Journal of Official Statistics
Vol. 8. No. 1, 1992, pp. 41–61

The Estimation of Instrument Effects on Data Quality in the Consumer Expenditure Diary Survey

Clyde Tucker[1]

Abstract: The study reported here examines the effects of procedural variations on nonsampling errors. Measures of errors due to both response and item nonresponse are created from information contained within the survey itself. A method for combining these measures into a single data quality indicator is illustrated. Important interactions between diary procedure and respondent characteristics are found.

Key words: Response error; item nonresponse; latent structure analysis; data quality; age effects.

1. Introduction

Diaries have been used extensively to collect data in fields as diverse as transportation and health (Roghmann and Haggerty 1972; Thompson, Carlson, Woteki, and Vagts 1980). Diaries also have been an important source of information on consumer spending (Flueck, Waksberg, and Kaitz 1971; Pearl and Levine 1971). The U.S. Bureau of Labor Statistics (BLS), in conjunction with the U.S. Bureau of the Census, conducted a consumer expenditure survey that included a diary during 1972 and 1973 in the United States. A similar survey has been ongoing in the United States since 1980.

Much research has been devoted to the topic of consumer expenditure diary methodology. Several studies have compared the differences in the estimates from personal interviews involving recall and those from diaries (Neter 1970; Stanton and Tucci 1982). Variations in diary procedures also have been examined (Grootaert 1986; Kemsley and Nicholson 1960; Sudman and Ferber 1971). The study reported here continues in the tradition of those which have considered the effects of procedural variations.

In 1985, BLS and the U.S. Census Bureau conducted a field test of three sets of diary procedures for collecting consumer expenditure information. These procedures differed with respect to the level of structure of the diary, the types of cues given the respondent, and the method for collecting recalled expenditures. In this paper, indicators of measurement errors arising from both response and item nonresponse using the three procedures are created, and a method for combining them into a single data quality indicator is illustrated. Using these indicators,

[1] Mathematical Statistician, Office of Prices and Living Conditions, Statistical Methods Division, U.S. Bureau of Labor Statistics, Washington, D.C. 20212, U.S.A.
Acknowledgement: The author thanks Mike Kaluba and Steve Hill for their programming efforts. Thanks are also extended to Adriana Silberstein, Leslie Miller, and the referees for their helpful comments. Any opinions expressed are those of the author and do not constitute policy of the U.S. Bureau of Labor Statistics.

the effects on nonsampling error of diary procedure, working in conjunction with other causal variables, are examined.

2. The Consumer Expenditure Diary Survey

The Consumer Expenditure Diary Survey (CED) is conducted by the U.S. Bureau of the Census for BLS and provides, along with the Consumer Expenditure Quarterly Interview (CEQ), the information needed to construct the cost weights for the Consumer Price Index. The data also are used for economic analysis. Although the diary was designed to collect all daily expenditures over a two-week period, it is especially effective for gathering information about small, frequently purchased items which are normally difficult to recall over an extended period. These expenditures include grocery items, meals eaten out, household supplies and personal care products and services. In addition to the expenditures, data also are collected on the income, work experience, and demographic characteristics of household members.

The unit of analysis in the CED, and the level at which most data are collected, is the consumer unit (CU). A CU is defined as one of the following: (1) the collection of all members of a household who are related by blood, marriage, adoption or other legal arrangement; (2) a person living alone or sharing a household with others or living as a roomer in a private home or lodging house or in a permanent living quarters in a hotel or motel, but who is financially independent; or (3) two or more persons who live together and pool their incomes to make joint expenditure decisions. To be considered financially independent, at least two of the three major expense categories (housing, food, and other living expenses) have to be provided by the respondent. For more information about the survey, see U.S. Department of Labor (1986).

3. The Diary Operational Test

Previous research on the CED has demonstrated that a number of factors influence the response (Tucker 1988). Respondent characteristics, in combination with environmental circumstances and the intervening survey procedures (including interviewer characteristics), affect both the respondent's attitudes toward the survey and his or her record-keeping behaviors. Attitudes are not directly responsible for the outcome of the survey process; the record-keeping behaviors are. Attitudes and record-keeping behaviors will coincide in many cases, but the research cited above has shown that this is not always true.

The analysis of this process until recently focused on the contributions of respondent and environmental characteristics to the shaping of the response. The role of survey procedures was ignored because they were assumed to be constant across all respondents. Or, in any case, what differences there might have been went unmeasured. The purpose of the 1985 Diary Operational Test was to evaluate the effects of procedural changes in the form of new diary formats and recall method. Each of these changes, outlined in Table 1, is not studied individually but, instead, is viewed as part of a total package of procedures.

A diary's format can either hinder or facilitate the reporting of expenditures. Research on item reporting rates from both the 1972/73 and 1980/81 CED indicated that explicit references to particular products in the diary increase the likelihood that these items will be reported, especially if the reporting rates are low to begin with (Jacobs 1983; Tucker 1984). In order to evaluate diary formats which provide more explicit

Table 1. Diary operational test design

	Current (Control) diary	Experimental diaries	
		Nonspecific	Specific
Sample	Retired CPS panel from 1979		
Interviewers	New to CED		
Content	– All daily expenditures	– Clothing, shoes, jewelry and some other non-food items excluded	
	– Quantity and weight	– No quantity and weight	
Recall data method	– Diary check for selected items, no scripted recall	– Scripted recall section	
Format	– Category titles and general descriptions on the left	– Category titles and general but more complete descriptions above	– Category titles only above
	– Blank lines for recording on the right	– Blank lines for recording below	– Lines with printed items to check below
Response to survey	*CUs*	*CUs*	*CUs*
– Completed both diaries	824	837	812
– Completed one diary	104	71	78
Other week:			
No one home	22	9	14
Temp. absent	19	15	15
Refused	39	24	31
Other	19	14	15
Out-of-scope	5	9	3
– Completed no diaries	208	192	219
No one home for both	29	21	23
Temp. absent for both	26	17	19
Refused both	119	118	145
Other	34	36	32
– Out-of-Scope both weeks	133	141	135
– Total attempted	1269	1241	1244

instructions as to the commodities to be reported, two new, experimental diaries were developed and compared to the current diary in a field test.

