Consumer Expenditure Survey Measurement Error Study Phase 1

Proposal on Tracking Measurement Error in the CE

Roger Tourangeau
Westat

Brandon Kopp and Lucilla Tan
Bureau of Labor Statistics

Final: 2013.03.12

Introduction

The primary objective of this proposal is to provide recommendations for the development of metrics that can be used on an ongoing basis to track measurement error in the Consumer Expenditure Survey (CE) over time. These recommendations draw heavily from the Study Team's preceding work on the *Report on the State of Knowledge of Measurement Error for the CE* (Tourangeau et al., 2013), specifically, from the review of past research approaches adopted to investigate measurement error in the CE. If the recommendations made here are adopted, the error measures may also help shed light on the sources of reporting error and the expenditures categories that are more or less susceptible to reporting error.

Although the intent of this effort is to assess and track the level of *measurement* error in CE data, some of the proposed indicators (e.g., comparing CE estimates with external benchmarks) reflect all sources of error in the CE (e.g., nonresponse errors). Still, we think these indicators will be useful in tracking *the overall level of error* over time and across any changes in CE protocols.

This proposal is organized into five sections:

I. a summary of the strengths and weaknesses of past approaches taken to investigate measurement error in the CE;
II. a description of a multi-method-indicators (MMI) approach to track measurement error;
III. an outline of issues for further research that should be conducted *prior* to implementation of the MMI methodology;
IV. an outline of research projects that should be conducted *following* implementation of the MMI to further refine the methodology; and
V. an assessment of the components of the MMI methodology against a set of criteria to assist with resource planning for implementing this proposal.

I. Strengths and weaknesses of past methods used to investigate measurement error

A variety of methods have been used in the past to investigate measurement error in the CE. (For a more detailed review of each method, see Section III of Tourangeau et al., 2013). Each method has its characteristic flaws:

- Indicators that are based solely on CE data or information about the data collection process are at best indirect measures of error. Consider, for example, the proportion of CE respondents consulting bills or other records during the Interview Survey; although we expect reports based on record-aided recall to be more accurate than other reports, we cannot say how accurate either type of report is;

- External indicators (comparing estimates from the CE to an external data source) are also flawed. While item comparability is generally accounted for in comparing CE estimates to other data sources, the sources and magnitude of error in the external data sources are often either unknown or not taken into account.

- Record check studies require considerable staff resources and are expensive to carry out. In addition, the respondent provided records are likely to give incomplete coverage of the universe of expenses incurred during the reference period.

Table 1 highlights the major strengths and weaknesses of methods that have been used in the past to investigate measurement error in the CE.

Table 1. Major Strengths and Weaknesses of Past Methods for Studying CE Measurement Error

Method	Strength	Weakness
1. Internal comparisons		
1a. Interview Survey vs. Diary Survey 1b. Within a Survey: comparing across waves, or groups within a wave	No added data collection costs incurred to obtain these data. Easier to establish concordance of expenditure categories between diary and interview since CE controls both instruments. Differences in estimates between two comparison groups within an instrument is indicates error in one or both groups.	This method lacks an objective "true value". The measures chosen to represent the underlying "true," or at least better, values (e.g., the Diary Survey, the most recent month of the reference period, the first wave of a panel survey, etc.) are likely to suffer from their own measurement errors, which are generally also unknown. Between the Interview and Diary Surveys, the Diary Survey is typically assumed to be more accurate, on the assumption that diary keepers promptly record expenses so there is little recall error. However, there has yet to be a definitive validation of this assumption. The use of internal data comparisons to track measurement error before and after a redesign may be problematic because the changes made to the survey could affect both the "true" value and the comparison, making any changes in the apparent level of error difficult to interpret.
1c. Latent class analysis	No added data collection costs incurred to obtain these data. Can be used to identify potential predictors of reporting error. Response error latent variable for classifying respondents was a better measure of underreporting than any of the observed indicators taken individually.	Less effective in predicting the level of reporting error across latent classes: • "true" values are model-based. • these models often rely on very strong assumptions that may not be met, in practice.

