency of purchase and the estimation of demand systems. J. Econom. 53, 53–85. http://dx.doi.org/10.1016/0304-4076(92)90080-B.

Natl. Res. Counc., 2013. Measuring What We Spend: Toward a New Consumer Expenditure Survey. http://dx.doi.org/10.17226/13520.

Neter, J., Waksberg, J., 1964. A study of response errors in expenditures data from household interviews. J. Am. Statist. Assoc. 59, 18–55. http://dx.doi.org/10.1080/01621459.1964.10480699.

Pudney, S., 1988. Estimating Engel curves: a generalisation of the P-Tobit model. Finnish Econ. Papers, 1.

Pudney, S., 1989. Modelling Individual Choice. Basil Blackwell, Oxford.

Runkle, D.E., 1991. Liquidity constraints and the permanent income hypothesis. J. Monet. Econ. 27, 73–98. http://dx.doi.org/10.1016/0304-3932(91)90005-9.

Silberstein, A.R., Scott, S., 1991. Expenditure diary surveys and their associated errors. In: Biermer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S. (Eds.), Measurement Errors in Surveys. Wiley, Hoboken NJ.

Statistics Canada, 1999. 1996 Food Expenditure Survey, Public-use Microdata Files. Statistics Canada, Income Statistics, Ottawa.

Stephens, M., 2003. "3rd of the Month": Do Social Security Recipients Smooth Consumption Between Checks? Am. Econ. Rev. 93 (1), 406–422. http://dx.doi.org/10.1257/000282803321455386.

Sudman, S., Ferber, R., 1971. Experiments in obtaining consumer expenditures by diary methods. J. Am. Statist. Assoc. 66, 725–735. http://dx.doi.org/10.1080/01621459.1971.10482336.

Sudman, S., Bradburn, N.M., Schwarz, N., 1996. Thinking About Answers. Jossey-Bass, San Francisco 10.2307/3152133.

Tanner, S., 1998. How much do consumers spend? Comparing the FES and National Accounts. In: Banks, J., Johnson, P. (Eds.), How Reliable Is the Family Expenditure Survey? Inst. Fisc. Stud, London, pp. 67–121.

Turner, R., 1961. Inter-week variations in expenditure recorded during a two-week survey of family expenditure. J. Roy. Statist. Soc., Series C 10 (3), 136–146. http://dx.doi.org/10.2307/2985205.