Respondent burden also was reduced to further enhance reporting. The experimental diaries cover fewer expenditure categories than the current diary. In particular, the clothing, shoes and jewelry section has been eliminated. These new diaries also are smaller and have more attractive covers. Respondents are not required to specify the quantity and weight of the items when completing the two experimental diaries, as they are in the current diary. What distinguishes the experimental diaries from one another is the specificity of the item descriptions within each section. Experimental diary A, the nonspecific diary, only has blank lines for recording purchases under each of the section headings, just like the current diary; but, in contrast to the latter, the section headings contain more complete, but still fairly general, descriptions of the items to be reported. Experimental diary B, the specific diary, has only category titles; however, the lines beneath each heading have specific items printed on them. Respondents need only check if an item was purchased and record the price.

These two formats were chosen as the most promising alternatives for improving data quality. Yet, they have their own shortcomings. The nonspecific diary still gives the respondent the freedom to describe purchases; but, like the current diary, this means a significant amount of writing. Much less writing is necessary in the specific diary, but the respondent must make classification choices which will take more thought and which are prone to errors (Tucker, Vitrano, Miller, and Doddy 1989). Respondents also must add expenditures for all items appearing on the same line.

Another feature of this experiment was a new method for collecting recalled expenditures. Currently, the interviewer records these expenditures directly into the diary using unscripted procedures and also asks a series of followup questions called diary check items about specific commodities which the respondent may have forgotten to report. The new procedures, used with the experimental diaries, involve a scripted recall section contained in the household characteristics questionnaire.

Finally, a diary assessment section was included in the test. This section contains questions which measure the respondent's attitudes and diary-keeping behavior. There are questions for both the respondent and interviewer to answer.

The research sample used in the test was a retired Current Population Survey (CPS) sample from 1979. Households in this sample had not been in the CPS since January or February of 1979. Twenty-two large cities or primary sampling units (PSUs) were surveyed. The sample was clustered, and the two experimental diaries and a control (the current diary and recall method) were interpenetrated within the clusters. Each of the three diaries was randomly assigned to a third of the sample units in each city.

In addition to the design information, Table 1 gives the results of attempts to place two diaries with the units in each of the three subsamples. Excluding those units which were out-of-scope both weeks (vacant dwellings, etc.), there were 3,345 consumer units which could have completed at least one diary. Of these, 2,473 CUs (about 74%) completed both diaries, and another 253 (almost 8%) completed one of the diaries. CUs which were temporarily absent, or out-of-town, both weeks were considered to have no expenditures while in the PSU, the only expenditures meant to be measured by the diary. Excluding these cases, the unit nonresponse (eligible CUs completing no

diaries) was 16% in both the control and nonspecific treatment groups and 18% in the specific group. The higher nonresponse rate in the specific subsample is due to the greater number of CUs refusing to keep either diary. About 73% of the eligible units in the control and specific groups completed both diaries while 76% of those in the nonspecific did.

4. The Analytical Design

4.1. Overview

An extensive analysis of aggregate performance measures for each procedure already has been conducted (Tucker and Bennett 1988). This analysis indicated that respondents to the specific diary performed better than those using the other two, but a micro-level examination of measurement errors has not been undertaken. Using measures of micro-level data quality developed in previous studies of the current diary (Tucker 1988), a more complete investigation of the survey process is carried out. This analysis is limited to the 2,473 CUs completing both weeks of the diary to ensure data comparability.

4.2. Independent variables

The procedural condition variable (defined by the three pairings of format with recall method) is the focus of the study, but its effect must be evaluated in conjunction with other independent variables. In this case, these variables are limited to those found to be related to data quality in previous analyses of the CED (Tucker 1988). Their distributions within each treatment cell are given in Table 2.

Although the three diaries were assigned randomly within the clusters, there are some significant differences between the demographic distributions across the three conditions. The respondents in the control condition are somewhat younger and better educated than those keeping the specific diary. On the other hand, the specific and control conditions have more single adult respondents in their samples than does the nonspecific. The nonspecific sample also has more respondents living in central cities. These demographic differences could be the result of the differential nonresponse already discussed and will be considered later in the context of the findings.

The statistically significant differences between treatments noted in the previous paragraph are small in absolute terms, and there is some question as to how meaningful they really are. The CPS design is complex; but, because the CPS clusters were interpenetrated with the three treatments, the design effect is usually less than 1.0 (where the variation generated from the complex design is compared to simple random sampling or SRS variation) when procedural condition is an independent variable. In fact, it can be as little as .2 for variables which are highly clustered, such as, region, degree of urbanization, and ethnicity. On the other hand, when procedural condition is not being considered, so that the analysis cuts across the interpenetration, the design effect is almost always larger than one. The half-sample replicate method in CPLX was used to carry out these analyses (Fay 1983).

4.3. Dependent variables

4.3.1. Response error

One dependent variable measures the response error in expenditure reports. This measure is based upon the following assumptions:

1. There are patterns in the information given by respondents which are related to the level of response error.
2. Various indicators of these response patterns can be developed.