Table 1. Major Strengths and Weaknesses of Past Methods for Studying CE Measurement Error

Method	Strength	Weakness
1d. Multilevel models	When validation data are available, this approach can be used to show which types of items are most susceptible to error and the item characteristics that can account for the variation in measurement error.	Lack of availability of validation data for sufficiently large, representative sample of respondents and item categories.
2. Comparison to external data sources		
2a. Personal Consumption Expenditure (PCE, National Accounts)	The PCE is the only *single* source of nationally representative data that covers the range of expenditure categories that the CE does.	Sources and magnitudes of error in the PCE are not well understood.
2b. Comparison to other household surveys	These data are based on household reports, like the CE. Some of these household surveys have built-in validation features – e.g., the Medical Expenditure Panel Survey (Household Component) uses medical provider information (regarded as less prone to error than household reports) to supplement or replace respondent reports; the Residential Energy Consumption Survey (RECS) obtains the household's energy billing data from the energy provider. Other surveys have high response rates which minimize non-response errors.	The sources and magnitude of error in these other surveys are not well understood. Adjustments may be needed to render estimates from the CE and the other household survey sources more comparable in scope and definition. Iterations of some surveys (e.g., RECS) are conducted too far apart to make them practical as an ongoing source of comparisons.
3. Validation (records check) of reported expenditures	A validation approach comes closest to obtaining true values for verifying a respondent's reported expenditure.	A representative sample is costly and resource intensive to conduct. Past CE validation studies have used small, convenience samples. It is difficult to ensure that every reported expense generates a record, or to verify reports that are omitted or included in error.
4. Check of balance between expenditures and income (after accounting for savings and debt)	The size of the gap between income and total expenditures is an easily understood indicator of the accuracy in overall reporting,	Reporting accuracy is not measured at the item or category level.

Table 1. Major Strengths and Weaknesses of Past Methods for Studying CE Measurement Error

Method	Strength	Weakness
	since the money coming into or going out of the household is accounted for.	Small scale lab study indicated that conducting real-time calculations based on respondent reports and providing interviewers and respondents with feedback useful for improving survey reports was not viable. Comprehensive measurements are required for all components of the balance check.

Source: Section III of Tourangeau et al., 2013

As we noted in our State of Knowledge report, the Geisen et al. (2011) study, which compared survey reports with respondent records, leads to different conclusions from the studies that compare CE estimates to external benchmarks. Their study suggests that overreporting and underreporting of expenditure amounts occur about equally often in the CE. In contrast, the CE-PCE comparisons almost uniformly suggest that the CE estimates are underestimates. The varied strengths and weaknesses among the methods imply no single method or indicator can adequately assess measurement error, and thus we advocate a multi-method-indicators (MMI) approach to track measurement error in the CE.

II. Proposed multi-method indicators (MMI) approach

The approach we propose to assess and monitor measurement error consists of three components:

1) Internal indicators, which are based solely on the CE data themselves and paradata (other information regarding the CE data collection process);
2) External indicators, which involve comparisons of key CE survey estimates to similar estimates from other data sources; and
3) Periodic but regular record check studies, including those incorporated into the production sample (e.g., via a methods panel).

The internal indicators and external indicators together make up a "vector of flawed indicators." The indicators are flawed because each of them is susceptible to measurement and other errors, but each provides useful, easily understood measures of overall error. In addition to identifying these indicators, ongoing research will also be needed to investigate the inter-relationship between the indicators.

The third component of the MMI approach, record check/validation studies, could be used to show which item categories are reported well or poorly in the CE. This information also provides a basis for confidence in findings from the comparison of these categories to external benchmarks, given no changes in the methodology of the benchmark source. For example, if the validation study suggests a given expenditure is underreported, we can be more confident that an increase in the CE/external benchmark ratio represents improved reporting in the CE. We believe it is important to move away from assuming that higher levels of reporting indicate higher levels of accuracy.

If all three components suggest that CE estimates are improving or getting worse in terms of overall level of error, it will provide greater confidence than trends based on any single approach can. Differences across indicators may suggest that the CE is getting more or less accurate for some expenditure categories but not for others, or may suggest that some indicators are too flawed to be useful and should be dropped. We describe the three components in more detail below and offer preliminary recommendations about indicators to include in each component.

II.1. Internal Indicators

The first component of the MMI approach encompasses internal indicators derived from CE data and paradata. A number of investigators have attempted to estimate the level of measurement error in the CE using data from the CE itself. For example, Yan and Copeland (2010) compared Wave 2 reports in the CE Quarterly Interview survey with those from later rounds (see also Shields and To, 2005, and Silberstein, 1990, for similar efforts). In a series of papers from 2004 to 2011, Tucker, Meekins, and Biemer (2004, 2008, 2010, 2011a, 2011b) have examined a variety of internal error indicators using latent class models; they have identified a set of variables that seemed to cluster together, allowing them to classify CE respondents into three groups with different overall levels of reporting error. Based on our review of this past work, we have identified several indicators that seem particular promising to us and that are relatively inexpensive to collect and monitor.

Selection criteria. The criteria for selecting internal indicators for inclusion in the MMI approach are the following:

1) Relevance to different sources of reporting error, such as recall error, conditioning through the survey experience itself, and satisficing;
2) Past research showing the indicators are highly predictive of reporting errors;
3) Availability of the data to produce the indicator repeatedly over time;
4) Usefulness for improving survey operations (we intentionally exclude respondent demographic characteristics even though some of these characteristics have been associated with poor reporting behavior)

An illustrative set of internal indicators might be:

- In the Diary Study, interviewer assessments of the diary keeper's level of diligence in recording of entries in the diary before pickup versus data collected by recall ;
- The ratio of the number of entries in diary week one and diary week two;
- The percentage of respondents who use records during the Quarterly Interviews;
- The average length of the interviews;
- The average number of contact attempts needed to complete Quarterly Interviews.

These five indicators reflect different potential sources of measurement error. The first and the third are measures of the likely level of recall error in the CE data. As we noted in the State of Knowledge report, respondents may forget a purchase entirely or may remember the purchase but forget the amount. Filling out the Diary ahead of time or consulting records during the Interview are likely to reduce the frequency of such errors. The decline in reporting over time (our second indicator above) is thought to reflect —conditioning," particularly for regularly purchased items; conditioning refers to the increased tendency for respondents to take shortcuts the longer they participate in a diary or panel survey. The final internal indicator — the average of contact attempts — presumably reflects the Consumer

Unit's (CU) overall reluctance to take part in the survey and the level of effort the members of the CU are likely to put into providing accurate data during the Quarterly Interview.

II.2. External Indicators

A common method for estimating the accuracy of a survey is to compare the survey estimates to some external benchmark and this forms the second component of the MMI approach. In the best case, the external benchmark is error-free or at least much less error-prone than the survey of interest. In reality, external benchmarks are likely to have error from various sources. One of the most commonly used benchmarks for the CE is the Personal Consumption Expenditure (PCE) component of the National Income and Product Accounts (NIPA). Unfortunately, the accuracy of the PCE estimates is not entirely clear. Secondly, disparities between the CE estimate and the external estimate reflect not only measurement error but also differences in coverage, nonresponse, sampling, and other sources of error.

Despite these problems, we believe it will be helpful to compare CE estimates to several external benchmarks. These comparisons have conventionally been reported as ratios (CE estimate/External estimate). The CE estimates have generally been lower than the NIPA PCE estimates for a given expenditure category, but are sometimes lower and sometimes higher than comparable estimates from other surveys, such as the American Community Survey (ACS) and the Medical Expenditure Panel Survey (MEPS) (see Tables 5a through 5f in Tourangeau et al., 2013). We recommend that BLS use several benchmark comparisons for any given category rather than relying on just one benchmark, giving more weight to the *trend* in the ratios than to their absolute values.