Table 2. Distributions for selected consumer unit characteristics by procedural condition

Characteristic	Specific recall $N^1 = 812$	Nonspecific recall $N^1 = 822$	Control diary check $N^1 = 839$
Age of Reference Person[2]			
Under 25	6.7%	6.4%	9.4%
25–44	44.4	49.2	45.4
45–64	29.5	27.2	27.4
65+	19.4	17.2	17.8
Education of Reference Person[2]			
Less Than H.S.	25.9	23.0	22.2
High School	30.5	28.6	30.3
Post High School	43.6	48.4	47.5
Ethnicity of Reference Person[3]			
Black or Hispanic	18.5	19.0	17.5
Other	81.5	81.0	82.5
Composition of CU[2]			
Husband/Wife	52.9	54.5	53.4
Single Parent/Single	36.2	32.7	36.5
Other	10.9	12.8	10.1
Degree of Urbanization[2]			
Central City	35.0	36.8	33.6
Other in SMSA	65.0	63.2	66.4
Region[3]			
Northeast	15.1	15.9	17.2
North Central	35.5	35.6	35.2
South	21.3	21.5	19.7
West	28.1	27.0	27.9
CU Tenure[3]			
Owner	58.4	58.1	58.0
Nonowner	41.6	41.9	42.0

[1] N = Number of CUs. The numbers of CUs differ from Table 1 because these are weighted.
[2] Chi-square significant at the 0.01 level.
[3] Chi-square not significant at the 0.05 level.

3. Reasonable judgments can be made about the substance of the relationships between the indicators and response errors.
4. The associations among these pattern indicators can be used to model an ordinal "latent" response error variable.

The central assumption here is that the level of response error can be determined from the manner in which the respondent reports information. Traditionally, reinterviews or independent sources have been used to identify response error (Corby and Miskura 1985; Groves and Magilavy 1984; Madow 1973; Sudman and Bradburn 1974),

but these methods have serious shortcomings. Reinterviews can be quite expensive and produce the same or different errors. Since reinterviewing is seldom done on the entire sample, inferences about those not reinterviewed must be based on what is often a self-selected sample. Furthermore, when the phenomena are transient, reinterviews may not be appropriate. Independent sources, on the other hand, may not be available or may be limited to an unrepresentative subset. Available sources may not be accurate or even truly comparable.

An alternative to these methods is the use of information from the survey itself. Patterns in an individual's responses can indicate the extent of response error for variables of interest. Little cost is incurred, and no new interviewing procedures need be developed nor independent sources found. Perhaps of greatest importance, generalizations from a subset are avoided and the problem of self-selection is eliminated.

Whether or not useful patterns can be identified depends on the particular survey. The search for response pattern indicators in the CED (and the Diary Operational Test, in particular) is made easier by the fact that a large body of information on consumer unit characteristics, expenditures, attitudes and behavior was collected. Furthermore, the expenditure information covers a period of time (two weeks) long enough to ascertain patterns in the reporting. Identifying the indicators, however, requires an understanding of both the psycho-social dynamics of the survey situation and the substantive nature of the response.

4.3.2. Item nonresponse

The following steps are carried out to construct a micro-level measure of item nonresponse:

1. Nonresponse is measured at the item level where a valid response either is present or is not.
2. Indicators of the amount of item nonresponse in each section of the CED are constructed using these indicator variables.
3. The various sections are weighted according to their assumed effects on substantive results.
4. An additive measure of the products of the item nonresponse indicators and weights is created.
5. Based on an inspection of its distribution, this summary measure is divided into categories representing meaningful distinctions in the level of item nonresponse.

Items are measured as present or absent, and respondents are differentiated according to the proportion of information provided (Thran, Marder, and Willke 1986; Tucker 1988). The more complex the survey, and the CED is complex, the more these proportions will vary. The weights for items or sections may or may not depart from unity, depending on the researcher's judgment. The measure described in step 5 is an ordinal one, like response error.

4.3.3. Data quality

The measures of response error in the expenditure report and item nonresponse are, themselves, quality measures. As such, they are analyzed as separate outcomes of the survey process, but a method for combining these measures also is investigated. This endeavor is a first step towards developing a measure of overall data quality. In combining a measure of response error and one of item nonresponse into a single indicator of data quality it is important to understand that response error and nonresponse both have a detrimental effect on quality, in that they both contribute to measurement error. The measurement error

usually should be greater for an adjusted nonresponse than for a response. This measurement error is related to the range of possible values and the frequency of their occurrence.

5. The Response Error Measure

5.1. *Response pattern indicators*

In developing response pattern indicators the assumption is made that most response errors in the diaries occur in the form of underreports. It is difficult to imagine an individual recording more items than were purchased or even consistently overreporting the price of items. On the other hand, the failure to report all items is quite likely given the time and effort required to keep the diary. Substantial information exists to support this assertion (Pearl 1979; Sudman and Ferber 1971).

Four response pattern indicators are formed from information contained in the survey. They are used to measure the error in what might be termed "typical" grocery expenditures and meals away from home. These expenditure items (shown in Table 3) are most of the ones which were collected on all three diaries. Grocery items include food and other purchases (e.g., personal care products, household supplies and nonprescription drugs) usually made at a grocery store. The items chosen are ones which most consumer units purchase frequently and are also those for which the diary was designed to collect information.

The first response pattern indicator compares the expenditures for the items in Table 3 which the respondent reported in the first week to those reported during the second week. Other consumer diary research (Pearl 1979; Silberstein and Scott 1991; Sudman and Ferber 1971; Turner 1961) has shown that first-week expenditure estimates tend to be higher than those in the second week, perhaps indicating underreporting, at least in the second week. The particular measure used here was computed by taking the difference between the expenditures for the two weeks (first week minus second

Table 3. Expenditure classes included in the analysis

I. Food at Home
   a. Flour, cereals, rice and other grain products
   b. Bakery products
   c. Beef
   d. Poultry
   e. Pork
   f. Other meats
   g. Fish, shellfish and other seafood
   h. Eggs
   i. Milk and other dairy products
   j. Fruits and fruit juices
   k. Vegetables and vegetable juices
   l. Sugar, sugar substitutes and sweets
   m. Fats, oils and dressings
   n. Nonalcoholic beverages
   o. Miscellaneous food at home
   p. Beer, wine and other alcoholic beverages
   q. Combined food and beverages at home

II. Food Away From Home
   a. Breakfast/brunch
   b. Lunch
   c. Dinner
   d. Snacks and nonalcoholic beverages
   e. Beer, wine and other alcoholic beverages
   f. Combined food and beverages away from home

III. Nonfood Items
   a. Tobacco products and smoking supplies
   b. Personal care products
   c. Housekeeping supplies
   d. Pet food and pet supplies
   e. Nonprescription drugs and certain medical supplies

week) and dividing by the sum of the two. This continuous variable was recoded into three discrete categories, with the middle or "relatively equal" category containing a zero difference. The 25th and 75th percentiles were chosen as the dividing lines for the bottom and top categories not only to allow for a reasonable amount of deviation from equality in the middle category but also to indicate large differences between classes of respondents.