Selection criteria for categories for comparison to external data. The categories selected should be representative as a set (that is, they should cover a range of different expenditure types) and should have the following additional features:

1) Cover categories that differ in the likely availability of records;
2) Include both regular (e.g., rent/mortgage, utilities) and irregular (e.g., clothing) expenditure categories ;
3) Cover both large and small expenditures;
4) Focus on categories in which the external source uses a definition that is reasonably consistent with the CE definition. Where possible, the sources of error in the external source and the magnitudes of these errors should be made explicit (to the extent that they have been identified and quantified).

The following is an illustrative list of potential external benchmarks; the percentage given in parenthesis indicates the average expenditure for each category as a share of average total expenditure from the official CE data table (*Table 52: Shares of average annual expenditures and sources of income, Consumer Expenditure Survey, 2011*):

1) Comparisons with other surveys

    a) ACS estimates for rent (6.1%) and mortgage (6.4%);
    b) ACS estimates for utilities and fuel (7.5%);
    c) Residential Energy Consumption Survey (RECS) estimates for utilities and fuel;
    d) MEPS estimates for hospitalization and health insurance (Healthcare 6.7%);
    e) MEPS estimates for medical and health;
    f) PSID estimates for medical and health.

2) Comparisons with PCE

   The proposed categories for comparisons with the PCE are among those that have the greatest comparability between the two sources in terms of definitional concordance (see SOK Report Table 6, Passero (2012)).

   a) Household appliances (major and small appliances 0.6%);
   b) Rent (6.1%) and utilities (7.5%);
   c) Food purchased offsite (Food away 5.3%);
   d) Women's and girl's clothing (1.5%);
   e) Men's and boy's clothing (0.8%).

II.3. Periodic Record Check Study

The final component of the MMI approach is to compare the survey reports to an external record at the respondent level. This information establishes a basis for confidence in findings from the comparison of these categories to external benchmarks. Geisen et al. (2011) used this approach to evaluate CE reports with a small convenience sample. They administered an abbreviated version of the CE Interview Survey, asked respondents to collect financial records for any expenditures covered by the survey, and then compared the reported answers with the values indicated on the records.

We propose that BLS adopt a record validation approach similar to that used by Geisen et al. (2011). Validation studies should be carried out regularly, though not necessarily with as high a frequency as the other methods proposed in this report. To help reduce costs, the respondent burden involved in collecting records, and the burden on BLS to process the data from those records, a subset of items or CE categories could be chosen for this periodic evaluation.

Selection criteria for categories to be included in validation study. By design, these categories should be similar to those used in the external benchmarking indicators because it would be useful to compare the conclusions derived from different methods for the same expenditure categories. Thus, following the illustrative categories proposed in Section II.2, the categories for periodic record check might include:

1) Women's and girl's clothing;
2) Men's and boy's clothing;
3) Rent and utilities;
4) Food purchased offsite; and
5) Hospitalization and health insurance.

III. Issues for Preliminary Research before Implementation

We recommend several lines of research before attempting to implement the collection and monitoring of these error indicators.

1) Investigate the feasibility and cost of collecting the internal and external indicators recommended here on a routine basis.

2) Conduct feasibility tests to develop better protocols for obtaining records for more expenditure categories and for a higher percentage of survey reports. While the Geisen et al. (2011) Records Study and the Sjoblom and Lee (2012) Records Information and Feasibility of Use Study provided useful information about the viability of this approach, several questions remain about how it can be used in a production or program evaluation environment. These questions can/should be addressed in smaller scale studies prior to full implementation. Some issues for further investigation for these preliminary studies include:

    a) Determine whether it is less burdensome to ask the respondent to collect records of all their expenses or to collect records only for select categories.

    b) Explore what expenditure categories and what types of records raise privacy concerns, making record-based data collection more difficult and make those records that are provided less representative of the population. Future research should address the issue of nonresponse in record collection and assess the potential for bias.