The second response pattern indicator measures the difference between the respondent's average weekly expenditure for grocery items as reported in the diary and a prior estimate of "usual" weekly expenditure given by the respondent at the beginning of the two-week diary period. Grocery items are those listed in Table 3, excluding the "Food Away From Home" categories. If the reported value is much lower than the usual expenditure, underreporting is likely. The difference between the reported expenditure and the expected or usual expenditure (reported minus expected) was divided by their sum. This variable was recorded in the same way as the weekly variation measure, and a difference of zero was again contained in the middle, "relatively equal" category. This choice allows the respondent considerable error in estimating the usual expenditure.

A third indicator is a measure of respondent style developed from the respondent's and interviewer's answers to the questions in the diary assessment section. The answers to these questions were recoded to reflect their presumed positive or negative relationships to accurate expenditure reporting. Rather than weighting the questions differently, each answer was assigned a value approximating the strength of its effect on diary keeping. Additive scales of both diary-keeping behavior and attitude toward the diary then were created using different respondent and interviewer questions in each. These scales were simplified by collapsing each into a dichotomy. With respect to accurate reporting, the categories of the attitude measure were labeled "favorable" and "unfavorable," and those of the behavior variable "desirable" and "undesirable."

At this point, the dichotomies were cross-tabulated, and a respondent style typology was developed. Those in the first category appear to be "model" respondents and are labeled the "accommodators." They had both positive attitudes and behavior and should have the least underreporting. Respondents in the next category, the "complainers," expressed dislike for the diary, but they kept it well anyway. Respondents in the third category are the "misleaders," having given a somewhat misleading picture of themselves. They reported positive attitudes, but their behavior was otherwise. Because behavior is more directly related to diary keeping than attitudes, misleaders should have greater underreporting than complainers. In the final category are those assumed to have the greatest amount of underreporting, the "resisters," who exhibited both unfavorable attitudes and undesirable behavior.

The final response pattern indicator measures to what extent the expenditure information was collected by way of a recall interview. The greater the extent of recall, the larger the underreporting problem that is expected. Interviewers were asked to indicate whether a particular week's expenditures were all recorded by the respondent, were partially obtained from recall, or were all obtained from recall. Based on this information, the classes of the response pattern variable were created.

If both diaries were completed by the respondent or no recall information was available on either diary, the amount of recall was considered to be "very small." A

"moderately small" amount of recall was assumed to be present for those with one diary completed by the respondent and the other diary completed partially by recall. When both diaries were partially completed by recall or one was completed by the respondent and the other was completed totally from recall, the amount of recall was coded as "moderately large." The amount of recalled information was considered to be "very large" if one diary was completed totally from recall and the other was completed either totally or partially from recall.

5.2. *Creation of the latent variable*

In order to combine the information from the four response pattern indicators to form a measure of response error, latent structure analysis, a technique for qualitative data which is similar to factor analysis, was used (Lazarsfeld and Henry 1968). A latent variable that is not observed directly is derived from manifest (observed) qualitative variables. This latent variable is taken to explain the relationships between the manifest variables. There can be any number of manifest variables and also more than one latent variable, just as factor analysis often produces more than one factor.

In mathematical terms, when variables A and B are not independent, the following relationship will *not* hold:

$$\pi_{ij}^{AB} = \pi_i^A \cdot \pi_j^B \tag{1}$$

where i indexes the classes of A, j indexes the classes of B, π_{ij}^{AB} is the probability an individual is in cell ij, π_i^A is the probability an individual is in class i, and π_j^B is the probability an individual is in class j.

For the above expression to be true, A and B must be independent. The purpose of the latent variable X is to achieve that independence. Thus, the following latent class model is desired

$$\pi_{ijt}^{ABX} = \pi_t^X \cdot \pi_{it}^{AX} \cdot \pi_{jt}^{BX} \tag{2}$$

where t indexes the classes of X, π_{ijt}^{ABX} is the probability of being in cell ijt of the unobserved ABX table, π_t^X is the probability that an individual is in one of the mutually exclusive and exhaustive classes of X, π_{it}^{AX} and π_{jt}^{BX} are the conditional probabilities that an individual is in a particular class of A and B, respectively, given that a person is in a certain class of X. Equation (2) indicates that, within a class of X, A and B are independent.

Goodman (1974) describes the procedure to be followed for identifying the classes of the latent variable. Clogg (1977) has developed a computer program (MLLSA) which uses Goodman's procedure to identify the latent structure model among a set of manifest variables. After the model has been defined, it is used to generate expected frequencies in the cells of the manifest table. Given these frequencies and the originally observed ones, a chi-square test is performed to determine the fit of the model. Other diagnostic statistics also are provided.

In this case, the response pattern indicators served as the manifest variables. Several latent structure models were examined before the one with three classes was selected as providing the best, most interpretable fit. Various starting points for estimating the parameters were used in the algorithm to ensure that the global solution had been reached. Table 4 indicates that, prior to the creation of the latent variable, there was a highly significant relationship between the four indicators. Afterwards, the relationship is still significant, but much less so. The chi-square value drops by 1,600 with the loss of only 20 degrees of freedom. Furthermore, the other measures of fit, lambda and the percent of cases correctly classified, indicate a good model. In this model, an individual case is assigned to a latent class based on the modal probability

Table 4. Results from the latent structure analysis creating the measure of response error in expenditure reports

	DF	Pearson χ^2	*p*
Without latent variable	133	1,747	.00
With latent variable[1]	113	153	.01
Lambda	.78		
Percent of cases correctly classified	94		
Final latent class probabilities			
Class	Probability		
Low Error	.71		
Moderate Error	.13		
High Error	.16		

[1] Degrees of freedom are 113 instead of 111 because two parameters were set to zero. Their estimates were very close to zero and negative.

associated with its cell in the manifest table. That is, respondents are placed in the latent class in which they are most likely to be, given their location in the original table. Based upon an examination of the pattern of cell allocation to the different latent classes, the three classes were given labels of "low," "moderate" and "high" response error.