    c) Determine the appropriate sample size for a records validation study that would be used to make inferences about the nature of measurement error in the CE and make recommendations for future changes. There is variability in respondents' ability to accurately report their expenditures, the likelihood the expenditures generate records, and the ability of the respondents to retrieve the records. To produce reliable results, a large sample will be needed. If finer detail is desired for demographic characteristics (e.g., income level, household size) then an even larger sample would be needed. The Geisen et al. (2011) and the Sjoblom and Lee (2012) studies had 115 and 152 participants, respectively. From these studies, it is clear that the sample should be much larger (≈1000 completed cases) and, if possible, it should be nationally representative. One possibility is to use a subsample of an outgoing cohort of the CE Interview Survey sample or to incorporate validation within an ongoing methods panel.

    d) Find out how we can obtain records for a higher percentage of reported expenditures. Geisen and her colleagues were able to compare only 36 percent of the reported expenditures to household records. Participants in that study were asked following the initial interview to collect records from the previous three months. They were not warned to do this ahead of time so could only collect the records they had held on to. The authors of that report suggest asking respondents in advance to gather relevant records prior to the interview. Future research should be directed toward developing procedures that maximize respondent participation in the record collection task (including the use of incentives) while minimizing the potential for bias.

    e) Attempt to access respondents' electronic records more effectively. It is also clear from both the Geisen et al. (2011) and the Sjoblom and Lee (2012) studies that many respondents were either unable or unwilling to provide printed copies of electronic financial records. These types of records are likely to become more common as time

goes on, so their collection would be of particular importance. Any future records validation study would need to motivate respondents to provide these records, either through monetary incentives or by making the process of printing or uploading the records easier and less costly for the respondent.

f) Find methods to improve on the Geisen et al. (2011) methodology to measure under- and overreporting of expenditures (as opposed to amounts). In the methodology used in that study, if a respondent failed to report an expenditure for which there was no record, there was no way to detect the underreporting of that expenditure. Poor memory for an expenditure and lack of a receipt are both likely when low-dollar value expenditures are made. Collecting secondary records, such as bank or credit card statements could identify some of these missing records/reports, but not those made using cash. To detect overreporting, it may require asking respondents to collect records from periods *before* the reference period used in the interview and using those records to check whether and how often expenses incurred prior to the reference period were erroneously reported.

IV. Issues for ongoing research after implementation

1) Once the external indicators begin to be collected, their interrelationships should be examined on an ongoing basis. Are the indicators unidimensional or multidimensional? One way to determine this would be to create a time series of indicators and to carry out a factor analysis of the indicator values. It may be that the relationships among the indicators are captured by a few factors. If so, the factor loadings could serve as weights for one or more composites that combine several or all of the indicators.

2) A similar type of factor analysis should also be conducted for the set of internal indicators.

3) Ongoing research to identify more effective internal indicators will be needed, especially if the CE survey design changes.

V. Additional considerations regarding implementation

The resources needed to carry out an approach like the one proposed here is a critical consideration for its implementation. As a first step towards acquiring this understanding, we offer a preliminary set of criteria and describe them for each component of the MMI in Table 2.

- **Cost:** What inputs are needed to develop each MMI component?
- **Duration for development**: How long will the development efforts take?
- **Applicability**: Is the component applicable only to the current CE design or will it remain applicable to other designs?
- **Periodicity**: How often can the indicators be tracked?

We emphasize that our estimates regarding the costs and durations in Table 2 are intended only to highlight the relative differences across the components of the MMI methodology and are not intended to represent estimates of actual cost and duration.

Table 2. Considerations for MMI Implementation Planning

	Internal indicators	**External comparisons**	**Periodic Record Check Studies**
Cost	Very low cost Primary resource need is staff time to conduct analysis of the data. Additional cost may be incurred if additional data are needed and require cognitive testing of new questions.	Very low cost Primary resource need is staff time to conduct analysis of the data.	High Cost (> $500k) Primary tasks involve: 1) Development of effective records collection protocols, 2) Interviewer training, 3) Instrument development, 4) Systems development to process collected records.
Duration for development	< 1 year	< 1 year for sources currently used. 1-2 years to find and analyze other sources for additional expenditure categories.	> 2 years
Applicable only to current CE design	No	No	Remains relevant if respondent recall remains a key feature of any future design
Periodicity	Quarterly	Annual (depends on the periodicity of data release for external benchmarks).	Every 3 to 5 years

Note: Cost and duration estimates in this table are intended only to highlight the relative differences between the components of the MMI methodology.