Table 5 shows how the manifest indicators are related to the latent variable. These relationships are quite striking. As might be expected, response error increases with the increase in the extent of recall. Given that both the accommodators and complainers have desirable behavior, it is not surprising that they dominate the low-error class. The misleaders and resisters, on the other hand, have larger levels of error, and the majority of resisters are in the highest error class. Most respondents reporting about the same or more expenditures compared to what they usually spend are in the low error class while fewer than half of the ones reporting much less are found in this class. The relationship between the weekly variation measure and the latent variable is particularly interesting and matches previous findings (Tucker 1988). One might have expected that respondents with higher expenditure reports during the second week would be the best ones. In fact, respondents reporting about the same amounts both weeks are the best; the other two groups look very similar with respect to the latent variable.

To corroborate the labeling of the classes of the latent variable, several analyses were done. Table 6 shows that the weekly expenditures for respondents in the three classes are in the expected direction. There may be some concern that the latent variable only succeeds in differentiating between CUs of different size and income. Results of analyses within income and CU size classes indicate that this is not the case. Within some income and CU size classes all three latent class means are significantly different and in the expected direction, and in all cases the mean for the low-error class is significantly larger than the high-error class mean (based on simple random sampling). Income and CU size explain 27% of the variance in mean weekly expenditure, but the latent variable contributes another 4-5%. This is over two-thirds of what the latent variable explains by itself, so much of its explanatory

Table 5. Relationships between the latent response error measure and the manifest pattern indicators (percent)

Latent response error	Extent of recall				Respondent style typology			
	Very small %	Mod. small %	Mod. large %	Very large %	Acc. %	Com. %	Mis. %	Res. %
Low error	89	74	47	1	99	96	45	0
Moderate error	3	12	17	44	0	0	15	45
High error	8	14	36	55	1	4	40	55
N	1,763	134	242	334	1,269	380	344	480

Latent response error	Expected vs. reported expen.			Week-to-week expen. variation		
	Reported high %	Rel. equal %	Expected high %	Week 2 high %	Rel. equal %	Week 1 high %
Low error	85	81	44	68	78	63
Moderate error	10	9	15	1	22	0
High error	5	10	41	31	0	37
N	595	1,188	690	611	1,210	652

power is unrelated to income and CU size.

Table 6. Means for weekly expenditures by classes of the latent response error measure

Latent response error	*N*	Mean[1]
Low error	1,782	$94.22
Moderate error	265	65.50
High error	426	44.40

[1] All means significantly different at the 0.01 level (simple random sampling).

5.3. *Causal analysis*

Table 7 displays the relationship between procedural condition and the response error variable. The respondents to the specific diary are somewhat more likely to fall in the low-error category, but a slightly larger proportion are also in the high-error class as compared to respondents in the control condition. These results are not impressive except in the sense that the proportion of respondents in the two highest error classes is reduced by about 10% with the specific diary. Näsholm, Lindström, and Lindkvist (1989) concluded that an increase in expenditure reporting of just 5% was enough to recommend the use of preprinted category headings in the 1985 Swedish Family Expenditure Survey. It is instructive at this point to note that the expenditure means for the specific, nonspecific, and control diaries are, respectively, $87.10, $81.13, and $79.49. The mean for the specific is significantly larger than the other two.

While not presented here, the relationships between the response error measure and the demographic characteristics are stronger than that between response error and procedural condition. The youngest and the least educated respondents perform relatively poorly. Blacks and Hispanics do not perform as well as those from other ethnic backgrounds. Respondents living in

Table 7. Relationship between the latent response error measure and procedural condition

Latent response error[1]	Specific recall $N = 812$	Nonspecific recall $N = 822$	Control diary check $N = 839$
Low error	74.1%	70.9%	71.1%
Moderate error	7.9	12.0	12.2
High error	18.0	17.1	16.7

[1] Chi-square significant at the 0.01 level.

central cities and those who rent (two characteristics which coincide to some extent) have relatively high levels of response error, and the husband/wife families perform better than other types of CUs.

The findings concerning the interactions between CU characteristics and procedural condition are of the greatest interest, and the one involving age, presented in Table 8, is the most striking. While the youngest respondents do best with the control diary, the oldest perform at their worst by far on the control. Those 25 to 44, like the youngest cohort, do best with the control, but they do the poorest with the nonspecific. Respondents in the 45-64 age group have the least response error when using the specific diary. These findings do not account for the small differences between the treatment conditions in the amount of response error. If the control diary had more elderly respondents, like the specific diary, it would have made little difference. They performed very poorly on the control diary.

Table 8. Procedural condition/age interactions affecting response error

Latent response error	Age of reference person[1]					
	Under 25			25–44		
	Specific	Non-specific	Control	Specific	Non-specific	Control
Low error	52.7%	58.7%	60.9%	73.2%	70.0%	75.3%
Moderate error	17.5	22.7	20.2	7.9	12.1	10.3
High error	29.8	18.6	18.9	18.9	17.9	14.4
Latent response error	Age of reference person[1]					
	45–64			65+		
	Specific	Non-specific	Control	Specific	Non-specific	Control
Low error	78.6%	71.3%	73.9%	76.5%	77.2%	61.8%
Moderate error	5.4	11.5	8.8	8.5	8.6	18.0
High error	16.0	17.2	17.3	15.0	14.2	20.2

[1] Significant at the .01 level.

6. The Item Nonresponse Measure

6.1. *Overview of item nonresponse in the Diary Operational Test*

Because only CUs responding both weeks are being analyzed, item nonresponse is the real concern. Item nonresponse affects not only univariate estimates but also the accurate analysis of bivariate and multivariate relationships (Ferber 1966). For example, the failure of a number of respondents to provide complete income reports results in a loss of over 21% of the sample when relationships involving income are studied. Other sensitive items which may go unanswered include some individual or household characteristics, screening questions in the recall or check-item section, and many in the assessment section. Also, given the complexity of the survey, it is not surprising that items are missed or that out-of-range codes are recorded or keyed; and these, too, will affect relationships when they do not occur at random.

6.2. *Sections of the survey*

Rather than consider each item in the survey separately for determining the amount of item nonresponse, the items were divided into groups by survey section. These sections are (1) household demographics used for weighting (race, household size and homeownership), (2) demographics of the CUs reference person, (3) a set of CU characteristics which includes housing information, vehicle ownership, a description of the frequency and content of grocery store purchases, and additional demographics, (4) income and work experience information used to compute the total income for the CU, (5) expenditure information from the two diaries (i.e., whether or not there were expenditures in each diary), (6) the screening questions in the recall or check-item section, (7) the assessment section plus the items measuring extent of recall in the two diaries, and (8) a record of house guests and CU members away during the two diary weeks.

The amount of item nonresponse in each section is defined, with one exception, by the proportion of valid responses. Other than for income, questions which are for individuals and, thus, depend on family size have been excluded. Therefore, only questions relating to the CU as a whole or the reference person are used. The measure of item nonresponse in the income section is based on an estimate of the amount of useful data present and then translated into a proportion. In sections where proportions for each week are calculated, the indicator value is the average of the two proportions.

6.3. *Rating the importance of the sections*

To fully assess the effect of the pattern of item nonresponse, the importance of each section in the creation of both univariate and multivariate estimates must be taken into account. Otherwise, only a count of missing items is produced without a sense of the magnitude of the effect. In fact, if measures of response error had been created for sections other than the diary (and they should be in the future), these measures also should be weighted. This is no different than evaluating the precision of a particular survey design by focusing on a few substantively important items except, in this case, an overall measure is developed as opposed to considering each section or item separately.

Ratings of the importance assigned to each section are given in Table 9. The higher the value is, the greater the importance. Arguments can be made for different orderings or, at least, different scale values for some sections; and others were considered. On the other hand, the diary information

Table 9. Ratings of the importance of different types of information in the diary

Section	Rating of importance
Expenditure information	15
Check-item or recall section screening questions	8
Weighting demographics (household level)	6
Income and work experience	6
Reference person demographics	5
Other consumer unit characteristics	4
Diary assessment section	3
Record of house guests and CU members away	1

clearly is the most important and should have a large influence on the overall score. All of the respondents recorded expenditures in both diaries, but this would not be true if unit nonrespondents and those completing only one diary had been included. In that case, the high rating or weight for the presence or absence of expenditure information would result in a sharp distinction being made between respondents who kept both, one, or no diaries. Even in the present situation, the high rating is needed to provide an accurate picture of the overall pattern of item nonresponse. Neither the recall nor the check-item section was designed to be the primary data collection vehicle. They are expected only to supplement the diary. Total recall interviews are conducted using the diary itself. Thus, while important, they are given a rating of just about half that of the diary.

Demographic information is differentiated on the basis of its importance to the development of correct estimates. The household characteristics (often formed from CU characteristics) are needed for weighting and are used later in analysis. The income information received the same rating as the weighting variables because of its central role in economic research. Some of the information about the consumer unit is captured in the demographics just discussed. In multi-CU households, however, CU size and information about home ownership may be different than that recorded for the household. Additionally, vehicle ownership and information about usual grocery expenditures are important, especially for evaluating the quality of the expenditure data. The assessment section is not used in economic analyses, but it does provide information for evaluating the quality of the data used in them as well as for making methodological improvements. The record of house guests and CU members away can shed further light on the information contained in the individual expenditure reports.

6.4. Construction of the item nonresponse measure

As the first step in the construction of the item nonresponse measure, the proportion of valid responses in each section was weighted by the section's rating. The sum of these weighted proportions was divided by the sum of the weights (48) to give an overall weighted proportion for each respondent. This latter measure was converted to an ordinal measure of level of item nonresponse. Given that only those respondents who reported expenditures in both diaries are used, it is not surprising that over a third are in the no missing data category. Almost half of the respondents have a weighted proportion that is between 0.9 and 1.0 and are considered to have "low" item nonresponse. Only a handful of respondents have a weighted proportion below 0.7; therefore, cases with a proportion less than 0.9 are classified as having a "moderate" amount of item nonresponse.

Table 10. Relationship between level of nonresponse and procedural condition

Level of nonresponse[1]	Specific recall $N = 812$	Nonspecific recall $N = 822$	Control diary check $N = 839$
None	41.9%	40.7%	31.9%
Low	42.1	46.0	52.6
Moderate	16.0	13.3	15.5

[1] Chi-square significant at the 0.01 level.

6.5. *Causal analysis*

Table 10 presents the distributions of level of item nonresponse for the three procedural conditions. Unlike the response error measure, differences clearly exist. Compared to the specific and nonspecific conditions, the control condition has approximately 10% fewer respondents falling in the class with no missing data. This results almost exclusively from the use of the check-item section instead of the recall section. A third of the control diary respondents have missing data in the screening items of the check-item section while less than 10% of the respondents in the other two conditions have missing data in the recall section's screening questions. The specific respondents are a little more likely to have a moderate amount of missing data compared to those in the nonspecific condition because there are somewhat more incomplete income reporters in the specific.

Strong relationships exist between the CU characteristics and the level of item nonresponse. As with the latent variable, the youngest respondents have the poorest performance. Respondents with the highest level of education have less item nonresponse than the other educational groupings. Most of the other relationships also mirror the ones reported earlier for response error.