Summary

Prior research to assess measurement error in the CE has used a variety of methods, ranging from small scale cognitive and records validation studies, to comparison with other data sources, to multivariate models of varying statistical complexity. Each of these methods focused on different aspects of measurement error (e.g., estimating the magnitude of measurement of error, distributional features of measurement errors, characteristics of respondents or items that are correlated with misreporting expenditures). The varied strengths and weaknesses of the methods imply that, although no single method or indicator can adequately assess measurement error, each method offers potentially useful information about the nature of measurement error in the CE surveys. Thus, we recommend a multi-method-indicators (MMI) approach to track measurement error in the CE. We recommend that this approach include three components: the use of CE data and paradata, comparisons with external benchmarks, and periodic but regular record validation studies. This approach offers the potential for convergent validity, or the triangulation of measurement error in specific item categories, that no single indicator can provide.

We do not provide a definitive list of internal and external indicators for the MMI approach, but instead provide criteria for their selection and offer recommendations for pre- and post-implementation research to provide additional evidence about the usefulness of the indicators that are ultimately selected.

Our proposed MMI methodology builds on previous research, as well as experience with small scale records collection studies. This increases the likelihood of the proposal to meet one of the overall project goals of *~~h~~aving in place, practical and replicable methods and /or metrics for monitoring and evaluating changes in measurement error by the end of 2014"* (see *Table 1. Overall goals of the CE Measurement Error, No. 4* in the project Statement of Work, dated July 23, 2012).

References

Geisen, E., Richards, A., Strohm, C., and Wang, J. (2011). *U.S. Consumer Expenditure Records Study Final Report*.

Shields, J. and To, N. (2005). Learning to Say No: Conditioned Underreporting in an Expenditure Survey. Paper presented at 2005 AAPOR.

Silberstein, A. (1990). First wave effects in the U.S. Consumer Expenditure Interview Survey. *Statistical Methods, Canada*, *16*, 293-304.

Sjoblom, M., and Lisa, L. (2012). *Records Information and Feasibility Use Study Final Report*.

Tourangeau, R., Fricker, S., Kopp, B., and Tan, L. (2013). *Report on the State of Knowledge of Measurement Error in the CE*. Washington, DC: Division of Consumer Expenditure Survey.

Tucker, C., Biemer, P., and Meekins, B. (2004). Estimating the level of under-reporting of expenditures among expenditure reporters: A micro-level latent class analysis. Presented at the Joint Statistical Meetings, August, 2004. Toronto, Canada.

Tucker, C., Biemer, P., and Meekins, B. (2008). A micro-level latent class model for measurement error in the Consumer Expenditure Interview Survey. In *Proceedings of the Section on Survey Methods Research* (pp. 2171-2178). Alexandria, VA: American Statistical Association.

Tucker, C., Biemer, P., and Meekins, B. (2010). Latent class analysis of Consumer Expenditure Reports. . In *Proceedings of the Section on Survey Methods Research* (pp. 5441-52). Alexandria, VA: American Statistical Association.

Tucker, C., Biemer, P., and Meekins, B. (2011a). Latent class analysis of measurement error in the Consumer Expenditure Survey. In *Proceedings of the Section on Survey Methods Research* (pp. 5218-5229). Alexandria, VA: American Statistical Association.

Tucker, C., Biemer, P., and Meekins, B. (2011b). Estimating underreporting of consumer expenditures using Markov latent class analysis. *Survey Research Methods*, 5(2): 39-51

Yan, T., and Copeland, K. (2010). *Panel Conditioning in Consumer Expenditure Interview Survey*. Final Report from NORC.