Again, there are important interactions between procedural condition and age which are displayed in Table 11. As one would expect, all age groups have more item nonresponse in the control condition; but there are some differences. For respondents under 25 or between 45 and 64, item nonresponse progressively increases as one moves from the specific to the nonspecific and then to the control. This is not the case for the other two age groups. Item nonresponse is still greatest in the control condition; but, it is, if anything, worse in the specific than in the nonspecific. Those 25 to 44 have less item nonresponse problems in the control compared to the other groups.

7. The Measure of Data Quality

7.1. *Constructing the measure*

The information about response error and item nonresponse is combined into a single indicator which represents a first attempt at a micro-level measure of overall survey quality. First of all, as response error increases, so does item nonresponse. It should be pointed out here that poor response in the recall or check-item sections not only can affect the level of item nonresponse but also response error since the failure to be asked or to answer the screening questions in these sections could result in no additional expenditures being reported. On the other hand, poor response to the

Table 11. Procedural condition/age interactions affecting level of nonresponse

Level of nonresponse	Age of reference person[1]					
	Under 25			25–44		
	Specific	Non-specific	Control	Specific	Non-specific	Control
None	36.0%	25.5%	22.3%	41.2%	43.4%	35.7%
Low	44.4	53.4	54.6	44.0	44.4	49.5
Moderate	19.6	21.1	23.1	14.8	12.2	14.8
Level of nonresponse	Age of reference person[1]					
	45–64			65+		
	Specific	Non-specific	Control	Specific	Non-specific	Control
None	42.1%	36.8%	27.5%	45.3%	44.9%	33.9%
Low	43.2	48.4	58.1	35.2	43.8	50.9
Moderate	14.7	14.8	14.4	19.5	11.3	15.2

[1] Significant at .05 level.

screening questions probably is indicative of poor response elsewhere as measured by variables like the respondent typology.

Based on the cross-classification of the two measures, respondents were assigned to one of four quality categories as defined in Table 12. Only those respondents with low response error and no missing data are classified as "very good." Otherwise, the classification scheme is fairly generous to the respondent.

7.2. Causal analysis

Table 13 presents the relationship between the data quality measure and procedural condition. Respondents in the specific condition again perform the best, but they are followed closely by the nonspecific respondents. Respondents in the control condition have the poorest quality, due in large part to their relatively high level of item nonresponse.

The bivariate associations between data

Table 12. Identification of quality categories (weighted percentage of total N)

Latent response error	Level of nonresponse		
	None	Low	Moderate
Low error	Very good (33%)	Good (33%)	Fair (6%)
Moderate error	Good (2%)	Fair (6%)	Poor (3%)
High error	Fair (3%)	Poor (8%)	Poor (6%)

Table 13. Relationship between the data quality measure and procedural condition

Quality[1]	Specific recall $N = 812$	Nonspecific recall $N = 822$	Control diary check $N = 839$
Very good	37.5%	35.1%	27.7%
Good	31.1	33.1	39.1
Fair	14.0	14.7	15.6
Poor	17.4	17.1	17.6

[1] Chi-square significant at 0.01 level.

quality and the CU characteristics, while often significant, do not add much new information. Significant interactions between procedural condition and several CU characteristics exist, and the interaction involving age is presented in Table 14. Based solely on response error, the youngest respondents perform most poorly in the specific condition; however, the quality measure shows a somewhat more complicated situation. The specific diary appears to be the best condition for those 45 to 64, but few distinctions can be made between the three diaries in the 25-44 age group. Those 65 and over still clearly do not do well in the control condition.

8. Conclusions

Three measures have been developed for assessing the effects on data quality which result from variations in the survey instrument. I believe that it has been demonstrated that useful data quality measures can be created from information contained within

Table 14. Procedural condition/age interactions affecting data quality

Quality	Age of reference person[1]					
	Under 25			25–44		
	Specific	Non-specific	Control	Specific	Non-specific	Control
Very good	28.5%	23.7%	19.8%	36.5%	37.3%	32.2%
Good	17.1	27.1	33.4	32.8	32.4	38.9
Fair	23.4	27.0	21.5	12.5	12.7	13.7
Poor	31.0	22.2	25.3	18.2	17.6	15.2
Quality	Age of reference person[1]					
	45–64			65+		
	Specific	Non-specific	Control	Specific	Non-specific	Control
Very good	38.8%	29.7%	25.1%	41.2%	41.7%	24.7%
Good	33.2	37.6	43.3	28.5	30.3	36.1
Fair	13.7	16.1	13.5	14.6	13.8	20.5
Poor	14.3	16.6	18.1	15.7	14.2	18.7

[1] Significant at the .02 level.

a survey. Of course, the measures developed here are incomplete. The measure of response error is imprecise, and the one for item nonresponse needs to be converted to a direct indicator of error. With better micro-level measures of nonsampling errors due to response and item nonresponse (as well as recording and keying), a more sophisticated indicator or set of indicators of the overall error in a respondent's data can be developed. These micro-level measures are needed because it is just too difficult to sort out the factors affecting survey quality using statistical controls at the aggregate level.

As for the substantive results in this case, the main effect of procedural differences on response error was marginal, with the specific respondents performing slightly better than the others. Nonetheless, this difference may be worth considering. A greater main effect (when comparing the control condition to the other two) was observed for the item nonresponse measure, but most of the difference was confined to the section for collecting recalled information. The main effect of procedural condition on the measure of data quality was largely a combination of the effects observed for the other two variables, although the quality indicator does provide a more complete picture of data problems and makes further distinctions between individual respondents.

The important story, however, is the joint effects of CU characteristics and procedural condition illustrated using the age variable. These results show that data quality not only varies according to individual CU characteristics but also that these individual differences can be affected by the type of diary instrument. When the effect of the recall method is factored out, the differences have to do largely with the amount of underreporting engendered by the three diary formats in various subpopulations.

The interactions between procedural condition and certain CU characteristics indicate performance might be improved by tailoring the instrument to the respondent. The reactions of the youngest and oldest respondents to the different diary formats provide an example of where this could be advantageous. The young may prefer the control or the nonspecific diary because they find it easier to write down their purchases rather than look for the correct line in the specific. At the other extreme, it appears that the elderly prefer the greater structure present in both the specific and the nonspecific. Psychologists have documented that problem-solving strategies differ between the young and old (Birren and Schaie 1985). Older adults often use more "primitive" strategies, and the greater structure in the two experimental diaries could make these strategies more effective. Younger respondents may be better able to develop their own structures.

9. References

Birren, J.E. and Schaie, K.W. (eds.) (1985). The Handbook of the Psychology of Aging. New York: Van Nostrand Reinhold.

Clogg, C.C. (1977). Unrestricted and Restricted Maximum Likelihood Latent Structure Analysis – A Manual for Users. Working Paper No. 1977–09, Population Issues Research Office, Pennsylvania State University.

Corby, C. and Miskura, S. (1985). Evaluating Data Quality in the Economic and Decennial Censuses. Proceedings of the First Annual Research Conference, U.S. Bureau of the Census, 159–175.

Fay, R.E. (1983). CPLX – Contingency Table Analysis for Complex Sample Designs, Program Documentation. Technical Report, U.S. Bureau of the Census.

Ferber, R. (1966). Item Nonresponse in a

Consumer Survey. Public Opinion Quarterly, 30, 399–415.

Flueck, J.A., Waksberg, J., and Kaitz, H.B. (1971). An Overview of Consumer Expenditure Survey Methodology. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 238–246.

Goodman, L.A. (1974). The Analysis of Systems of Qualitative Variables When Some of the Variables are Unobservable: Part I - A Modified Latent Structure Approach. American Journal of Sociology, 79, 1179–1259.

Grootaert, C. (1986). The Use of Multiple Diaries in a Household Expenditure Survey in Hong Kong. Journal of the American Statistical Association, 81, 938–944.

Groves, R.M. and Magilavy, L. (1984). An Experimental Measurement of Total Survey Error. Proceedings of the Section on Survey Research Methods, American Statistical Association, 698–703.

Jacobs, C.A. (1983). 1983 CES-Analysis of Reporting Completeness. U.S. Bureau of Labor Statistics Memorandum to John L. Marcoot.

Kemsley, W.F.F. and Nicholson, J.L. (1960). Some Experiments in Methods of Conducting Family Expenditure Surveys. Journal of the Royal Statistical Society, Ser. A, 123, Part 3, 307–328.

Lazarsfeld, P.F. and Henry, N.W. (1968). Latent Structure Analysis. Boston: Houghton-Mifflin.

Madow, W.G. (1973). Net Differences in Interview Data on Chronic Conditions and Information Derived from Medical Records. Series 2, No. 57, Vital and Health Statistics.

Neter, J. (1970). Measurement Errors in Reports of Consumer Expenditures. Journal of Marketing Research, 7, 11–25.

Näsholm, H., Lindström, H., and Lindkvist, H. (1989). Response Burden and Data Quality in the Swedish Family Expenditure Survey. Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, 501–514.

Pearl, R.B. (1979). Reevaluation of the 1972-73 U.S. Consumer Expenditure Survey. Technical Paper No. 46, U.S. Bureau of the Census.

Pearl, R.B. and Levine, D.B. (1971). A New Methodology for a Consumer Expenditure Survey. Proceedings of the Business and Economic Statistics Section, American Statistical Association, 254–259.

Roghmann, K.J. and Haggerty, R.J. (1972). The Diary as a Research Instrument in the Study of Health and Illness Behavior: Experience with a Random Sample of Young Families. Medical Care, 10, 143–163.

Silberstein, A.R. and Scott, S. (1991). Expenditure Diary Surveys and Their Associated Errors. In Measurement Errors in Surveys, ed. P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman, New York: John Wiley, 303–326.

Stanton, J.L. and Tucci, L.A. (1982). The Measurement of Consumption: A Comparison of Surveys and Diaries. Journal of Marketing Research, 19, 274–277.

Sudman, S. and Bradburn, N.M. (1974). Response Effects in Surveys. Chicago: Aldine Publishing Company.

Sudman, S. and Ferber, R. (1971). Experiments in Obtaining Consumer Expenditures by Diary Methods. Journal of the American Statistical Association, 66, 725–735.

Thompson, W.L., Carlson, L.T., Woteki, T.H., and Vagts, K.A. (1980). Improving the Quality of Data from Monthly Gasoline Purchase Diaries. Proceedings of the Section on Survey Research Methods, American Statistical Association, 652–654.

Thran, S.L., Marder, W.D., and Willke, R.J. (1986). Probability of Response: A Multivariate Probit Analysis. Paper presented at the meeting of the American Medical Association, September.

Tucker, C. (1984). Decision-Making in Diary Research Planned for 1985. U.S. Bureau of Labor Statistics Memorandum to Curtis A. Jacobs.

Tucker, C. (1988). Estimating Measurement Error in Surveys. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Toronto, Canada, May 19–22.

Tucker, C., Vitrano, F., Miller, L., and Doddy, J. (1989). Cognitive Issues and Research in the Consumer Expenditure Diary Survey. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, FL, May 20–23.

Tucker, C. and Bennett, C. (1988). Procedural Effects in the Collection of Consumer Expenditure Information: The Diary Operational Test. Proceedings of the Section on Survey Research Methods, American Statistical Association, 256–261.

Turner, R. (1961). Inter-week Variations in Expenditures Recorded During a Two-week Survey of Family Expenditures. Applied Statistics, 10, 136–146.

U.S. Department of Labor (1986). Consumer Expenditure Survey: Diary Survey, 1982–1983. Bulletin 2245, Bureau of Labor Statistics.

Received March 1991
Revised January 1992